Optimisation de l'apprentissage par récupération en mémoire pour promouvoir la rétention à long terme de nouvelles connaissances

Alice Latimier

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

Optimisation de l'apprentissage par récupération en mémoire pour promouvoir la rétention à long terme de nouvelles connaissances

Soutenue par

Alice LATIMIER

Le 22 Novembre 2019

Ecole doctorale n° 158

ED3C : Cerveau, Comportement, Cognition

Spécialité

Sciences cognitives

Composition du jury :

Éric JAMET
Professeur, Université Rennes 2 — *Président et Rapporteur*

Fabien MATHY
Professeur, Université Nice Sophia Antipolis
Rapporteur

Emilie GERBIER
Maitre de Conférences, Université Nice Sophia Antipolis — *Examinatrice*

Bérengère GUILLERY-GIRARD
Maitre de Conférences, Université Caen — *Examinatrice*

Franck RAMUS
Directeur de recherche CNRS, Ecole Normale Supérieure
Directeur de thèse

Roberto CASATI (absent)
Directeur de recherche CNRS, Ecole Normale Supérieure
Co-directeur de thèse

ENS | PSL

A mes grands-parents, Hélène et Georges

Abstract

Learning in formal school contexts usually unfolds in a study phase consisting in studying and memorizing the contents of a lecture followed by an assessment phase that allows teachers to evaluate their students' knowledge. Within this configuration, the assessment phase is confined to a pure testing of the learning without playing any further role in the learning process. However, research in cognitive science has provided strong evidence for the use of tests during the study phase to enhance long-term memory. The effect of testing one's knowledge with retrieval practice is actually more efficient than merely repeated reading (testing effect). In addition, inserting a time interval between episodes of the study sessions promotes better consolidation than massed practice that consists in learning the same piece of information in a repetitive fashion without inter-study interval (spacing effect). As opposed to commonly used learning strategies such as reading and cramming, retrieval practice and spaced learning have been shown to promote better long-term retention among a great variety of populations and contents. However, to this day, teachers and students still rarely employ these effective strategies.

The start-up company Didask designed a teaching platform with the same name that incorporates the results of research to offer an evidence based learning tool. In my PhD project, I have used Didask to raise the following issue: how can one optimize the benefits of retrieval practice relative to the exposure to learning contents of a lecture?

The first chapter of this dissertation is a general introduction that contextualizes the present research project. The second chapter presents three meta-analyses conducted on the effect of spaced retrieval practice. Spaced retrieval practice consists of repetitions of the same retrieval event distributed through time. Results of the first meta-analysis indicated a significant benefit of spaced retrieval practice on retention relative to massed retrieval practice ($g = 0.74$); while the second meta-analysis did not show a significant benefit of the spaced retrieval practice relative to spaced restudying ($g=0.46$). The last meta-analysis indicated no difference between the expanding schedule (i.e., spacing intervals between repeated reading episodes increase with every repetition) and the uniform one (spacing intervals are kept constant throughout the study phase; $g = 0.032$). Together, these analyses support the advantage of spaced retrieval practice, but do not support the wide belief that intervals should be progressively increased.

Chapters 3, 4, and 5 refer to three experiments run on Didask. These experiments investigated the benefits of different placements and schedules of retrieval practice episodes on long-term retention, one week and one month after the learning session. Experiment 1 demonstrated that learning with retrieval practice was better when retrieval practice episodes were placed after rather than before reading the learning contents. Moreover, the benefit of post-testing over pre-testing transferred to untrained information during learning. Experiment 2 demonstrated that the granularity of the learning contents is especially of interest when learning with successive readings: short reading passages led to a better retention than longer readings. However, when learning with retrieval practice, granularity did not matter for long-term retention. Finally, Experiment 3 replicated the testing effect but not the spacing effect. Contrary to our predictions, the effect of the combination of both learning strategies was not significant but using spaced retrieval practice led to better retention in average than massed retrieval practice, spaced reading, and massed reading. Overall, the results of this thesis give a better understanding on how learners should use retrieval practice and suggest new research questions in optimizing learning.

Résumé

L'apprentissage scolaire implique généralement une phase d'étude des cours suivie d'une phase d'évaluation pour mesurer l'efficacité de la première. Dans cette conception, la phase de test sert à quantifier la réussite de l'apprentissage mais n'est pas envisagée comme outil d'apprentissage. Pourtant, de nombreuses études ont montré l'importance des tests comme processus actif pour consolider des connaissances à long terme. Une autre pratique peu utilisée est la distribution d'un même apprentissage dans le temps. Alors que le bachotage favorise la mémorisation à court terme, l'espacement des révisions favorise la consolidation sur le long terme. A l'heure actuelle, ces méthodes sont méconnues des enseignants et des élèves alors que les bénéfices de l'entrainement par récupération en mémoire et de l'espacement ont été répliqués de manière robuste avec des populations et contenus variés.

La start-up Didask a créé la plate-forme d'enseignement numérique du même nom en incorporant les résultats de ces recherches menées sur l'apprentissage. En collaboration avec Didask, ma thèse s'est articulée autour de la problématique suivante : quel est l'agencement optimal des phases de récupération en mémoire par rapport à la présentation des contenus d'un cours ?

Le premier chapitre est une introduction générale pour contextualiser cette recherche. Le second chapitre compile les résultats de trois méta-analyses portant sur l'effet de l'apprentissage avec tests répétés et espacés dans le temps. La première méta-analyse a montré un bénéfice significatif de l'apprentissage avec tests espacés dans le temps sur la rétention en mémoire par rapport à l'apprentissage avec tests massés dans le temps ($g = 0.74$). La seconde méta-analyse suggère en revanche un bénéfice non significatif de l'apprentissage avec tests espacés par rapport à l'apprentissage avec relectures espacées dans le temps ($g=0.46$). La dernière méta-analyse n'a pas démontré de différence entre un planning expansif d'apprentissage avec tests espacés (accroissement progressif de l'intervalle de temps entre les sessions d'apprentissage) et un planning uniforme (maintien du même intervalle entre les sessions, $g = 0.032$). L'ensemble de ces résultats confirme le net avantage de l'apprentissage avec tests espacés dans le temps, mais le planning d'espacement optimal n'est pas nécessairement expansif.

Les chapitres 3, 4, et 5 présentent trois expérimentations menées sur Didask. L'objectif était de mesurer les bénéfices de différents emplacements et planning de tests d'apprentissage sur la rétention en mémoire une semaine à un mois après la session d'apprentissage. Les résultats de l'Expérience 1 ont démontré qu'il est préférable de faire des tests d'apprentissage après la lecture du cours pour une meilleure mémorisation plutôt qu'avant. De plus, l'avantage de faire des tests après la lecture du cours était observé sur les informations non testées lors de l'apprentissage. Les résultats de l'Expérience 2 indiquent que le degré de granularité des contenus importe lors de l'apprentissage par relectures successives : un découpage fin permet une meilleure mémorisation qu'un découpage plus grossier. Par contre, l'importance de la granularité disparait dans la condition avec des tests d'apprentissage. Enfin, l'Expérience 3 a répliqué l'effet de récupération en mémoire mais pas l'effet d'espacement. Contrairement aux hypothèses de départ, l'effet de la combinaison des deux stratégies d'apprentissage n'était pas significatif. Néanmoins, l'apprentissage avec tests espacés permettait en moyenne une meilleure mémorisation que l'apprentissage avec tests massés, avec relectures espacées, et avec relectures massées dans le temps. Les résultats de cette thèse apportent une meilleure compréhension sur la manière d'utiliser l'apprentissage par les tests. Ils suggèrent également de nouvelles pistes de recherche sur l'optimisation des apprentissages pour promouvoir la consolidation de nouvelles connaissances.

Remerciements

Je tiens à remercier chaleureusement mon jury, Mesdames Emilie Gerbier et Bérengère Guilley-Girard, ainsi que Messieurs Eric Jamet et Fabien Mathy, qui m'ont fait l'honneur d'évaluer mon travail. Je ne pouvais pas espérer mieux pour parler *testing* et *spacing effects* ! J'en profite pour remercier Bérengère et Fabien pour leurs précieux conseils dans le cadre de mon comité de suivi de thèse.

Au pays des merveilles, Alice a vécu des aventures extraordinaires. Pendant mon séjour de 5 ans au Département d'Etudes Cognitives, j'ai moi même vécu des aventures, mais plus de l'ordre du scientifique que du fantasmagorique. Néanmoins, tout comme la Alice de Lewis Carroll, j'ai moi aussi fait des rencontres exceptionnelles qui m'ont permises de voir le monde autrement et dont je garderai un souvenir impérissable.

Franck et Roberto, merci de m'avoir permis d'être l'heureuse élue pour réaliser ce beau projet de recherche. Franck, je tiens à te remercier pour ta confiance et pour la grande liberté que tu m'as donnée durant ces trois ans tout en posant ton regard bienveillant et rigoureux sur mon travail. Ton rationalisme et ta force tranquille vont me manquer. C'était extrêmement formateur de travailler dans ton équipe (et quelle équipe !) et je renouvellerai l'expérience les yeux fermés ! Roberto, même si nous avons eu de rares occasions pour interagir, ces moments étaient très agréables. Merci pour tes précieuses relectures en fin de course.

Ce projet de recherche n'aurait pu se faire sans l'équipe incroyable de l'entreprise Didask, ma deuxième maison professionnelle après le labo. Merci pour votre accueil chaleureux, la porte était toujours ouverte et vous m'avez véritablement intégré à la vie de l'entreprise pour que je m'y sente bien, c'était très réussi. J'ai passé d'excellents moments en votre compagnie et je vous souhaite le meilleur pour la suite (qui s'annonce déjà sensationnelle !). Arnaud et Thierry, voir l'évolution de votre « bébé » était une véritable expérience en soi. Je suis fière de l'aboutissement de cette collaboration, cela n'aurait pas été possible sans votre dévouement et sans votre conviction pour produire des données empiriques sur votre plateforme et améliorer l'apprentissage.

Un grand merci à l'association Synlab (à Joyce, Aude, et Florence) pour le suivi du projet Parcours Connectés. Je tiens également à remercier les autres partenaires de la partie recherche du projet pour les échanges constructifs lors de nos réunions, Maud et Niluphar pour le LP3C, Benoît et Fabrice pour le LRI.

Avant de commencer la thèse au Laboratoire de Sciences Cognitives et Psycholinguistique, j'ai eu la chance de travailler dans d'autres labos du DEC. Je tiens à remercier Jean Lorenceau, qui a l'époque m'a fait confiance pour continuer un projet de recherche appliqué passionnant au Laboratoire des Systèmes Perceptifs, alors que je finissais tout juste mon master à Jussieu. Ensuite, j'ai eu la chance d'intégrer l'équipe de Tiziana Zalla à l'Institut Jean Nicod. Pendant un an et demi, cette expérience de recherche fut très enrichissante aussi bien sur le plan scientifique que sur le plan humain. Je regrette de ne pas avoir pu lui partager ma reconnaissance quand cela était encore possible. Un grand merci à mes incroyables collègues de cette période : Marina, Marco, Laura, Marine, Laora, Isabelle, Marie, Klara et Doriane. Nathalie et Vincent, merci beaucoup pour votre gentillesse, votre professionnalisme et votre bonne humeur.

Pendant la thèse j'ai pu compter sur mes collègues du LSCP, tous aussi fantastiques et brillants les uns que les autres. Je ne te remercierai jamais assez Hugo pour le temps que tu as pris pour travailler avec moi, c'était très stimulant de faire des papiers ensemble ! Un grand merci à

Benjamin, Matthieu, et Jean-Maurice pour la très bonne ambiance dans notre premier bureau au pavillon jardin, je vous souhaite le meilleur pour cette dernière année. Je remercie mes anciens collègues d'équipe Mikel et Raphaël pour leur gentillesse et leurs conseils. Georgia, Monica, Camila, Camille, Ghislaine, Cécile, Leticia et Baptiste merci pour votre soutien, l'ambiance de notre bureau va terriblement me manquer ! Prenez soin de nos belles plantes ! Mention spéciale à Ava et Gerda, mes deux piliers de la thèse. Qu'est-ce que j'aurais fait sans votre écoute, votre générosité, vos encouragements, votre humour, votre expertise indiscutable en statistiques, votre vivacité d'esprit, votre sourire, et sans les litres de thé bus ensemble ? Qu'est-ce que j'aurais fait sans vous tout simplement ?

Au tout début de ma thèse, j'ai rencontré la brillante équipe de l'association Cog'Innov, dans laquelle j'ai eu grand plaisir à m'investir pour partager mes connaissances et organiser des événements de vulgarisation de grande qualité. Judith, je vais suivre de près l'évolution de ce beau projet que tu portes à merveille. Continue de croire en tes idées ! Gaëtan, Marie et Emma, merci pour votre agréable compagnie dans les locaux de Didask, très bonne continuation pour Cog'X ! Pour avoir largement participé à ma formation à la vulgarisation scientifique, je remercie toute l'équipe de médiatrices de la Cité des Enfants de la Cité des Sciences. Que des bons souvenirs de mes animations, très bien encadrées par Florence. Depuis 2015, j'ai la chance de collaborer avec l'équipe passionnée et passionnante des Savanturiers au CRI. Ange, Margaux, et Julie, un grand merci pour vos encouragements permanents. C'était un vrai plaisir de suivre l'évolution de votre mission. Grâce à vous j'ai fait de très belles rencontres avec des enseignants engagés et pu partager le fruit de mon travail avec eux.

Au delà de mon entourage professionnel exceptionnel, j'ai aussi pu compter sur le soutien de mes amis. Pour tous les bons moments de déconnexion avec vous, et pour vos épaules précieuses sur lesquelles je peux me reposer les yeux fermés n'importe quand : infinis remerciements à Stéphanie, Charlotte, Dorian, Ava, Svetlana, Benoît, Marina, à ma montréalaise préférée Flavie, et à mon globe-trotter préféré Yvain. Egalement un grand merci à mes amis Melunais et Balnéolaises pour les balades en forêt à Fontainebleau, les vacances « découvertes » de notre beau pays, et autres soirées-concerts bien arrosées ! Vous allez tous me manquer, visitez-moi à Dijon quand vous voulez.

J'aimerais finir ces remerciements émouvants par ma famille. Je ne serai jamais assez reconnaissante envers mes parents pour le dévouement démesuré pour leurs enfants, je vous dois ma réussite. Pour le lien unique que vous avons et les innombrables fous rires vécus ensemble, je partage également ma réussite avec mon frère Antoine et ma sœur Adèle, mes plus grands fans. Un immense merci à mon oncle et ma tante toujours à l'écoute et rassurants quand j'avais des doutes sur mon avenir et ma capacité à réussir dans la recherche. Je souhaite dédier ce travail à mes grands-parents alsaciens et charentais, sans qui je n'aurais pas pu mener mes études confortablement et avec qui j'aurais tellement aimé partager cet aboutissement. Enfin, un grand merci à ma belle famille, sans qui la vie parisienne serait vraiment triste !

Grégoire, il me semble avoir déjà communiqué tout l'amour que je te porte mais je le renouvelle. Tu mérites autant que moi le titre de docteur, cette thèse c'est aussi la tienne. Maintenant, allons au bout du monde pour célébrer !

Publications et valorisation des travaux de recherche

- **Articles scientifiques peer reviewed**

Latimier, A., Peyre, H., Ramus, F. A quantitative review of the "spacing-of-tests" effect (in prep).

Latimier, A., Riegert, A., Ly, S.T., Ramus, F. Retrieval practice promotes long-term retention irrespective of grain size (in prep).

Latimier, A., Riegert, A., Ly, S. T., Ramus, F. Do spacing and retrieval practice effects interact? (in prep).

Latimier, A., Riegert, A., Peyre, H., Ly, S.T., Casati, R., & Ramus, F. (2019). Does pre-testing promote better retention than post-testing? *Npj Science of Learning*, *4*(1), 1–7. https://doi.org/10.1038/s41539-019-0053-1

- **Articles de vulgarisation**

Latimier A. (juin 2018). « Parcours connectés » : accompagner la formation des enseignants dans la transformation numérique - Média *The Conversation France.*

Latimier A., Meyer S. (octobre 2017). Et si les réformes de l'éducation s'appuyaient sur les données de la science? - Média *The Conversation France.*

Latimier A. (octobre 2017). « 10 idées reçues sur l'apprentissage » – Interview pour le dossier de couverture du magazine *Ça m'intéresse*, numéro d'Octobre 2017.

Duchamp L., Latimier, A. (juillet 2017). L'illusion de maîtrise, un danger pour l'apprentissage – Blog *DisDonc Didask*

Latimier A. (juin 2017). Notre cerveau peut apprendre à tout âge (Il n'est jamais trop tard pour s'y mettre !) – Publication simultanée sur *Cog'Innov* et le média *The Conversation France.*

Latimier A. (juin 2017). Les difficultés désirables au service des apprentissages durables – Blog *DisDonc Didask.*

Latimier A. (juin 2017). Il n'y a pas d'âge pour apprendre – Blog *DisDonc Didask*

- **Posters et communications dans des congrès**

Latimier, A., Peyre, H., Ramus, F. (08/2019). A quantitative review of the "spacing-of-tests" effect**.** *18th Biennial European Association for Research on Learning and Instruction Conference (Aachen).*

Latimier, A., Riegert, A., Ly, S. T., Casati, R., Ramus, F. (07/2019). What is the optimal grain size when learning with study-retrieval practice? *Twenty-sixth International Conference on Learning (Belfast).*

Choffin, B., Latimier, A., Ahmadi, N., Ramus, F., Popineau, F. (05/2019). Making sense of learner behavioral, cognitive and demographic characteristics to improve learner modeling**.** *3rd International Congress on Technologies in Education (Paris).*

Latimier, A., Riegert, A., Peyre, H., Ly, T., Casati, R., Ramus, F. (03/2019). Does pre-testing promote better retention than post-testing? *International Convention of Psychological Science (Paris).*

Latimier, A. (20/02/2019). Workshop: Learning through questioning. *World Innovation Summit of Education (Centre de Recherche Interdisciplinaire, Paris).*

Latimier, A., Riegert, A., Peyre, H., Ly, S.T., Casati, R., Ramus, F. (2018). To enhance long term retention, should students be tested before or after exposure to learning contents? *29th International Congress of Applied Psychology (Montréal).*

Latimier, A., Riegert, A., Peyre, H., Ly, S.T., Casati, R., Ramus, F. (2018). To enhance long term retention, should students be tested before or after exposure to learning contents? *29th International Congress of Applied Psychology (Montréal).*

Latimier, A., Riegert, A., Peyre, H., Ly, S.T., Casati, R., Ramus, F. (2017). To enhance long term retention, should students be tested before or after exposure to learning contents ? *58ème Congrès de la Société Française de Psychologie (Nice).*

TABLE DES MATIERES

`And how many hours a day did you do lessons?' said Alice, in a hurry to change the subject.
`Ten hours the first day,' said the Mock Turtle: `nine the next, and so on.'
`What a curious plan!' exclaimed Alice.
`That's the reason they're called lessons,' the Gryphon remarked: `because they lessen from day to day.'
This was quite a new idea to Alice, and she thought it over a little before she made her next remark. `Then the eleventh day must have been a holiday?'
`Of course it was,' said the Mock Turtle.

— Lewis Carroll
Alice's Adventures in Wonderland (1865, 1869), Chapter IX *The mock turtle story*.

1 Chapitre 1 : Introduction générale

1.1 Vers une éducation fondée sur des preuves

L'éducation fondée sur des preuves s'inscrit dans le mouvement plus large de la pratique fondée sur des preuves dont le but principal est « *d'utiliser les meilleures preuves disponibles pour produire des effets désirables, et réciproquement pour éviter des effets indésirables* » (Kvernbekk, 2017). En écho à la médecine fondée sur des preuves (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996), l'éducation fondée sur des preuves vise à intégrer les connaissances produites empiriquement dans les pratiques pédagogiques plutôt que de se baser uniquement sur des intuitions sans fondements scientifiques (Hammersley, 2007; Slavin, 2002, 2004). L'éducation fondée sur des preuves a depuis longtemps fait sa place dans les pays anglo-saxons, mais peine toujours à se faire une place en France. Pourtant, cette démarche est très bien documentée (ex., Gorard, See, & Siddiqui, 2017; Laurent et al., 2009) et la création de ponts entre recherche et éducation représente une piste très prometteuse pour réformer l'école (Farley-Ripple, May, Karpyn, Tilley, & McDonough, 2018). En France, et restreinte aux neurosciences, cette mouvance a pris le nom de « neuro-éducation » sous l'impulsion de quelques chercheurs mais sa légitimité pour offrir à elle seule des applications directes pour optimiser l'apprentissage est désormais remise en question (Bowers, 2016). Dans une perspective plus interdisciplinaire, la recherche en sciences cognitives a sa place dans cette démarche. En effet, au-delà des résultats solides qu'elle fournit sur le fonctionnement cérébral et le comportement humain (Dessus & Gentaz, 2006); la démarche expérimentale qu'elle utilise est très pertinente pour comprendre comment améliorer l'apprentissage et pour s'intégrer dans des projets de « recherche-action » dans l'enseignement (Andler & Guerry, 2008; Bruer, 2008; Mesnier & Missotte, 2003; Pasquinelli, 2013). Le projet de recherche présentée ici s'inscrit dans cette mouvance qui vise à transformer les méthodes d'apprentissage en se fondant sur des données issues des travaux de recherche en psychologie (Davies, 1999; Pasquinelli, 2011).

1.1.1 Pourquoi aller vers une éducation fondée sur des preuves ?

La démarche d'une éducation fondée sur des preuves reste encore timide en France car il existe une tension entre les évaluations des pratiques pédagogiques, qui demandent à être conduites à grande échelle ; et la nécessité de prendre en compte des éléments fins du processus

éducatif, un sujet qui a fait l'objet de recherches prometteuses en sciences cognitives (Brown, Roediger III, & McDaniel, 2014; Hattie, 2008; Weinstein, Madan, & Sumeracki, 2018; Coalition for Evidence-Based Policy, 2013). Cette nouvelle façon d'envisager l'enseignement répond pourtant à un véritable besoin de transformation de la pédagogie et des méthodes d'apprentissage communément utilisées par les apprenants et les enseignants. En effet, certaines méthodes d'apprentissage usuelles n'ont pas démontré de réelle efficacité pour une maitrise durable des connaissances, alors qu'un des objectifs principaux de l'école est de former « pour toute la vie » (Avis & Fisher, 2018; Demirel, 2009). Aussi, ces méthodes classiques ne permettent pas de résoudre les problèmes d'hétérogénéité dans certaines classes (Galand, 2009). Les enseignants doivent prendre en compte les différences individuelles de leurs élèves tout en favorisant un climat de classe optimal (ex, différences de milieux socio-économiques, différences de niveaux sur les acquis des fondamentaux, troubles des apprentissages; Bianco, 2016; Marzano, 2001). Dans ce contexte, les enseignants n'ont pas toujours la possibilité ni les moyens de faire du « cas par cas » pour s'adapter au rythme et besoins de chaque élève (Cnesco[1], 2017). Enfin, et cela va de pair avec ces raisons, les enseignants sont à la recherche d'outils et bonnes pratiques pour accompagner leur pédagogie et dont ils peuvent avoir la certitude qu'elles sont efficaces. Malheureusement, il n'est pas évident pour eux de savoir où chercher ni comment s'informer sans tomber dans le piège des « neuromythes » qui séduisent par leur simplicité (Dekker, Lee, Howard-Jones, & Jolles, 2012; Pasquinelli, 2012; Tricot, 2017). Un exemple bien connu de neuromythe est la théorie des styles d'apprentissage (typiquement auditif, visuel, kinesthésique). Elle a longtemps été considérée comme véridique et nombre d'enseignants formés avec ce type de connaissances ont cherché à appliquer cette théorie en classe pour s'adapter aux particularités individuelles de leurs élèves (dans le cadre d'un enseignement différencié, Landrum & McDuffie, 2010). Pourtant, cette théorie est critiquée par la communauté scientifique car les résultats de la recherche n'ont pas démontré l'existence de ces styles d'apprentissage (Pashler, McDaniel, Rohrer, & Bjork, 2008; Rohrer & Pashler, 2012).

Dans le domaine 2 du nouveau socle commun proposé en 2016 par le Ministère de l'Éducation Nationale[2] et intitulé « *Méthodes et outils pour apprendre* » ; il est spécifié que l'objectif est de permettre à tous les élèves « *d'apprendre à apprendre* ». Ils doivent ainsi « *savoir apprendre une leçon […] s'entraîner en choisissant les démarches adaptées aux*

[1] http://www.cnesco.fr/fr/differenciation-pedagogique/
[2] Socle commun de connaissances, de compétences et de culture : https://eduscol.education.fr/cid98781/le-numerique-et-le-socle-commun.html#lien1

objectifs d'apprentissage préalablement explicités ». Dans cette démarche, il parait donc plus que nécessaire de développer les compétences métacognitives liées à l'apprentissage chez les élèves, à savoir leur capacité à évaluer l'efficacité de leurs méthodes d'apprentissage et à les réguler pour optimiser leur réussite. Une des raisons de l'échec de certains étudiants serait un excès de confiance dans leur capacité à connaitre leur cours (Hacker, Bol, Horgan, & Rakow, 2000) et donc une tendance à sous-estimer la vitesse à laquelle ils risquent d'oublier les connaissances apprises (Koriat, Bjork, Sheffer, & Bar, 2004). Hors, comme nous allons le développer dans les sections suivantes, certaines méthodes d'apprentissage créent des « illusions de maitrise » sans réellement permettre une consolidation à long terme mais qui renforcent le sentiment de confiance dans la maitrise des connaissances (Kirschner & van Merriënboer, 2013). Plusieurs études ont démontré que les étudiants avaient tendance à privilégier ces méthodes de révisions qui maximisent la mémorisation à court terme et dont les bénéfices sur le long terme sont au contraire faibles (ex: Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler et al., 2007). Par conséquent, il existe un vrai décalage entre les pratiques des élèves au quotidien et les résultats de la recherche sur l'optimisation des apprentissages.

Considérant ces constats, adopter des méthodes pour apprendre dont l'efficacité est véritablement basée sur des données issues de la recherche parait une approche prometteuse ; plutôt que de se baser sur de simple intuition sans fondement empirique. La section 1.1.2 *Les stratégies d'apprentissage efficaces* (p11) aborde de manière plus exhaustive ce décalage entre les pratiques communément utilisées par les élèves et les résultats des expérimentations en laboratoire sur l'optimisation des apprentissages.

1.1.2 Comment aller vers une éducation fondée sur des preuves ?

Un premier levier pour privilégier une éducation fondée sur des preuves est de former les futurs enseignants, professeurs, et plus généralement tous les responsables de formations aux résultats de la recherche en sciences cognitives. Les institutions responsables de la formation des futurs enseignants doivent mettre en place les moyens pour répondre aux directives du socle commun, et donc proposer des modules spécifiques sur les connaissances théoriques en psychologie cognitive et leurs applications potentielles dans les pratiques pédagogiques. Proposer des moyens pour donner aux enseignants la possibilité d'expérimenter eux-mêmes certaines pratiques pédagogiques pour mesurer leurs effets sur l'apprentissage est

également un objectif important à remplir dans la démarche de transformation et d'amélioration de leurs pratiques au quotidien. En lien avec cet objectif, il est également important de créer de véritables ponts entre laboratoires de recherche et établissements scolaires. D'une part afin de tester des méthodes en contextes d'apprentissage réalistes en prenant en compte toute la complexité de l'environnement scolaire ; et d'autre part pour mieux intégrer le corps enseignant dans la démarche scientifique lors de la mise en place de telles interventions à grande échelle puisqu'ils sont les premiers concernés.

Un second levier pour aller vers une éducation fondée sur des preuves, et qui découle du premier, est de former explicitement les élèves aux méthodes « qui marchent » par rapport à celles « qui ne marchent pas » tout en leur fournissant des outils concrets pour mettre en place de nouvelles méthodes correctement (en lien avec l'axe « *Apprendre à Apprendre* » du domaine 2 du socle commun, Bjork & Yan, 2014). Cela aurait pour double objectif de stimuler leur métacognition pour qu'ils soient capables de réguler en autonomie leur processus d'apprentissage, indirectement d'être plus engagés dans ce processus ; et d'autre part de leur permettre d'obtenir de meilleures notes.

Enfin, ces actions sont indissociables des travaux issus de la recherche fondamentale dans le domaine de la psychologie cognitive. Les résultats ainsi produits peuvent mener à des recommandations pratiques et à des applications concrètes pour l'enseignement (Dunlosky & Rawson, 2019; Roediger, Finn, & Weinstein, 2012; Rohrer & Pashler, 2010). Cela passe notamment par une méthode dite quantitative, c'est à dire qui mène à la production de mesures précises de tailles d'effets (Simpson, 2017) dans le cadre d'expérimentations contrôlées et randomisées (Styles & Torgerson, 2018) et avec des échantillons de participants les plus conséquents possibles. De plus, des réplications des effets mesurés permettent de renforcer la fiabilité des résultats (Francis, 2012; Makel & Plucker, 2014). Lorsqu'un certain nombre d'expérimentations ont été conduites pour mesurer un même effet, il est ensuite indispensable de réunir ces études pour réaliser ce qu'on appelle une méta-analyse, afin d'apporter de la robustesse à cet effet. Les revues systématiques de littérature qui compilent toutes les études mesurant un même effet sont aussi une étape indispensable de la démarche de production de preuves robustes et généralisables (Arthur, Waring, Coe, & Hedges, 2012; Hattie, Rogers, & Swaminathan, 2014). Sans ce type de travaux fournissant les tailles d'effet de différentes méthodes d'apprentissage, il paraît très compliqué de recommander leur utilisation. Cette

démarche quantitative se nourrit des études qualitatives qui sont elles aussi indispensables dans une démarche plus exploratoire de transformation de la pédagogie (Kozleski, 2017).

Pour résumer, une interaction forte entre classes et laboratoires semble être la clé pour promouvoir une éducation fondée sur des preuves. Ce pont ne peut se construire sans l'intervention des acteurs de l'éducation nationale et de l'enseignement, et plus largement sans celle des décideurs politiques. Ces derniers ont pour rôle de créer et maintenir une dynamique d'échange entre la communauté enseignante qui vit les problématiques du terrain, et les chercheurs qui peuvent mettre à profit leur expertise pour comprendre la source de ces problèmes. A l'étranger, plusieurs initiatives mènent déjà des projets de recherche pour promouvoir l'interaction de toutes ces actions. Par exemple au Royaume-Uni, l'Education Endowment Foundation[3] (soutenu par le gouvernement anglais et identifié comme étant un « What Works Centre for Education ») finance des programmes de recherche incluant des interventions en classe tout en proposant des synthèses de la recherche pour fournir des kits pédagogiques[4] aux enseignants et les aider à intégrer les résultats significatifs dans les pratiques. Outre-Atlantique, le What Works Clearinghouse[5] a le même objectif : classer les méthodes d'apprentissage en fonction des preuves scientifiques qui les soutiennent puis les disséminer et les promouvoir auprès de la communauté enseignante. En France, nous avons peu d'équivalent à ce type d'instances alors que les enseignants sont à la recherche de ressources et de nouveaux outils pédagogiques pour améliorer leurs pratiques, tandis que les chercheurs souhaitent quant à eux développer des projets à grande échelle sur le terrain. La direction de l'éducation et des compétences de l'OCDE[6] (Organisation de Coopération et de Développement Economiques) ainsi que le CNESCO[7] (Conseil National d'Evaluation du Système sCOlaire) sont des initiatives du ministère de l'éducation qui remplissent une fonction d'évaluation du système éducatif et des politiques publiques liées à l'éducation, mais elles n'ont pas vocation à diffuser largement les résultats de ces évaluations auprès des enseignants, ni à promouvoir une éducation fondée sur des preuves.

En dehors des initiatives gouvernementales, de nombreux chercheurs ont également développé des initiatives pour informer et former le corps enseignant et les étudiants aux

[3] https://educationendowmentfoundation.org.uk
[4] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit
[5] https://ies.ed.gov/ncee/wwc/
[6] http://www.oecd.org/fr/education/
[7] http://www.cnesco.fr/fr/accueil/

méthodes d'apprentissages fondées sur des preuves sous forme de ressources et boites à outils. A titre d'exemple, l'équipe de The Learning Scientists[8] propose des fiches récapitulatives[9] des résultats de la recherche en psychologie à destination des enseignants, des élèves, mais aussi des parents ; une partie de leurs ressources ont d'ailleurs été adaptées en français. Autre exemple, John Hattie (2012) a quant à lui adapté sa « méta-analyse de méta-analyses » sur la réussite scolaire à destination pour les enseignants. Même si les recommandations proposées dans ce type de synthèse ne sont pas à prendre aux pieds de la lettre, la vulgarisation de ces travaux de recherche a au moins le mérite de mettre la lumière sur la diversité et la complexité des facteurs qui peuvent influencer la réussite des apprentissages scolaires (John Hattie propose cing grandes catégories de facteurs : l'élève lui-même, son environnement familiale, l'école, les programmes scolaires, et enfin l'enseignant). En France, la fondation la Main à la Pâte[10] publie des dossiers de synthèses[11] rédigés par le groupe de recherche *Sciences cognitives pour l'éducation*, à destination des enseignants ; de même le groupe Compas[12] assure une valorisation des résultats de la recherche pour les diffuser auprès du milieu de l'éducation et s'intéresse aux mutations de l'école avec l'utilisation de nouvelles technologies. Récemment, l'éducation nationale a créé un conseil scientifique ayant pour mission d'informer les décideurs politiques sur les résultats robustes de la recherche tout en proposant des projets de recherche en lien avec la réalité du terrain (Conférence internationale sur le rôle de l'expérimentation dans le domaine éducatif, Février 2018[13]).

Clôturons cette première partie consacrée à l'éducation fondée sur des preuves sur une représentation de la méthodologie qui devrait l'accompagner en réunissant les trois acteurs principaux de cette démarche. D'une part, les problèmes et questionnements issus du terrain en partie explorés par les enseignants de manière qualitative peuvent être problématisés pour être étudiés en laboratoire de manière contrôlée. Cela peut être l'effet d'un paramètre, l'impact d'un nouveau dispositif éducatif, ou le changement dans l'organisation de la classe sur la réussite des élèves. D'autre part, les chercheurs en sciences cognitives produisent des résultats comportementaux issus d'expérimentations sur l'apprentissage. Ces connaissances théoriques peuvent avoir des applications plus ou moins directes en lien avec les problématiques

[8] https://www.learningscientists.org
[9] https://www.learningscientists.org/downloadable-materials/
[10] https://www.fondation-lamap.org
[11] https://www.fondation-lamap.org/fr/page/23574/sciences-cognitives-et-education
[12] https://www.compas-etc.org
[13] https://www.reseau-canope.fr/conference-internationale-sur-le-role-de-lexperimentation-dans-le-domaine-educatif/programme1247.html#bandeauPtf

quotidiennes des enseignants. Le relais entre les deux se ferait par le biais d'instances gouvernementales, qui guideraient vers une démarche empirique et quantitative de production de données. Lorsqu'un niveau de preuve suffisamment élevé est atteint en laboratoire pour comprendre l'effet d'un paramètre précis sur l'apprentissage (en menant des réplications et méta-analyses notamment), une sortie des résultats hors du laboratoire est envisageable. Cela doit mener dans un premier temps à des expérimentations à grande échelle et dites "écologiques", afin de mesurer l'impact d'un dispositif éducatif en tenant compte de toute la complexité du terrain. La finalité à long terme étant de pouvoir formuler des recommandations pratiques précises pour apporter des pistes d'améliorations et d'innovation pédagogique pour l'enseignement. Cela passe à nouveau par le relais des décideurs politiques qui vont repenser la formation des enseignants (initiale et continue) pour intégrer de nouvelles pratiques et ainsi assurer le transfert des résultats probants sur le terrain (Figure 1).

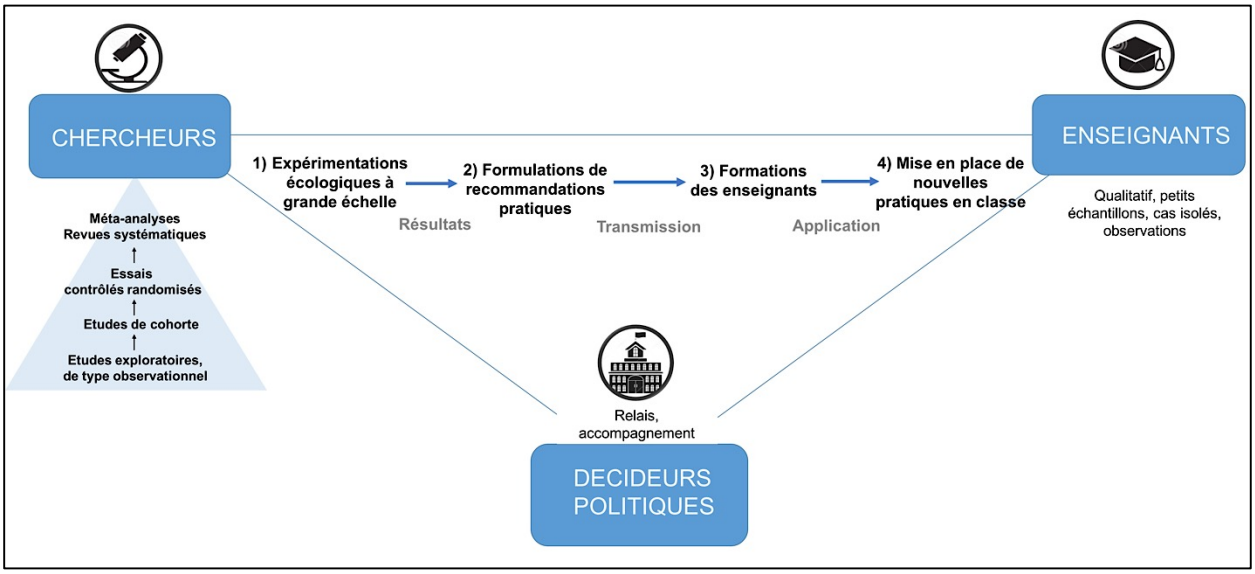


Figure 1. *Les trois acteurs impliqués dans une démarche d'éducation fondée sur des preuves. L'interaction entre les deux pôles « Chercheurs » et « Enseignants » peut se faire par le biais des « Décideurs politiques », qui pilotent leurs actions respectives et servent de médiateur. Elle peut aussi se faire directement dans le cadre de projets de recherche-action par exemple. Les contributions de chaque acteur doit pouvoir mener à des expérimentations à grande échelle dans des établissements pour produire des résultats robustes sur l'efficacité de certains outils sur la réussite des élèves. Cela peut ensuite mener à la formulation de recommandations qui seront transmises aux instances responsables de la formation des enseignants pour proposer l'utilisation des nouveaux outils pédagogiques dont on connait véritablement l'efficacité (« Transmission » et « Application » sur la figure).*

Mon travail de thèse se situe au cœur des actions menées par le laboratoire (pôle « Chercheurs » sur la Figure 1). L'objectif était de produire des résultats solides sur l'optimisation des apprentissages en suivant la méthodologie propre à la psychologie cognitive. Nous avons ainsi recueilli des données comportementales par le biais d'expérimentations contrôlés et randomisés dans le but de comparer les effets de différentes stratégies d'apprentissage sur la mémorisation à long terme. La partie suivante décrit les connaissances théoriques développées par la psychologie sur le mécanisme d'apprentissage.

1.1.3 Le processus d'apprentissage selon la psychologie cognitive

L'apprentissage est un processus dynamique qui correspond à un changement plus ou moins permanent de notre comportement et de notre cognition par le biais de l'acquisition de nouvelles informations qui sont stockées en mémoire (Korte, 2013; Mazur, 2015). La mémoire est intimement liée à ce processus puisqu'elle correspond à la persistance des informations acquises lors de nouvelles expériences dans le but d'une utilisation ultérieure. Même si les définitions de l'apprentissage ne sont pas mutuellement exclusives et dépendent des disciplines et des théories (Barron et al., 2015) ; en psychologie cognitive le processus d'apprentissage peut être représenté par trois étapes successives qui sont l'encodage, le stockage, et la récupération (Eysenck & Keane, 2010; McDermott & Roediger, 2018; Melton, 1963).

La première étape correspond au point d'entrée des informations nouvelles, ce processus initial mène à une représentation mentale de l'information dans la mémoire à court terme. Nous récupérons des informations caractéristiques du stimulus par notre système perceptif. Dans le cas d'un stimulus verbal, l'encodage se fait par les canaux visuels et auditifs, le système accède ensuite à la dimension sémantique de l'information dans notre lexique mental (Glenberg, 1979). Au-delà de l'encodage concernant des éléments relatifs à l'information d'intérêt (c'est-à-dire composants descriptifs), des composants dits contextuels indissociables de celle-ci sont également intégrés automatiquement. Ce sont des informations qui relèvent du contexte dans lequel a lieu l'encodage et peuvent être temporelles et spatiales mais elles correspondent également aux états émotionnels et cognitifs de l'apprenant (Bower, 1972; Godden & Baddeley, 1975). Ces informations participent à la création de la trace mnésique de l'information d'intérêt (Hintzman, 1988). Le stockage se matérialise quant à lui par la création d'une trace mnésique durable. Il suppose des changements à court et long terme dans les structures cérébrales (Dudai, 2004; Dudai, Karni, & Born, 2015). En effet, à la suite de l'encodage, la trace mnésique se trouve en mémoire de travail, elle est labile, fragile, et encore soumise aux effets d'interférence (Squire, 1986). Ces interférences peuvent être proactives (la formation de nouveaux souvenirs

serait alors perturbée par le traitement en cours d'informations précédemment encodées, Postman & Keppel, 1977) ou au contraire rétroactives (les informations les plus récentes perturberaient celles qui sont en cours de traitement, c'est-à-dire que les souvenirs nouvellement acquis et en cours de consolidation seraient vulnérables à l'arrivée d'un nouvel apprentissage, Tulving & Watkins, 1974). Si l'information n'est pas consolidée en mémoire à long terme, alors sa trace initiale est vouée le plus souvent à s'estomper ou tout simplement disparaître (McGaugh, 2000)**.** La consolidation renvoie à la période durant laquelle nous allons répéter plus ou moins consciemment en fonction de nos buts une information jusqu'à ce qu'elle soit suffisamment ancrée dans notre mémoire pour être retenue durablement (Frankland & Bontempi, 2005). Grâce à ce processus de consolidation, l'information devient de plus en plus stable et moins sujette aux interférences proactives ou rétroactives. D'un point de vue cérébral, la consolidation est un processus lent qui consiste en un transfert de l'information des régions de stockage des souvenirs de l'hippocampe vers le néocortex (Alvarez & Squire, 1994; Roediger, Dudai, & Fitzpatrick, 2007; McClelland, McNaughton, & O'Reilly, 1995). Alors que les étapes d'encodage et de stockage peuvent être relativement proches dans le temps, celles du stockage et de la récupération peuvent être temporellement éloignées (la dernière étape n'étant pas toujours prévisible). La récupération est donc l'aboutissement de tous les efforts précédents qui consiste en l'action de retrouver une information en mémoire dans un contexte particulier nécessitant son utilisation. En une fraction de seconde nous avons accès à l'information stockée précédemment. Si une information est déjà activée et donc en mémoire de travail, alors le processus de récupération sera d'autant plus facile mais l'information n'étant pas consolidée et donc labile, cet avantage a une limite de temps parce que l'activation s'estompe assez rapidement. En revanche, quand l'information est consolidée et doit être récupérée en mémoire à long terme, cela va mettre plus de temps et des stratégies de recherche devront certainement être mises en place mais l'information en question peut être récupérée à n'importe quel moment. Cette information transite de la mémoire à long terme à la mémoire de travail lors de son activation pour la rendre disponible et utilisable (Dudai et al., 2015).

L'école inclut principalement des connaissances verbales (avec une part non négligeable de connaissances procédurales telles que l'apprentissage de l'écriture, les travaux pratiques dans les cours de sciences, les cours d'éducation sportives) et le cadre de l'apprentissage est majoritairement explicite (même si une part d'implicite existe). L'encodage correspondrait à l'étape où les nouvelles connaissances sont délivrées, découvertes par les élèves pour la première fois (Figure 2). Elles sont entendues ou lues, sous forme de cours donné par un enseignant ou bien par le biais d'un manuel scolaire. L'étape de stockage suit généralement

celle de l'encodage et coïncide avec la mise en pratique des connaissances, c'est-à-dire leur mobilisation (via des exercices par exemple). Mais cette étape d'ancrage en mémoire se fait principalement par des séances de révisions en dehors de la salle de classe, et leur déroulement est à la charge de l'élève. Entre ces deux étapes, les nouvelles connaissances abordées en classe restent sujettes à l'oubli. Ebbinghaus est l'un des premiers à s'intéresser à l'apprentissage et a démontré que l'oubli suivait une courbe exponentielle qui pouvait être minimisée par certains facteurs au moment de l'encodage. En effet, la réussite de l'apprentissage explicite est conditionnée par plusieurs facteurs tels que :

- le niveau d'attention portée à l'information (Szpunar, Moulton, & Schacter, 2013; Unsworth, McMillan, Brewer, & Spillers, 2012; Yoncheva, Blau, Maurer, & McCandliss, 2010) **;**
- le niveau d'expertise de l'apprenant pour comprendre et intégrer les nouvelles informations et le niveau de difficulté des connaissances elles-mêmes (Kalyuga & Renkl, 2010; Webb & Chang, 2015);
- la profondeur de traitement de l'information (Craik & Tulving, 1975; Roediger, 2008),
- la qualité de l'environnement dans lequel l'information est présentée (Cheryan, Ziegler, Plaut, & Meltzoff, 2014) ;
- l'engagement ou l'effort mis en place pour comprendre (Auble, Franks, Soraci, Soraci, & Soraci, 1979; Zaromb, Karpicke, & Roediger, 2010) ;
- enfin la motivation et les affects (Deci & Ryan, 2000; Isen, 2001; Lin, McKeachie, & Kim, 2003; Unsworth & McMillan, 2013). La récupération se produit quant à elle le plus souvent dans le cadre d'une évaluation, d'un examen, d'une restitution des connaissances qui est réalisée de manière plus ou moins différée par rapport à l'encodage et qui dans la grande majorité des cas aboutit à une note qui juge la qualité de l'encodage et du stockage (Figure 2).

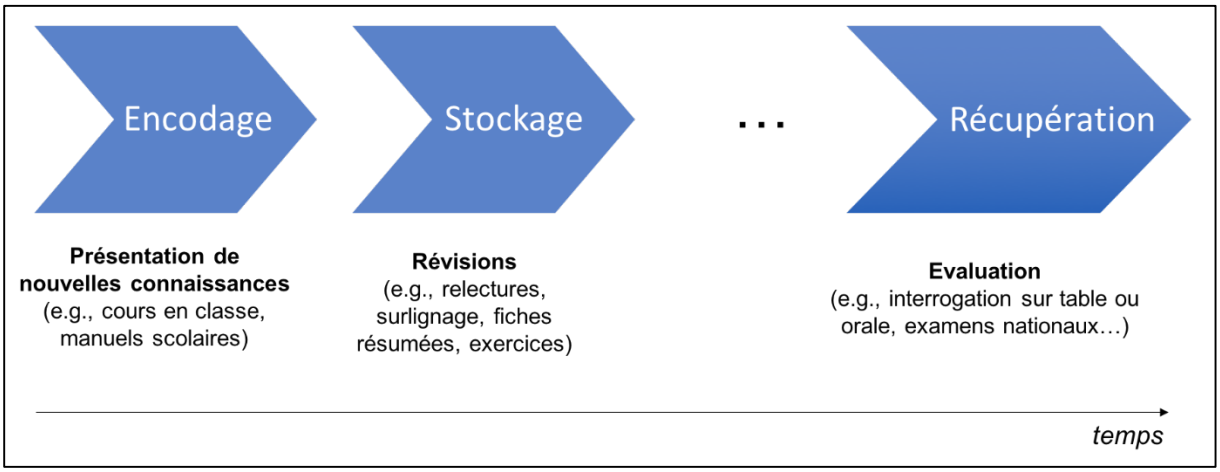


Figure 2. *Schéma représentant les trois étapes du processus d'apprentissage tel qu'il est décrit en psychologie cognitive. Ici, le parallèle est fait entre chacune de ses étapes et une activité de l'apprentissage scolaire tel qu'il est généralement envisagé.*

Mon projet de recherche est directement lié à la question suivante : comment optimiser les deux premières étapes du processus d'apprentissage (encodage et stockage) pour maximiser la dernière étape (récupération) ?

1.1.4 Les stratégies d'apprentissage efficaces

Dans cette conception de l'apprentissage à trois étapes dépendantes les unes des autres, les stratégies mises en place au moment de l'encodage et du stockage sont donc décisives pour garantir la récupération d'une information apprise. Un apprentissage est dit « réussi » si le niveau de performance est satisfaisant au moment de la récupération des connaissances apprises en cours. Cette réussite se mesure très souvent par l'obtention d'une bonne note à un examen mais aussi à la capacité à transférer les connaissances lors de situations nouvelles. Dans le contexte scolaire, le but de tout apprentissage est donc de garantir la consolidation des nouvelles connaissances afin de 1) ralentir leur oubli dans le temps tout en 2) augmentant la probabilité qu'elles soient disponibles et accessibles rapidement quand ces connaissances sont nécessaires ; et 3) de favoriser au maximum leur transfert dans d'autres contextes. Comment promouvoir ces objectifs fondamentaux ? Le succès d'un apprentissage est principalement dépendant de la réactivation régulière des traces mnésiques d'une information et dans des situations variées pour favoriser le transfert des connaissances (Leberman, McDonald, & Doyle, 2006). Répéter ces informations apparait donc comme une pratique réactivatrice idéale pour participer à la consolidation des connaissances. La question qui se pose ici est comment « bien » répéter efficacement l'information pour qu'elle soit récupérée de manière optimale le

moment venu ? (c'est-à-dire le jour de l'examen ou à tout moment dans la vie dans laquelle elle s'avère pertinente). Les travaux précurseurs de Robert Bjork (Bjork & Bjork, 2011; Bjork, 1988; Whitten & Bjork, 1977) sur les stratégies d'apprentissage efficaces ont donné naissance à tout un champ de la psychologie cognitive sur l'optimisation des processus qui rendent possible la rétention de nouvelles connaissances. Cette recherche est ancrée dans la démarche de production de résultats pour une éducation fondée sur des preuves. Cette section ne présente qu'une partie limitée de cette vaste littérature sur les stratégies d'apprentissage efficaces ; deux d'entre elles sont l'objet d'étude de cette thèse : l'effet de récupération en mémoire et l'effet d'espacement.

Il y a 10 ans, Karpicke, Butler, & Roediger (2009) ont réalisé un sondage auprès de 177 étudiants à l'Université Washington (St. Louis) sur les méthodes employées pour réviser des cours. Leurs résultats indiquent que 84% des étudiants interrogées ont listé la lecture répétée comme méthode utilisée, et 55% ont favorise cette méthode comme étant la plus efficace selon eux. S'entrainer avec des exercices, utiliser ce qu'on appelle des « flashcards » (un concept clé sur le recto d'une fiche, une information le concernant sur le verso), recopier les notes d'un cours, ou encore réviser en groupe étaient les autres méthodes d'apprentissage rapportées par moins de la moitié des étudiants interrogés. Une autre pratique particulièrement répandue consiste à réviser les mêmes connaissances de manière massée dans le temps, c'est à dire en une seule session sans aucune interruption ; et se combine très bien avec le bachotage, c'est à dire réviser à la dernière minute avant un examen. Même si cette enquête n'est pas récente et a été réalisée sur un échantillon restreint, la méthode qui consiste à relire plusieurs fois un même cours de manière massée la veille d'un examen semble être celle qui convient le plus aux étudiants, et notamment ceux qui seraient le plus en difficulté (Hartwig & Dunlosky, 2012; Rohrer & Pashler, 2010; Y. Weinstein, Lawrence, Tran, & Frye, 2013).

Nous voici face à un problème : la recherche sur l'optimisation des apprentissages a démontré que ces stratégies ne permettent pas de consolider durablement les connaissances à long terme et procurent un sentiment de maitrise qui donne l'illusion de savoir (Bembenutty, 2009). Il s'avère que les étudiants ont tendance à penser que les méthodes qui demandent peu d'effort, qui rendent l'apprentissage facile, et qui donnent une impression de stockage immédiat en mémoire sont celles qui vont effectivement leur permettre d'avoir de bonnes performances lors d'un examen (Bjork & Bjork, 2011). Cependant, des décennies de travaux sur le sujet ont montré que les stratégies d'apprentissage qui promeuvent une rétention des connaissances à long terme sont en général celles qui engendrent un certain niveau de difficulté durant les étapes d'encodage et de stockage (Brown et al., 2014; Hattie, 2008). Malheureusement, ce sont aussi

celles qui vont avoir tendance à décourager les apprenants et leur à donner la sensation qu'ils ne savent pas.

Dans une étude devenue une référence sur le sujet, Dunlosky et collaborateurs (2013) ont recensé un total de 10 méthodes d'apprentissage telles que surligner des passages importants, faire des fiches de synthèse, activer l'imagerie mentale, ou encore utiliser l'alternance des contenus, ainsi que certaines méthodes citées plus haut dans l'enquête de Karpicke, Butler, & Roediger (2009). Les auteurs ont ainsi réalisé une synthèse des études ayant démontré des bénéfices de chacune de ces méthodes pour la mémorisation à long terme en fonction i) des conditions d'apprentissage, ii) du contenu à étudier, iii) des caractéristiques des apprenants, et iv) des outils utilisés pour mesurer les performances finales. L'objectif principal de ce travail était d'attribuer un niveau d'efficacité à chacune de ces méthodes en se basant sur les résultats issus de la littérature, et donc en fonction du niveau de « preuves » existantes en faveur de chaque méthode d'apprentissage. La conclusion majeure de leur synthèse est instructive : les méthodes les plus utilisées sont les moins efficaces d'après leurs critères, alors que les moins utilisées seraient au contraire les plus bénéfiques pour maximiser l'apprentissage. Le Tableau 1, reprise de cet article, présente les niveaux d'efficacité attribués à chacune des 10 méthodes d'apprentissage sur la base de différents critères et sur la possibilité de généraliser leur utilisation. Par exemple, une note positive (P) sur les évaluations finales indique qu'il y a un niveau de preuves suffisamment élevé pour conclure qu'une méthode est réellement bénéfique pour maximiser les performances des étudiants, et ce pour une grande variété de types d'évaluations et d'intervalle de rétention (plus ou moins différé). A l'inverse, une note négative (N) indique que le nombre de résultats scientifiques ayant testé l'efficacité de la méthode en question est suffisant pour démontrer qu'elle n'est pas bénéfique pour la rétention. Par exemple : le surlignage n'a pas d'effet bénéfique sur les performances finales des étudiants, alors qu'elle est pourtant très populaire (ex, Bell & Limber, 2010; Lonka et al., 1994). Les résultats mitigés des expérimentations ayant mesuré ses effets sont plus plutôt dissuasifs (ex, Todd and Kessler, 1971 ; Peterson, 1992). La note intermédiaire (Q pour *qualified*) indique qu'il y a des effets bénéfiques mais seulement dans certaines circonstances, ce qui empêche de généraliser cette méthode. Enfin, une notre qualifié d'insuffisante (I) indique que le niveau de preuves n'est pas assez conséquent pour conclure sur l'efficacité d'une méthode, et que des recherches doivent être menées pour avoir plus de certitude. La Figure 3 résume en partie ces résultats en classant les méthodes selon leur niveau d'efficacité et le degré de preuves disponibles sur leur efficacité pour la rétention.

Tableau 1. *Table extraite de Dunlosky et collaborateurs. (2013). Evaluation de l'utilité et du degré de généralisation de chaque stratégies d'apprentissage (Tableau 4, p45).*

Technique	Utility	Learners	Materials	Criterion tasks	Issues for implementation	Educational contexts
Elaborative interrogation	Moderate	P-I	P	I	P	I
Self-explanation	Moderate	P-I	P	P-I	Q	I
Summarization	Low	Q	P-I	Q	Q	I
Highlighting	Low	Q	Q	N	P	N
The keyword mnemonic	Low	Q	Q	Q-I	Q	Q-I
Imagery use for text learning	Low	Q	Q	Q-I	P	I
Rereading	Low	I	P	Q-I	P	I
Practice testing	High	P-I	P	P	P	P
Distributed practice	High	P-I	P	P-I	P	P-I
Interleaved practice	Moderate	I	Q	P-I	P	P-I

Note: A positive (P) rating indicates that available evidence demonstrates efficacy of a learning technique with respect to a given variable or issue. A negative (N) rating indicates that a technique is largely ineffective for a given variable. A qualified (Q) rating indicates that the technique yielded positive effects under some conditions (or in some groups) but not others. An insufficient (I) rating indicates that there is insufficient evidence to support a definitive assessment for one or more factors for a given variable or issue.

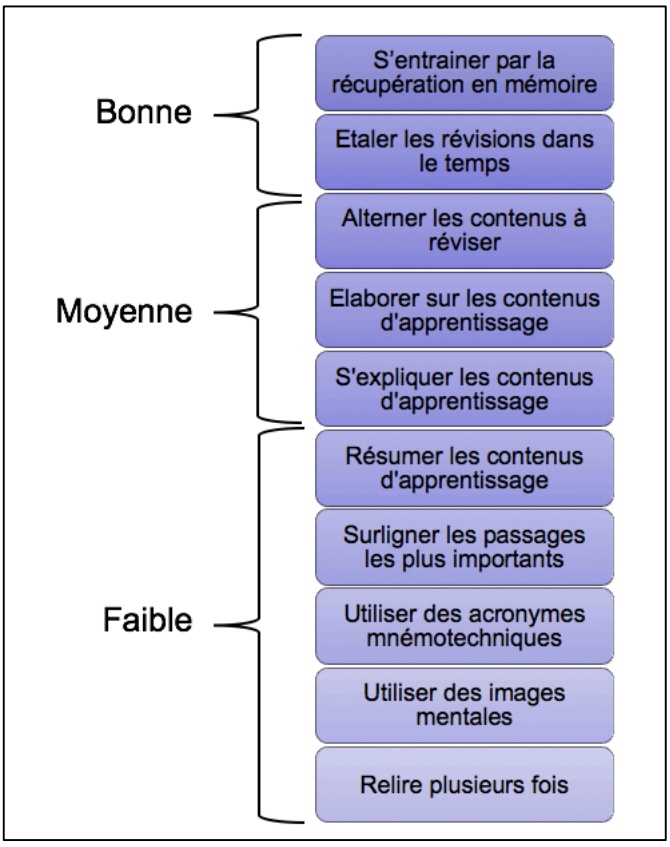


Figure 3*. Schéma récapitulatif des résultats principaux de la synthèse de Dunlosky et collaborateurs (2013). Les différentes stratégies d'apprentissage sont regroupées en fonction de leur efficacité (bonne, moyenne, ou faible) pour optimiser la rétention à long terme de nouvelles connaissances, et ce d'après les résultats fournis par la recherche en psychologie cognitives.*

Que faut-il retenir de ces deux articles ? Ils illustrent bien le décalage entre les méthodes principalement mises en place par les apprenants et ce que démontre la recherche sur l'efficacité réelle de ces mêmes méthodes. Cela montre également qu'il est difficile pour les étudiants de juger à quel point les méthodes mises en place durant leurs révisions optimisent véritablement la récupération ultérieure au moment d'un examen. Prendre les bonnes décisions quant à la manière de réviser est une capacité modulée par les compétences métacognitives (Finley, Tullis, & Benjamin, 2010; Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003). Si nos jugements sur l'efficacité de ces compétences sont biaisés, les décisions prises sur les méthodes pour réviser vont mener à un apprentissage a priori sous optimal (Atkinson, 1972; Kornell & Bjork, 2008; Tullis & Benjamin, 2011).

Dans le cadre ce projet de recherche, je me suis principalement intéressée à deux stratégies d'apprentissage dont les bénéfices sur la consolidation de connaissances nouvelles ont été démontrés de manière robustes dans divers contextes (différentes populations et différents contenus de formation) mais qui sont justement peu utilisées par les étudiants et enseignants pour apprendre de nouvelles connaissances. La première stratégie est **l'effort de récupération en mémoire** ; la seconde est **l'espacement des séances d'apprentissage dans le temps** (respectivement *retrieval practice* et *spaced* ou *distributed learning* en anglais). Dans la synthèse de Dunlosky et collaborateurs (2013) ces pratiques (« S'entrainer par la récupération en mémoire » et « Etaler les révisions dans le temps », Figure 3) sont celles qui auraient la plus grande efficacité, mais également un important niveau de preuves issues de la recherche en leur faveur. Les sous parties suivantes présentent une description succincte de ces deux stratégies d'apprentissage efficaces ainsi que leurs bénéfices et mécanismes cognitifs sous-jacents. Pour une description plus exhaustive, le Chapitre 2 propose une revue systématique de la littérature sur ces effets (p30-66). D'autres stratégies d'apprentissage reconnues comme efficaces existent en dehors de celles-ci mais ne font pas l'objet de ce projet de recherche. Il existe notamment l'alternance (*interleaving effect* en anglais) du type de compétences sur lesquelles nous nous entraînons (pratiquer différents types de problèmes mathématiques en les alternant plutôt que de les apprendre par bloc ; Birnbaum, Kornell, Bjork, & Bjork, 2013; Brunmair & Richter, 2019) ; ou encore le double codage (combiner un support verbal et un support visuel pour représenter une même connaissance à apprendre, Mayer & Anderson, 1992; Paivio, 2007).

1.1.4.1 L'effet de récupération en mémoire (testing effect en anglais)

De nombreuses études réalisées en laboratoire et en classe ont montré l'importance de l'entrainement par récupération en mémoire comme processus actif pour encoder et stocker de nouvelles connaissances et les consolider en mémoire à long terme (Adesope, Trevisan, & Sundararajan, 2017; Hattie, 2008; Rowland, 2014). Cette pratique d'apprentissage dite « active » a été comparée à la méthode classique de révision par relectures successives, considérée comme « passive ». Le design expérimental dans lequel ce *testing effect* (tel qu'il est appelé en anglais) est mesuré consiste en une première exposition aux connaissances sous forme de lecture ou bien plus rarement sous la forme d'un cours qui est délivré oralement. Puis les apprenants sont invités à réviser le contenu pour le consolider en mémoire dans le but de passer un examen de manière plus ou moins différée dans le temps. C'est cette phase de révision qui est manipulée expérimentalement : elle prend la forme d'un entrainement par récupération en mémoire des connaissances présentée précédemment, ou bien elle consiste à les relire. La réussite de la phase de révision est mesurée par la quantité de connaissances correctement rappelées lors d'un test final de mémorisation. Une abondante littérature a donc démontré qu'utiliser des stratégies de type rappel libre ou rappel indicé (récupérer à l'aide d'indices les éléments mémorisés sous forme de flashcards ou de QCM) favorisent un meilleur stockage des nouvelles connaissances tout en minimisant leur oubli avant une prochaine récupération (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger & Karpicke, 2006; Zaromb & Roediger, 2010).

Cette méthode peut être facilement mise en place par les enseignants sous forme d'évaluation formative pour préparer aux évaluations sommatives. En 2007, Pashler et al. (2007) avaient déjà publié un rapport sur les bénéfices de cette méthode d'apprentissage avec des recommandations pratiques à destination des enseignants pour les intégrer en classe et transmettre les bonnes méthodes de travail à leurs élèves. Depuis ce rapport, les recommandations s'adressant aux enseignants et formateurs pour intégrer l'entrainement par récupération dans leur cours ne se dénombrent plus, même en France (à titre d'exemple : Dehaene (2018), *Cerveau : les quatre piliers de l'apprentissage*[14], le MOOC « Apprendre et enseigner avec les sciences cognitives[15] », et le livre « Mets-toi ça dans la tête » traduction française du livre de (Brown et al., 2014). La promotion de cette méthode de révision se fait

[14] https://www.lexpress.fr/actualite/societe/les-quatre-piliers-de-l-apprentissage_2031330.html
[15] https://www.fun-mooc.fr/courses/course-v1:drhatform+124001+session02/about

également auprès des étudiants par plusieurs chercheurs en psychologie cognitive (Dunlosky et al., 2013; Putnam, Sungkhasettee, & Roediger, 2016) et à travers des formations en ligne (à titre d'exemple le MOOC « Learning How To Learn » proposé par l'Université McMaster et l'Université de Californie à San Diego[16]). Dans leur enquête, Karpicke, Butler, & Roediger (2009) confirmaient la très faible utilisation de l'apprentissage par récupération en mémoire puisque seulement 11 % des étudiants interrogés listaient cette méthode parmi celles utilisées dans leurs révisions. Et la grande majorité de ces étudiants révélaient utiliser cette méthode d'apprentissage dans le but de s'autoévaluer plutôt que pour réviser, phénomène déjà retrouvé dans une enquête précédente du même type (Kornell & Bjork, 2007).

D'un point de vue cognitif, lorsque nous devons répondre à une question, nous devons produire un effort pour rechercher l'information pertinente afin d'y répondre correctement : c'est ce qu'on appelle l'effort de récupération. Par exemple, répondre à un quiz demande à l'apprenant de faire un effort de recherche de l'information cible en mémoire pour générer une réponse (Carpenter & DeLosh, 2006; Glover, 1989; Kang, McDermott, & Roediger III, 2007). Mis en place tout au long de l'apprentissage, les tests de récupération en mémoire vont aussi permettre de renforcer les « chemins » menant à l'information nouvelle, en activant notamment des connaissances préalables fortement connectées (Antony et al. 2017). Cette stratégie va également permettre de multiplier le nombre de ces « chemins » possibles en générant naturellement de nouveaux « indices » qui pourront ultérieurement servir d'amorce pour mobiliser une connaissance cible (Carpenter, 2009; Carpenter & DeLosh, 2006; Rowland & DeLosh, 2014). Une autre explication qui a été avancée sur les bénéfices cognitifs est la notion de similitude entre processus cérébraux mis en place lors de l'encodage, et ceux mis en place lors de la récupération dans le cadre d'un examen (*TAP theory* pour *Transfer Appropriate Processing* en anglais; Adesope, Trevisan, & Sundararajan, 2017; Johnson & Mayer, 2009).

En dehors de ces effets directs de l'apprentissage par récupération en mémoire sur la rétention à long terme, de nombreux bénéfices indirects ont été démontrés dans la littérature(Benjamin & Pashler, 2015). Par exemple, répondre à des quiz fréquemment maintient les élèves constamment engagés et attentifs à l'étude du contenu (Karpicke, 2009). En effet, quelques études ont démontré les effets bénéfiques des quiz sur la fréquence du « *mind wandering* »[17] (Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013).

[16] https://www.coursera.org/learn/learning-how-to-learn
[17] Ou « vagabondage de l'esprit » en français. Il se rapporte au fait de penser à autre chose que ce que l'on est en train de faire à un instant donné.

S’autoévaluer régulièrement encourage les apprenants à étudier en permanence tout au long du l’année et les oblige à revenir sur des connaissances qui ont été acquises il y a longtemps. Des quiz fréquents sont aussi efficaces pour la métacognition. En répondant régulièrement à des quiz, l'apprenant vérifie ce qu’il a vraiment compris grâce à des retours d'information correctifs (Butler & Roediger, 2008). Des explications qui suivent le quiz lui permettent effectivement de réajuster ses connaissances si besoin, et de revenir sur les points clés du cours qui ne sont pas maitrisés (Raaijmakers, Baars, Paas, van Merriënboer, & van Gog, 2019). En permettant à l’apprenant de faire le point sur l'état de ses connaissances, l’entrainement par récupération en mémoire suivi de feedback contribue ainsi à éviter les illusions de maitrise, le sentiment de « savoir » et donc éviter d’être trop confiant sur ses propres capacités à réussir lors d’un examen (Bjork, Dunlosky, & Kornell, 2013; Fernandez & Jamet, 2017; Miller & Geraci, 2014). Enfin, des études ont démontré que se tester sur des connaissances immédiatement après leur encodage permettraient d’améliorer l’encodage de nouvelles connaissances présentées ultérieurement. Cet effet appelé *foward testing effect* en anglais (Arnold & McDermott, 2013; Cho, Neely, Crocco, & Vitrano, 2017; Yang, Potts, & Shanks, 2018) a été répliqué plusieurs fois dans le contexte d’un apprentissage par récupération en mémoire en alternance avec la lecture de chapitres issus d’un même cours (*interpolated testing* en anglais).

Au-delà des effets positifs de cette stratégie d’apprentissage sur les compétences métacognitives des apprenants, se soumettre à des tests régulièrement peut aider à réduire l’anxiété face aux évaluations (Agarwal, D’Antonio, Roediger, McDermott, & McDaniel, 2014; A. M. Smith, Floerke, & Thomas, 2016) tout en maintenant un certain niveau d’attention et de motivation chez les étudiants à condition d’intégrer suffisamment de feedback (Healy, Jones, Lalchandani, & Tack, 2017; Kole, Healy, & Bourne, 2008). Enfin, du côté des enseignants, utiliser l’apprentissage par récupération en mémoire en classe est un excellent format d'évaluation formative puisqu’ils peuvent les renseigner sur les points du cours qui posent des difficultés et ces résultats peuvent les guider dans la préparation des cours suivants.

La Figure 4 extraite de l’article de Weinstein, Madan, and Sumeracki (2018) récapitule sous forme de schéma conceptuel une partie des bénéfices directs et indirects de l’apprentissage par les tests.

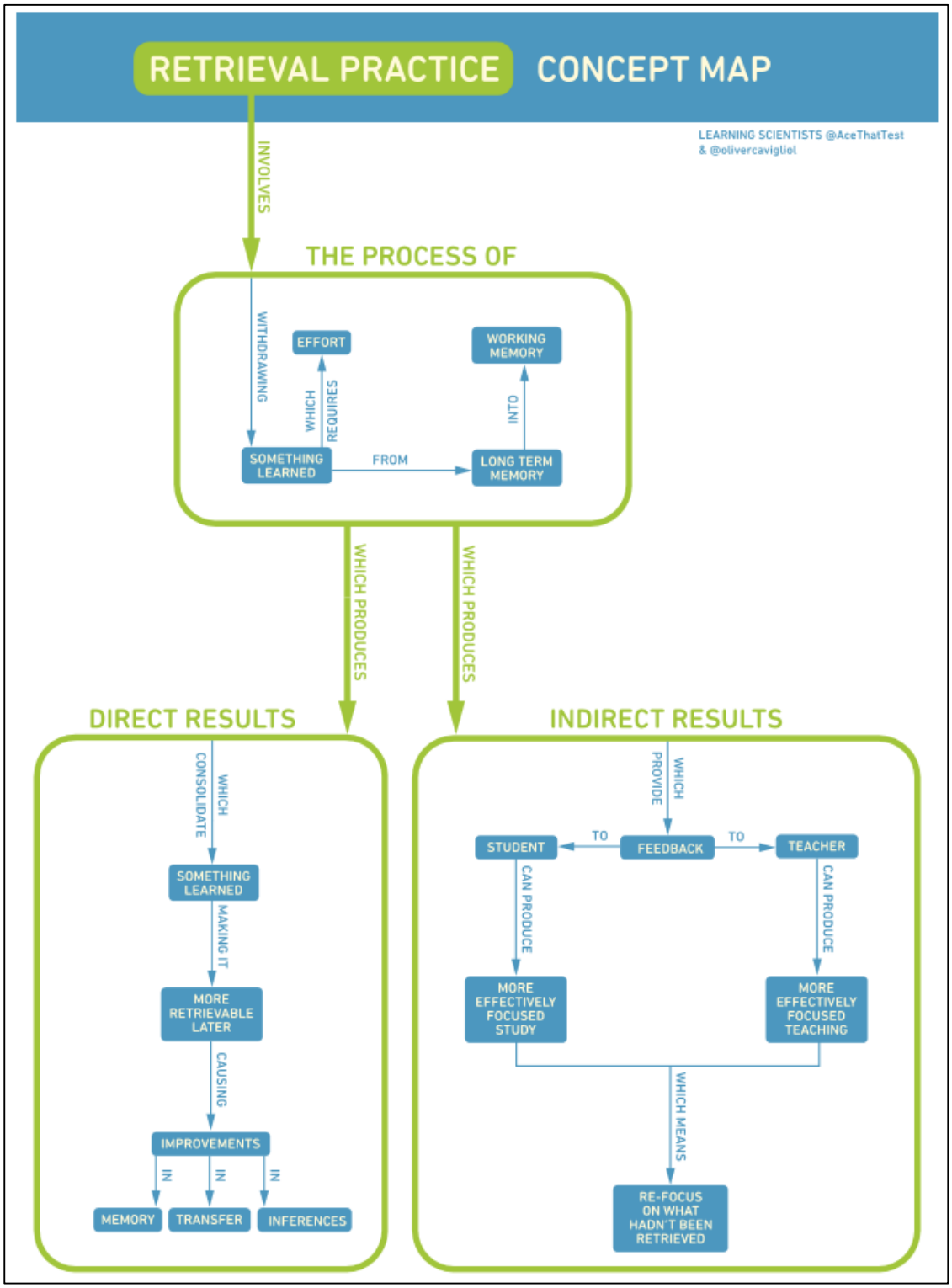


Figure 4. *Par Weinstein, Madan, and Sumeracki (2018, p 4). Cette figure illustre le processus de l'apprentissage par récupération en mémoire et les bénéfices qui en résultent. L'apprentissage par récupération en mémoire implique de récupérer une information précédemment apprise en mémoire à long terme, puis elle est transférée en mémoire de travail, ce qui nécessite un certain effort. Cela engendre des bénéfices directs via la consolidation de*

l'information apprise, qui sera d'autant plus facile à récupérer plus tard et permet une amélioration de la capacité à mémoriser, transférer, et réaliser des inférences. L'apprentissage par récupération en mémoire produit également des effets positifs indirects tels que la production de retours informatifs, à la fois aux étudiants et aux enseignants, ce qui en retour améliore la manière de réviser et d'enseigner, avec une attention particulière donnée aux informations non correctement récupérées. Copyright note : cette figure est extraite d'un article du blog The Learning Scientists rédigé par Megan Smith, Yana Weinstein, & Oliver Caviglioli[18].

1.1.4.2 L'effet d'espacement (spacing effect en anglais)

La répartition dite « massée » consiste à délivrer et à apprendre l'information en une fois sans interruption entre les expositions répétées d'une même connaissance. Elle est la pratique majoritairement utilisée aussi bien par les enseignants que par les étudiants. Une répartition des apprentissages qui est à l'inverse peu utilisée est l'espacement dans le temps. Pourtant, depuis les travaux pionniers d'Ebbinghaus sur l'oubli (Ebbinghaus,1885/2013), des centaines d'études ont démontré les bénéfices de la répartition espacée pour consolider sur le long terme et mieux gérer l'oubli des informations apprises au fur et à mesure que l'on se rapproche d'un examen final (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Delaney, Verkoeijen, & Spirgel, 2010; Smith & Scarf, 2017). Le délai entre deux répétitions (*interstudy interval* en anglais) peut varier de quelques secondes à plusieurs semaines selon les études et ce délai peut être vide de contenu à apprendre (espacement lié au temps) ou bien il peut comprendre la répétition d'autres informations (espacement lié aux items à apprendre). Ce bénéfice appelée *spacing effect* en anglais a fait l'objet de plusieurs méta-analyses (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003). Le design expérimental typique dans lequel cet effet est étudié est très similaire à celui utilisé pour mesurer le *testing effect.* Après une première exposition au contenu à apprendre, la phase d'entrainement consiste en une ou plusieurs séances de révision massée(s) ou espacée(s) dans le temps par rapport à la phase initiale. Ces révisions peuvent prendre la forme de relectures répétées ou bien consister en des exercices de récupération en mémoire selon les protocoles Après un certain intervalle de temps, la rétention en mémoire des informations apprises est mesurée dans la phase d'évaluation (Figure 5). La littérature s'est principalement intéressée à deux types de répartitions ou planning d'espacement des sessions

[18] http://www.learningscientists.org/blog/2016/4/1-1

de révisions jusqu'à l'échéance fixée pour un test final : le planning expansif, qui consiste à accroitre progressivement l'intervalle de temps entre deux sessions de révisions ; et le planning uniforme qui consiste à maintenir le même intervalle entre chaque révision.

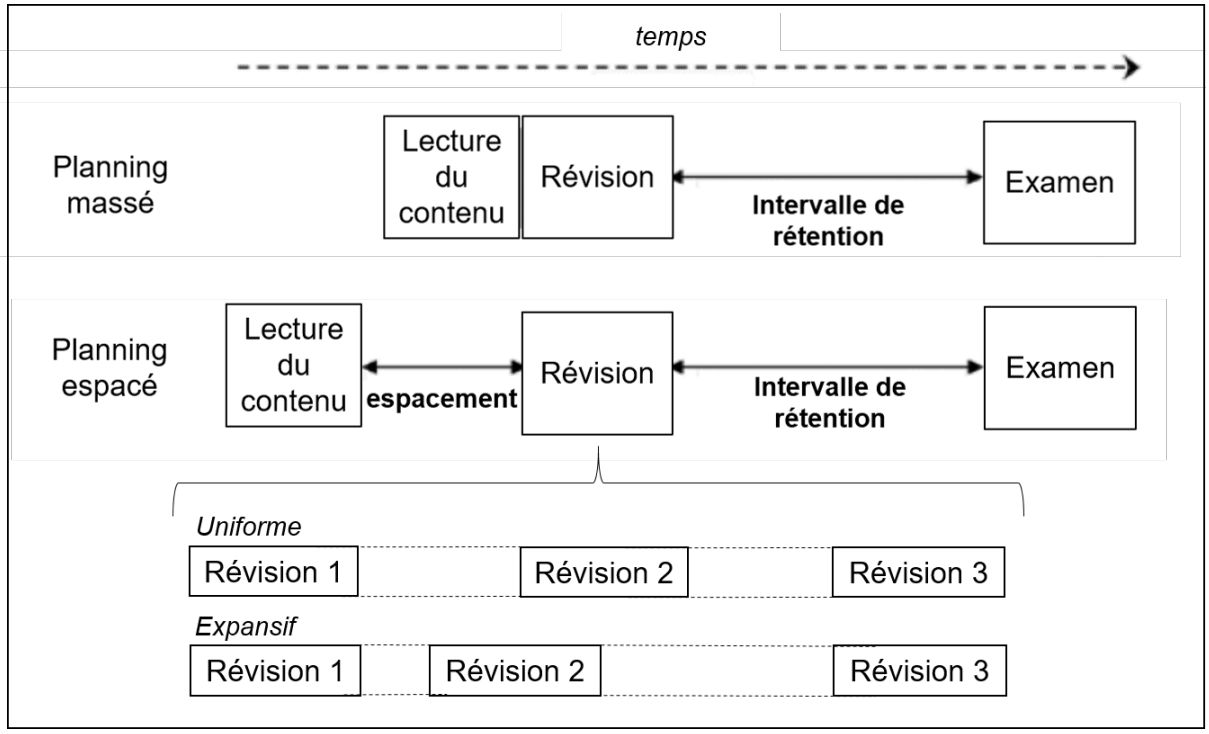


***Figure 5**. Schéma expérimental typique pour étudier l'effet d'espacement. La phase initiale de présentation du contenu d'apprentissage (lecture) est suivie d'une phase de révision qui peut avoir lieu immédiatement après (planning massé) ou bien de manière différée (planning espacé). Après un certain intervalle de temps, la réussite de la phase de révision est mesurée par un examen de mémorisation. Si plusieurs séances de révisions sont programmées, leur répartition peut être espacée de manière uniforme, ou bien expansive jusqu'à la date de l'examen. Ce schéma est inspiré de celui proposé par Kang (2016, p13).*

Malgré les effets positifs notables de l'espacement des apprentissages, les apprenants privilégient la pratique massée, notamment parce qu'elle procure un sentiment de tout avoir en mémoire comme lors de la relecture (Kornell & Bjork, 2007). D'ailleurs, dans le cas de révisions avec exercices de récupération en mémoire, les étudiants ont tendance à prédire de meilleures performances à un examen après un planning massé par rapport à un planning espacé. En effet, les bonnes performances obtenues durant la session d'apprentissage massée les induisent en erreur sur la réussite future (McCabe, 2011; Metcalfe & Finn, 2008; Metcalfe & Xu, 2016; Roediger & McDermott, 2018). Le fait que l'apprentissage espacé induise plus d'erreur durant les révisions, et donc un sentiment d'oubli plus important, joue en défaveur de

l'espacement car les étudiants ont du mal à en apprécier les bénéfices pour planifier leurs révisions. Du côté des enseignants il est aussi difficile de promouvoir l'utilisation de l'espacement des sessions de cours sur un même concept car cela oblige à planifier à l'avance les répétitions. Etant donnée la quantité très importante de contenus à enseigner dans les programmes scolaires, les aspects logistiques deviennent prioritaires (Barzagar Nazari & Ebersbach, 2018; Delaney et al., 2010). Pourtant, l'utilisation plus fréquente de tests cumulatifs pourrait induire une forme d'espacement dans la révisions des connaissances (Lindsey, Shroyer, Pashler, & Mozer, 2014). A titre d'exemple, des chercheurs ont proposé un algorithme très simplifié d'espacement sous Excel pour permettre aux enseignants d'étaler dans le temps la répétition des connaissances à l'échelle d'un trimestre ou même de l'année scolaire (Weinstein-Jones, F., & Weinstein, Y., 2017)[19].

Plusieurs théories ont été proposées pour justifier les bénéfices de l'espacement des apprentissages (Smolen, Zhang, & Byrne, 2016; Toppino & Gerbier, 2014). Une des explications du bénéfice de l'espacement rejoint celle proposée pour expliquer le bénéfice de l'entrainement par récupération en mémoire : la notion d'effort de récupération. L'apprentissage espacée dans le temps « permet » d'oublier les connaissances apprises entre deux sessions de révision ce qui augmente en retour l'effort cognitif pour se rappeler les connaissances lors de la deuxième révision, cet effort va lui même être bénéfique pour consolider les connaissances en renforçant la trace en mémoire. Laisser s'écouler un certain temps entre deux répétitions impose de faire cet effort de récupération en mémoire à long terme, ce qui n'est pas le cas lors d'un apprentissage massé. En effet, à chaque nouvelle session de révision espacée de la précédente, il faut récupérer et réactiver la trace en mémoire à long terme, ce qui renforce la trace au fur et à mesure des récupérations. Dans la répétition massée, la trace en mémoire n'a pas à être réactivée puisqu'elle est toujours présente en mémoire de travail, ce qui explique la très bonne performance de récupération en mémoire lors de la session d'apprentissage (*study-phase retrieval theory* en anglais; Braun & Rubin, 1998). De plus, l'espacement permettrait la mise en place de processus responsables de l'ancrage mémoriel à l'échelle du neurone. La consolidation permet de stabiliser progressivement la nouvelle information dans nos réseaux neuronaux (Kramár et al., 2012). Elle fait appel à des mécanismes biochimiques et moléculaires qui prennent du temps : plusieurs heures, voire plusieurs jours peuvent être nécessaires avant d'aboutir à la formation d'une trace mémorielle facilement ré-activable lors du rappel (Litman & Davachi, 2008). En s'entraînant de manière espacée dans le

[19] Topic spacing spreadsheet for teachers [Excel macro]. Zenodo. http://doi.org/10.5281/zenodo.573764.

temps, la consolidation est facilitée dans notre mémoire à long terme. En revanche, en s'entraînant de manière massée, ce processus neuronal n'a pas le temps de stabiliser la trace qui reste sensible à l'oubli et peut donc être rapidement perdue. Une autre explication intéressante sur le bénéfice de l'espacement a été avancée : l'amplitude de la fluctuation du contexte d'apprentissage entre deux sessions de révisions. Comme nous l'avons vu dans la section « Le processus d'apprentissage selon la psychologie cognitive», en apprenant des connaissances particulières nous encodons et stockons des caractéristiques du contexte d'apprentissage. Ces éléments participent également à l'étape de récupération. Ainsi, il semblerait que plus l'espacement entre deux sessions de révision est important et plus ces éléments contextuels seront divers et nombreux pour participer favorablement au processus de récupération des connaissances lors d'un examen *differential storage* en anglais ; Bjork & Allen, 1970; Emilie Gerbier, 2011). Enfin, dans le cas d'espacement de l'ordre d'un jour, le bénéfice du sommeil a été avancé dans le processus de consolidation des connaissances, mais cette piste reste peu étudiée à ce jour pour en tirer des conclusions claires (Frankland & Bontempi, 2005; Mazza et al., 2016). Tout comme l'usage de l'apprentissage par récupération en mémoire, la promotion de cette méthode de planification des révisions ne cesse d'être faite auprès des enseignants et des étudiants par le biais de plusieurs canaux (Kang, 2016; Weinstein, Madan, & Sumeracki, 2018; Vidéos de vulgarisation par la chaîne Osmose[20]).

Que doit-on retenir de ces deux effets de récupération en mémoire et d'espacement ? Bjork (1994) parle de « difficultés désirables » pour l'apprenant puisque comme nous l'avons vu ces stratégies d'apprentissage sont efficaces en rendant le processus d'apprentissage plus difficile que ne le rendent les stratégies d'apprentissage plus passives telles que la relecture. Ce type d'apprentissage plus facile a tendance à berner les apprenants puisqu'ils se basent justement sur cette facilité de stockage de l'information pour prédire leur future réussite (Koriat & Ma'ayan, 2005; Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Autrement dit, ils ont tendance à considérer que ce qui est facilement appris sera aussi facilement récupéré au moment de l'examen (Koriat, 2008; Miele & Molden, 2010). Comme nous allons le voir dans les chapitres suivants, malgré l'intérêt croissant pour l'étude des mécanismes des effets de test et d'espacement, plusieurs questions restent encore peu explorées sur la manière de les intégrer dans les séances de révisions. Les Chapitres 3 et 4 s'intéressent à la manière d'optimiser l'apprentissage par récupération en mémoire et les Chapitres 2 et 5 s'intéressent à la combinaison de ces deux pratiques.

[20] https://www.youtube.com/watch?v=cVf38y07cfk

1.2 Questions de recherche de la thèse

1.2.1 Cadre du projet et objectifs généraux

La recherche présentée dans ce manuscrit s'insère au sein d'un projet plus large appelé « Parcours connectés »[21] et financé par la Caisse des Dépôts et Consignations dans le cadre du programme e-FRAN[22] (pour Espaces de Formation, de Recherche et d'Animation Numérique). Cet appel à projets a été lancé par le *Programme d'Investissement d'Avenir*. Ce projet réunit un consortium de quatre laboratoires, l'association SynLab (https://syn-lab.fr/lassociation/) qui est porteuse du projet et pilote les activités du consortium, et enfin la start-up Didask qui a mis au point la plate-forme d'enseignement numérique du même nom (http://www.didask.com). Deux autres doctorants sont également impliqués dans Parcours Connectés (en informatique et en psychologie cognitive). Un premier axe du projet se concentre sur des expérimentations à l'ESPE de l'académie de Créteil à la fois sur la partie formation initiale des étudiants qui se préparent au métier de professeurs des écoles, mais aussi sur l'accompagnement des titulaires qui ont tout juste obtenu leur diplôme. Cette thèse concerne le second axe de recherche de Parcours connectés : améliorer les apprentissages en exploitant et en paramétrant au mieux les effets de récupération en mémoire en utilisant Didask comme terrain d'expérimentation. Cette plateforme a un fonctionnement particulier basé sur un contenu pédagogique fractionné en unités élémentaires pour optimiser leur assimilation et sur l'utilisation systématique de quiz pour évaluer les acquis mais surtout comme outils d'apprentissage. Les quiz sont programmés et répétés dans le temps pour optimiser la consolidation en mémoire sur le long terme. Didask a donc cherché à incorporer les effets de récupération en mémoire dans le cœur de la plateforme. Mon projet de thèse est ainsi un projet de collaboration à l'interface entre recherche fondamentale et appliquée en psychologie cognitive en lien avec l'éducation.

La spécificité de ce projet réside à la fois dans ses objectifs qui vont au-delà de considérations purement scientifiques mais qui sont aussi pratiques ; et dans la méthodologie utilisée visant à recréer des situations d'apprentissage les plus écologiques et réalistes possibles tout en assurant une rigueur expérimentale indispensable lors de mesures comportementales.

L'objectif principal de ma thèse était de comprendre les paramétrages permettant d'optimiser les apprentissages avec l'effet de récupération en mémoire dans le cadre

[21] https://www.caissedesdepots.fr/sites/default/files/medias/projet_18_dp_cdc.pdf
[22] https://www.caissedesdepots.fr/espaces-de-formation-de-recherche-et-danimation-numeriques-e-fran

d'expérimentations comportementales contrôlées, randomisées, et en ligne. Pour cela, nous avons utilisé la plate-forme numérique Didask décrite précédemment pour contrôler les modalités exactes de présentation et de test des contenus pédagogiques afin de comprendre comment maximiser les bénéfices des quiz d'apprentissage. En lien avec cet objectif principal, nous avions des objectifs secondaires tels que promouvoir les méthodes d'apprentissage qui fonctionnent dans une plateforme d'enseignement numérique, apporter des résultats dans une démarche d'éducation fondée sur des preuves empiriques, et apporter de nouvelles pistes de développement de la plateforme Didask.

1.2.2 Les questions précises en lien l'effet de récupération en mémoire

Une revue de littérature lors des premiers mois de thèse ont permis de cibler des questions peu étudiées sur l'optimisation de l'effet de récupération en mémoire. En effet, de nombreux résultats moins bien établis nécessitaient des réplications, et, s'ils étaient confirmés, nécessitaient des études complémentaires afin de mieux comprendre leur nature et leurs limites. Par le biais de trois expériences et une méta-analyse j'ai cherché à répondre à la problématique suivante : **comment optimiser le bénéfice des tests d'apprentissage pour une meilleure rétention à long terme en manipulant leur emplacement ?** Plus précisement, je me suis intéressée aux effets d'interaction entre l'apprentissage par récupération en mémoire et le positionnement, la granularité, et l'espacement des périodes de révisions (Chapitres 2-5). Les questions précises de l'ensemble de ce travail sont brièvement décrites ci-dessous pour introduire les chapitres associés à chaque expérience et à la méta-analyse. Tous ces chapitres constituent des articles à part entière (celui qui correspond au Chapitre 3 est publié). A l'exception de la discussion générale en français qui clôt ce manuscrit, tous les chapitres à suivre sont rédigés en anglais.

1.2.2.1 Problématique du Chapitre 2 (Méta-analyse)

Produire des mesures quantitatives suffisantes pour justifier la mise en œuvre d'une stratégie d'apprentissage plutôt qu'une autre nécessite une méthodologie particulière appelée méta-analyse. Combinant les résultats d'une série d'études indépendantes sur un problème donné, la méta-analyse permet une analyse plus précise d'un effet par l'augmentation du nombre de cas étudiés afin de tirer une conclusion globale. Etant donnés les bénéfices non négligeables des effets de récupération en mémoire et d'espacement pour maximiser la rétention à long terme, plusieurs travaux se sont intéressés à la façon dont les deux méthodes interagissaient.

Deux questions sont posées par cette méta-analyse : **i) Quelle est ampleur du bénéfice de révisions avec récupération en mémoire réalisées de manière espacée dans le temps sur la rétention de nouvelles connaissances ? ii) Quel est le régime d'espacement préférable pour planifier les épisodes de récupération en mémoire ?**

Pour répondre à la première question, j'ai compilé les études qui ont comparé les révisions avec récupération en mémoire espacées i) par rapport aux révisions avec récupération en mémoire massées dans le temps ; et ii) par rapport aux révisions avec relectures espacées dans le temps. Le but étant de quantifier le bénéfice d'une stratégie correspondant à la combinaison des deux difficultés désirables par rapport à une stratégie correspondant à une seule des difficultés désirables. La seconde question de recherche visait à compiler les études ayant comparé les deux régimes d'espacement, expansif et uniforme, dans le cas de révisions avec récupération en mémoire et de manière espacée dans le temps. Plusieurs variables ont été identifiées *a priori* comme pouvant modérer la taille des effets observés, ces modérateurs ont donc été inclus dans le modèle des analyses. Enfin, les implications théoriques et pratiques de ce travail sont discutées.

1.2.2.2 Problématique du Chapitre 3 (Expérience 1)

Une question qui a été soulevée dès le départ par Didask était celle de l'emplacement optimal des quiz d'apprentissage par rapport au contenu de formation à lire. Alors que la majorité des études sur l'effet de récupération en mémoire s'intéressent au bénéfice de cette stratégie pour le stockage quand les exercices de récupération étaient placés après l'exposition au contenu (appelé *post-testing effect* dans le chapitre 3) ; quelques études se sont intéressées à l'effet de la récupération en mémoire avant même d'avoir accès aux connaissances traitées par le cours et un *pre-testing effect* a été mesuré. En effet, les tests réalisés avant l'accès aux contenus à apprendre apporteraient des bénéfices sur les performances finales par rapport à un apprentissage contrôle de type relecture, et ce même quand le taux de réussite au pré-test est très faible. L'hypothèse est que le pré-test participerait à la mobilisation de l'attention du sujet durant l'apprentissage. De plus, la tentative de récupérer une information en mémoire même de façon erronée peut renforcer les "voies" de récupération potentielles entre la question et la réponse. Une nouvelle expérimentation paraissait nécessaire pour mieux comprendre cet effet de pré-test sur la mémorisation à long terme de contenus pédagogiques (connaissances scientifiques abordées au niveau lycée) ; tout en le comparant directement au *post-testing effect.* Les questions suivantes ont donc fait l'objet de la première expérience du projet de thèse :

i) De quelle l'ampleur sont les bénéfices de ces deux types de positionnement des tests d'apprentissage et dans quelles conditions sont-ils obtenus ? ii) Pour mieux mémoriser, l'apprentissage par des tests doit-il se faire avant ou après la lecture des contenus pédagogiques ?

1.2.2.3 Problématique du Chapitre 4 (Expérience 2)

Dans la continuité de la question abordée dans la première expérience, et toujours en lien avec la manière dont les contenus de formation sont présentés sur Didask, le découpage des épisodes de récupération en mémoire (appelé aussi granularité) a été étudié. Une première stratégie consiste à découper le contenu à lire en plusieurs sections de connaissances élémentaires, et de répartir les phases de récupération en mémoire après la lecture de chacune des sections, permettant ainsi une alternance de phase de lecture et de phase de récupération en mémoire. C'est la méthode utilisée par Didask : les cours sont découpées en grains très fins qui correspondent à des unités élémentaires de connaissance. Les tests d'apprentissage inclus dans chaque grain ne dépassent pas 5 à 6 questions. Une seconde stratégie est d'étudier l'intégralité du contenu à lire, puis de réaliser l'ensemble des tests associés à chaque section du contenu comme un tout. Cette question du degré de granularité des sessions de récupération en mémoire a fait l'objet de peu d'études. Celles-ci ont comparé ces deux niveaux de granularité et ont trouvé une supériorité d'un niveau de granularité fin par rapport à l'absence de découpage du contenu, mais seulement durant la phase d'entraînement et non lors de la phase d'évaluation des connaissances, et ce qu'importe l'intervalle de rétention entre phase d'entraînement et test final (une seule étude a trouvé un effet bénéfique du niveau de granularité fin sur la rétention). La question de la granularité des tests d'apprentissage paraissait donc très pertinente notamment le cadre d'un apprentissage de contenu dense et complexe. Dans l'expérience 2, nous nous sommes posés la question suivante : **Y a-t-il une interaction entre le niveau granularité des contenus (fin, moyen, grossier) et la stratégie d'apprentissage utilisée pour les apprendre (récupération en mémoire sous forme de tests vs relectures) ? Si oui, dans quelles conditions ?**

1.2.2.4 Problématique du Chapitre 5 (Expérience 3)

En écho aux résultats de la méta-analyse (Chapitre 2) et pour continuer à répondre à la question de l'optimisation de l'apprentissage par récupération en mémoire, une dernière expérience a consisté à étudier l'interaction de cette stratégie avec l'espacement. A notre connaissance, aucune étude n'a encore comparé dans un même design expérimental un

apprentissage avec la combinaison des deux effets (des tests d'apprentissage espacés dans le temps) à un apprentissage avec espacement sans tests d'apprentissage (c'est à dire des relectures espacées dans le temps) ; et à un apprentissage avec tests d'apprentissage massés dans le temps. Il paraissait donc intéressant de proposer une expérience réunissant ces situations d'apprentissage afin de quantifier leurs effets respectifs par rapport à une situation d'apprentissage contrôle par relectures massées ; mais aussi et surtout pour les comparer entre elles et quantifier leur bénéfice les uns par rapport aux autres. La problématique de cette dernière expérience est complémentaire à celles abordées dans les deux autres expériences puisque nous avons à nouveau cherché à comprendre comment maximiser le bénéfice de l'apprentissage par récupération en mémoire. Aussi, et dans la logique de fournir des résultats quantitatifs sur l'efficacité de différentes stratégies d'apprentissage, il paraissait pertinent de réunir dans un même design expérimental quatre stratégies d'apprentissages très étudiées dans le domaine de la psychologie cognitive appliquée à l'éducation. Enfin, nous avons évalué la métacognition des apprenants par le biais de « jugements d'apprentissage ». Les participants devaient prédire les notes qu'ils pensaient obtenir aux deux examens de connaissances, de façon prospective et rétrospective.

Cette dernière expérimentation sur Didask abordait quatre questions :

i) **De quelle ampleur est le bénéfice de chacune de ces différentes stratégies d'apprentissage sur la rétention de nouvelles connaissances ?**

ii) **Quel est l'effet d'un apprentissage combinant l'utilisation de tests et l'espacement des sessions de révisions dans le temps ?**

iii) **Dans quelle mesure la réponse à ces questions diffère-t-elle entre la rétention à court-terme (24h après l'apprentissage) et la rétention à long-terme (une semaine après l'apprentissage) ?**

iv) **Enfin, quelle est l'influence de ces différentes stratégies d'apprentissage sur la capacité des apprenants à prédire leurs performances aux examens finaux ?**

‘Curiouser and curiouser!’ cried Alice
(she was so much surprised, that for the moment she quite forgot how to speak good English).

— Lewis Carroll
Alice's Adventures in Wonderland (1865, 1869), Chapter II *The pool of tears.*

2 Chapitre 2 : “A meta-analysis of the effect of spaced retrieval practice on memory retention”

This section constitutes the following manuscript: Latimier, A., Peyre, H., Ramus, F. A meta-analysis of the effect of spaced retrieval practice on memory retention (in prep).

The data from the included studies and R script for analyses with robumeta package are available on OSF with the following link:
https://osf.io/jbmq4/?view_only=c05f2fcb93bc4d77b68fb11b1561d220.

2.1 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 effect of spaced retrieval practice. The total dataset was divided into three subsets to investigate the following questions: i) Does spaced retrieval practice induce better memory retention than massed retrieval practice? ii) Does spaced retrieval practice induce better memory retention than spaced restudy? iii) Are some schedules of spaced retrieval practice better than others? Using meta-regression with Robust Variance Estimation, 39 effects sizes were aggregated in subset 1, 16 in subset 2, and 54 in subset 3. Results from subset 1 indicated a strong benefit of spaced retrieval practice in comparison to massed retrieval practice ($g = 0.74$). Limited by the small number of aggregated effect sizes, comparing spaced retrieval practice to spaced restudying (subset 2) revealed a non-significant benefit of spaced retrieval practice ($g=0.46$). Finally, results from subset 3 indicated no difference between expanding and uniform spacing schedules of retrieval practice ($g = 0.032$). Moderator analyses in subset 3 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 to the uniform one. Overall, these meta-analyses support the advantage of spaced retrieval practice, but do not support the wide belief that inter-retrieval intervals should be progressively increased until a retention test.

2.2 Introduction

Research on learning practices typically consists in studying the effects of the conditions of a study phase on performance in a test phase. During the study phase, participants are exposed to the learning content, and they 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, also described as "desirable difficulties" (Bjork & Bjork, 2011; Brown et al., 2014). First, the effect of testing one's knowledge with retrieval practice during the study phase leads to better retention than just re-reading or re-exposure to the material. Second, 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). The present paper aims to provide a systematic literature review and meta-analysis of these effects and their interaction.

Retrieval practice effect

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. The retrieval practice effect, also known as the testing effect, has mainly been shown in the context of recall after being exposed to learning contents. However, a pretesting effect (i.e., practice before being exposed to learning contents) has also emerged from the recent literature and deserves further investigation (Carpenter & Toftness, 2017; Richland, Kornell, & Kao, 2009). Numerous studies have shown that retrieval practice enhances long-term retention, compared to re-reading (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006). Two recent meta-analyses have summarized these results. They have shown a strong and positive mean effect of using retrieval practice during learning compared to 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. (2017a) 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 applied (e.g., classroom) settings (Bangert-Drowns, Kulik, & Kulik, 1991; Karpicke & Grimaldi, 2012; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011a). Moreover, this learning strategy seems to provide knowledge transfer to untested but related information under certain circumstances (Chan, McDermott, & Roediger, 2006a; McDaniel, Thomas, et al., 2013; Pan & Rickard, 2018a).

Main moderators of the retrieval practice effect

The meta-analyses cited above have investigated the influence of several moderators on the magnitude of the retrieval practice effect. Rowland (2014) reported that retrieval practice 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 (2017a) suggested that the population of secondary students benefited more from retrieval practice than younger and older students populations ; and that classroom experiments showed similar benefits relative to laboratory ones. The authors also found that the strongest retrieval practice effect was obtained when training exercises and final exams had similar formats (i.e., validation of the *Transfer Appropriate Processing theory*, Franks, Bilbrey, Lien, & McNamara, 2000). 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 retention interval before a final assessement should be long enough to promote long term benefits without being too distant from the end of the learning phase (i.e., between 1 day and one month).

Spacing effect

Spacing refers to the deliberate insertion of lags between learning episodes, with the aim to optimally counteract forgetting and improve the consolidation of newly learned information (Cepeda et al., 2009; McDaniel, Fadler, & Pashler, 2013). Spacing is typically contrasted with massed learning, i.e., the repetitive exposure to the same contents without any time lag in-between. The spacing effect has been demonstrated in typical experiment where earners 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). 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 focussed 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 (the effect size d=0.96 was computed by 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 (2008a) conducted a

meta-analysis of these meta-analyses and provided a string mean effect size of d=0.71 for the spacing effect. A well-known meta-analysis by Cepeda et al. (2006) focussed 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. These authors did not estimate the mean effect size of the spacing effect itself. Thus, there is a need for a new meta-analysis of the spacing effect with various moderators and learning contexts (e.g. online learning).

As hinted above, the spacing effect has been shown in different domains (Dail & Christina, 2004; Mumford & And Others, 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, Vlach, & Sandhofer, 2014; Vlach & Sandhofer, 2012). Spacing effects also occur across age groups from children (Fritz, Morris, Nolan, & Singleton, 2007; Kalenberg, 2017) to healthy aging and individuals with memory impairments (D. A. Balota, Duchek, Sergent-Marshall, & Roediger, 2006). Spacing effects have been shown from very short inter-study intervals such a few seconds (e.g., Whitten and Bjork, (1977); to much larger intervals such as days or weeks (e.g., Dobson & al., (2016). Spaced learning also provides substantial improvements in long-term memory for learners in real educational settings (Carpenter et al., 2012; Karpicke, Blunt, & Smith, 2016; Larsen, 2018; Mettler, Massey, & Kellman, 2016; Seabrook, Brown, & Solity, 2005; Sobel, Cepeda, & Kapler, 2011).

Main moderators of the spacing effect

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 stydy episodes led to larger spacing effect sizes. However, more recent studies showed that for a given retention interval the effect of the lag between the study episodes follows an inverted and non-symmetrical u-shape (i.e., both too short and too long lags decrease the benefit of the spacing effect; Cepeda et al., 2009; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008).

From these reviews, it also emerged that the optimal lag between learning episodes may depend on the task: brief lags yielded stronger effects for simple tasks (i.e., perceptual or motor tasks), but longer lags appeared to be more beneficial for more complex tasks (i.e., text comprehension). Janiszewski et al. (2003) found larger spacing effects when using stimuli that are more meaningful and complex for the learners. Spacing also seemed more beneficial for explicit than for implicit learning. 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; Gerbier, Toppino, & Koenig, 2015; Godbole, Delaney, & Verkoeijen, 2014; Kapler, Weston, & Wiseheart, 2015). However, recent work showed 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 (Roediger III & Karpicke, 2011; Rohrer & Taylor, 2006). Recently, in a study conducted in young children with educational contents, Greving and Richter (2019) found a difference between massed and spaced rereading only at short term (massed outperformed spaced) but not at long term. Thus, the influence of the retention interval on the magnitude of the spacing effect remains unclear.

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. Two schedules of review have been frequently 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 between 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 favourable to promote long term retention, although it has showed benefits on short-term retention (Küpper-Tetzel, Kapler, & Wiseheart, 2014; Mozer, Pashler, Cepeda, Lindsey, & Vul, 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, Phelan, & Gerbier, 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 & Bjork (2012) found a superiority of the expanding schedule spaced presentation over the uniform one only at delayed final test but not at immediate final test (i.e., expanding and uniform were equivalent in term of final performances). Cepeda et al. (2006) meta-analysed 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 (William L. Cull, Shaughnessy, & Zechmeister, 1996; Kang, Lindsey, Mozer, & Pashler, 2014; Maddox, Balota, Coane, & Duchek, 2011; Storm, Bjork, & Storm, 2010), but other studies found the opposite result (W. L. Cull, 2000; Karpicke & Roediger, 2007; Logan & Balota, 2008, Toppino, Phelan, & Gerbier, 2018 in a high level initial training condition). In addition, a non-negligible part of the literature reports no difference between the two schedules (S. K. Carpenter & DeLosh, 2005; W. L. Cull, 2000; William L. Cull et al., 1996; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2010; Logan & Balota, 2008; Pyc & Rawson, 2007; Storm et al., 2010; Terenyi, Anksorus, & Persky, 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 good learning strategies (Brown et al., 2014; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Moreira, Pinto, Starling, & Jaeger, 2019; Rosenshine, 2010; Weinstein et al., 2018). Nevertheless, research on learning strategies has tended to examine the effect of various strategies in isolation from one another (Miyatsu, Nguyen, & McDaniel, 2018). Thus, it is time to investigate the combination of these effective learning strategies to measure their interaction and potential additive effects. In light of the gaps identified in the preceding literature review, of the dates of previous meta-analyses and of the vast number of studies published since them, the present study proposes three new meta-analyses focusing on the following questions:

i) What is the benefit of spaced retrieval practice relative to massed retrieval practice? (Subset 1)

ii) What is the benefit of spaced retrieval practice relative to spaced reading? (Subset 2)

iii) What are the relative benefits of expanding and uniform spacing schedules for retrieval practice? (Subset 3)

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.

2.3 Methods

Search Strategy and coding procedure

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 (David A. Balota, Duchek, & Logan, 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 (Google Scholar alerts).

Our searches generated 3948 results. To select eligible studies, we applied a first screening based on titles and abstracts. 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.

After eliminating duplicates and after reading each references' title and abstract to determine whether they fitted the research topic, 42 articles were selected in the initial round (Figure 1).

Inclusion criteria for the article-level review stage

The population of interest was a neurotypical population with no age limit and no specific education level. We included laboratory as well as classroom experiments. Thus, the second round of screening was based on full text reading, with the following criteria for eligibility:
i) The experimental set up included a training phase followed by an assessment phase (i.e. final test(s)).
ii) The training phase included repeated retrieval events over time with minimum two retrieval attempts for the same item.
iii) The design included a control group with massed retrieval practice or spaced restudying. 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 learning is considered to be studying when no retrieval or elaborative task was assigned to the control group (i.e., reading).

iv) AND/OR the design included an experimental manipulation of spacing, with at least expanding and uniform spacing schedules.

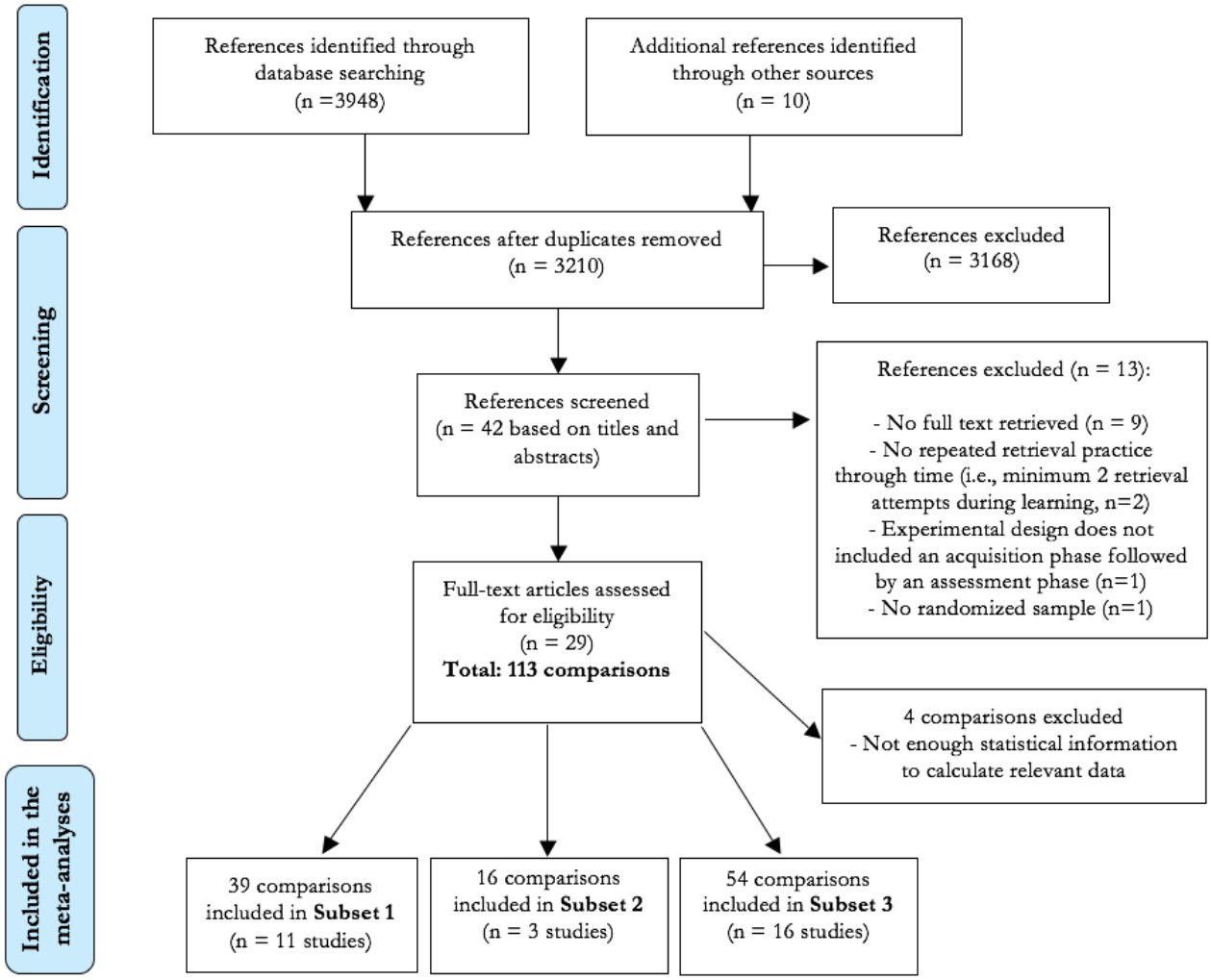
v) Participants were randomized into the experimental conditions.

vi) All necessary information for effect size calculations must have been reported or derivable (sample sizes with means + SD or SE), either in tables or in figures. When necessary and when available, graphs were digitized using the WebPlotDigitizer software to recover missing information such as standard error, standard deviations and means.

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, 31 satisfied all criteria, including 109 comparisons in total (Figure 1).

To answer our three main research questions, we separated the data into three subsets. Subset 1 included 39 comparisons between spaced and massed retrieval practice; Subset 2 included 16 comparisons between spaced retrieval practice and spaced restudying; and Subset 3 included 54 comparisons between an expanding and a uniform spacing schedule.

Several comparisons came from each study, and experiments/comparisons from a given study could be included in several subsets.



Note: n refers to individual studies.

Figure 1. PRISMA group flow diagram depicting study inclusion criteria (Moher, Liberati, Tetzlaff, & Altman, 2009). For each stage, we provide the number of included and excluded references, and the reason why we excluded some of them.

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 selected the following 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 (preschool, elementary and high school) and 12 years and more.

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, 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, 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 retrieval 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 and 2 we coded the variable as between two and four exposures vs more than 4 exposures. For Subset 3 we coded the variable as four exposures vs more than 4 exposures.

Type of spacing schedule (specific to Subsets 1 and 2) - In relation to the first research question, another moderator was coded: the type of spacing schedule compared to the massed schedule. It was coded as a categorical variable: expanding versus uniform.

Timing of the first retrieval event - 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 vs different timing.

2.4 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; between spaced restudying and spaced retrieval conditions for Subset 2; and between expanding and uniform schedule conditions for Subset 3. 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, standard deviation *s* was calculated as:

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

2) Cohen' s d was computed as:

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

Where the pooled standard deviation for a within-subject design was:

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

In addition, the pooled standard deviation 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:

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

In addition, the standard error of the effect size for a between-subject design was computed with:

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

M is the mean proportion correct for a given condition, *n* is the sample size for a condition, *s* is the standard deviation 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 spaced reading 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's g for each of the included effect sizes in order to correct for this bias, following this formula (Hedges, 1981):

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

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 (packages: robumeta (Fisher & Tipton, 2015)) was used to conduct the meta-analysis. Each subset of comparisons was analysed separately. The method we used to synthetize 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, Tipton, & Johnson, 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 have 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 (SD). In addition, we reported the estimated between-study heterogeneity T^2 that provides an estimate of the variance in the true effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2009); and the magnitude of the homogeneity I^2 (in %) among studies (as for the random effects model, Higgins, Thompson, Deeks, & Altman, 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 analysed separately. Finally, we reported publication bias analysis based on using a funnel plot inspection and using Egger's regression test (Egger, Davey Smith, Schneider, & Minder, 1997).

2.5 Results

General study characteristics

Subset 1: spaced vs. massed retrieval practice (Table 1)

We identified n=11 studies involving k=39 effect sizes for this comparison. 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 one day to more than one 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: spaced restudying vs spaced retrieval practice (Table 2)

We identified only three studies involving 16 comparisons for this subset. They were two studies on adults and one study on upper-secondary school adolescents, and all were set in classroom environment. The material used as the learning contents was more complex and ecological than laboratory settings, as it was extracted from textbooks and curricula (biology, courses for future driver, medical courses). Training sessions and final assessments were usually assessed using the same test type (free recall in majority), and the spaced retrieval condition included at least four retrieval attempts for each item. Retention intervals were quite long in comparison to laboratory settings (one week to several weeks).

Subset 3: comparing different spacing schedules (Table 3)

We identified 16 studies involving 54 effect sizes for this comparison. 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 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 3 used longer retention intervals than studies in Subset 1 (usually one 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.

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	21600	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	10080	UC	NR
Carpenter & 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 & 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 & Balota (2008)												
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 (short 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	57148	2	Un
Experiment 1	2.10	86	C	12+	BS	O	CR	Yes	CR	57148	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	40320	6	Con

Note. Effect sizes are indicated in Hedges's 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 spaced restudying and spaced retrieval practice

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	Spacing schedule
Pagliarulo (2011)												
Experiment 1	0.28	98	C	12+	BS	O	CR	No	CR	50400	5	Un
Experiment 1	0.11	98	C	12+	BS	O	CR	No	CR	50400	5	Un
Stenlund et al. (2016)												
Experiment 1	0.47	54	C	<12	WS	O	CR	Yes	CR	5	4	NR
Experiment 1	1.00	54	C	<12	WS	O	FR	Yes	CR	5	4	NR
Experiment 1	0.62	54	C	<12	WS	O	CR	Yes	FR	5	4	NR
Experiment 1	0.62	54	C	<12	WS	O	FR	Yes	FR	5	4	NR
Experiment 1	1.20	54	C	<12	WS	O	CR	Yes	CR	10080	4	NR
Experiment 1	0.98	54	C	<12	WS	O	FR	Yes	CR	10080	4	NR
Experiment 1	0.99	54	C	<12	WS	O	CR	Yes	CR	40320	4	NR
Experiment 1	0.95	54	C	<12	WS	O	FR	Yes	CR	40320	4	NR
Dobson (2016)												
Experiment 1	0.39	60	C	12+	WS	O	FR	No	FR	0	4	Con
Experiment 1	0.35	60	C	12+	WS	O	FR	No	FR	1440	4	Con
Experiment 1	0.37	60	C	12+	WS	O	FR	No	FR	40320	4	Con
Experiment 1	0.22	60	C	12+	WS	O	FR	No	FR	0	6	Con
Experiment 1	0.24	60	C	12+	WS	O	FR	No	FR	1440	6	Con
Experiment 1	0.27	60	C	12+	WS	O	FR	No	FR	40320	6	Con

Note. Effect sizes are indicated in Hedges's 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 3. Description and moderator information for studies included in Subset 3 comparing different spacing schedules

Subset 3 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 & 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	0.12	38	L	12+	WS	P	CR	No	CR	5	6	ST
Experiment 3	-0.17	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 & 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 & Rawson (2007)												
Experiment 1	0.02	64	L	12+	BS	P	CR	Yes	CR	40	4	DT
Logan & Balota (2008)												
Experiment 1	0	80	L	12+	WS	P	CR	NR	CR	0	4	DT
Experiment 1	0.32	66	L	12+	WS	P	CR	NR	CR	0	4	DT
Experiment 1	-0.37	24	L	12+	WS	P	CR	NR	CR	0	4	DT
Experiment 1	-0.28	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	10080	5	DT
Experiment 1	-0.15	88	C	12+	BS	TP	FR	No	CR	10080	5	DT
Experiment 2	0.99	30	C	12+	BS	TP	FR	No	FR	10080	5	DT
Experiment 2	0.67	30	C	12+	BS	TP	FR	No	CR	10080	5	DT
Storm et al. (2010)												
Experiment 3	0.64	16	C	12+	WS	TP	CR	No	CR	10080	4	DT
Experiment 3	-0.21	18	C	12+	WS	TP	CR	No	CR	10080	4	DT

Subset 3 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
Karpicke & Roediger (2010)												
Experiment 1	0.07	40	L	12+	BS	TP	FR	No	FR	10080	4	ST
Experiment 1	-0.25	40	L	12+	BS	TP	FR	No	FR	10080	4	ST
Experiment 2	0.08	32	L	12+	BS	TP	FR	Yes	FR	10080	5	ST
Experiment 2	0.12	32	L	12+	BS	TP	FR	Yes	FR	10080	5	ST
Experiment 2	-0.38	32	L	12+	BS	TP	FR	Yes	FR	10080	5	ST
Experiment 2	0.08	32	L	12+	BS	TP	FR	Yes	FR	10080	5	ST
Karpicke & Bauernschmidt (2011)												
Experiment 1	0.08	48	L	12+	WS	P	CR	No	CR	10080	4	DT
Experiment 1	0.24	48	L	12+	WS	P	CR	No	CR	10080	4	DT
Experiment 1	-0.13	48	L	12+	WS	P	CR	No	CR	10080	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	41760	5	ST
Experiment 1	0.67	93	L	12+	BS	TP	CR	No	CR	41760	5	DT
Experiment 1	0.22	103	L	12+	BS	TP	CR	No	CR	41760	5	DT
Experiment 1	0.47	99	L	12+	BS	TP	CR	No	CR	41760	5	DT
Experiment 1	0.76	196	L	12+	BS	TP	CR	No	CR	41760	5	NR
Dobson (2013)												
Experiment 1	-0.13	91	C	12+	BS	O	CR	No	CR	NR	10	ST
Kang al. (2014)												
Experiment 1	0.20	37	L	NR	WS	P	CR	Yes	CR	80640	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	10080	UC	DT
Experiment 1	0.07	34	L	12+	BS	P	CR	Yes	CR	50400	UC	DT
McGregor (2014)												
Experiment 1	-0.52	91	L	12+	BS	P	CR	No	FR	1020	NR	ST
Mettler et al. (2016)												
Experiment 1	0.25	36	L	12+	WS	P	CR	Yes	CR	0	4	DT
Experiment 1	0.05	36	L	12+	WS	P	CR	Yes	CR	10080	4	DT
Kalenberg (2017)												
Experiment 1	-0.20	32	C	3 to 4	BS	P	CR	Yes	CR	10080	4	DT

Note. Effect sizes are indicated in Hedges's 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

Effect size analyses [23]

Subset 1: Spaced retrieval practice vs. others learning strategies

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 = 1.02 (95% CI: [0.68, 1.36], p < 0.0001) with an estimated between-study SD of 0.15 (Table 4). Varying the assumed within-study effect size correlation (ρ) had no impact on g, ranging from 1.020 to 1.021 (Appendix 1) and a minimal impact on the estimated between study-variance (T^2), which ranged from 0.219 when ρ = 0 to 0.230 when ρ = 1. Heterogeneity was moderate (Higgins' I^2 = 51.06 %).

Table 4. Summary of the weighted mean effect sizes for Subset 1, 2, and 3.

Subset (k=number of effect sizes)	g	SD	df	*p*	95% CI
Subset 1: Spaced retrieval practice vs. massed retrieval practice (k = 39)	1.02	0.15	9.78	<0.01***	[0.68, 1.36]
Subset 1: Trim-and-fill correction (k = 49)	0.74	0.09	19.4	<0.01***	[0.55, 0.92]
Subset 2: Spaced retrieval practice vs. spaced restudying (k = 16)	0.46	0.2	1.93	0.15	[-0.41, 1.33]
Subset 3: expanding vs. uniform retrieval practice schedule (k = 54)	0.032	0.06	13.7	0.62	[-0.10, 0.17]

Note. Weighted mean effect size in terms of Hedges' g; SD: between-study standard error; df: adjusted degrees of freedom; CI: confidence interval. Results are not reliable when df < 4. Signif. Codes: < .01 *** < .05 ** < .10 *.

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, df = 37, p < 0.0001) confirmed that the funnel plot was asymmetrical (Figure 2, A). This makes it likely that publication bias has occured: some studies with negative or non-significant findings were probably not published and therefore were not included in this meta-analysis. Thus, the mean

[23] All data from included studies and R script for analyses with robumeta package are available on OSF with the following link https://osf.io/jbmq4/?view_only=c05f2fcb93bc4d77b68fb11b1561d220.

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 4, see Appendix 2 for the trim-and-fill funnel plot). The overall weighted mean effect size was reduced to g = 0.74 (95% CI: [0.55, 0.92], $p < 0.0001$), with a moderate heterogeneity ($I^2 = 48.19$ %) too.

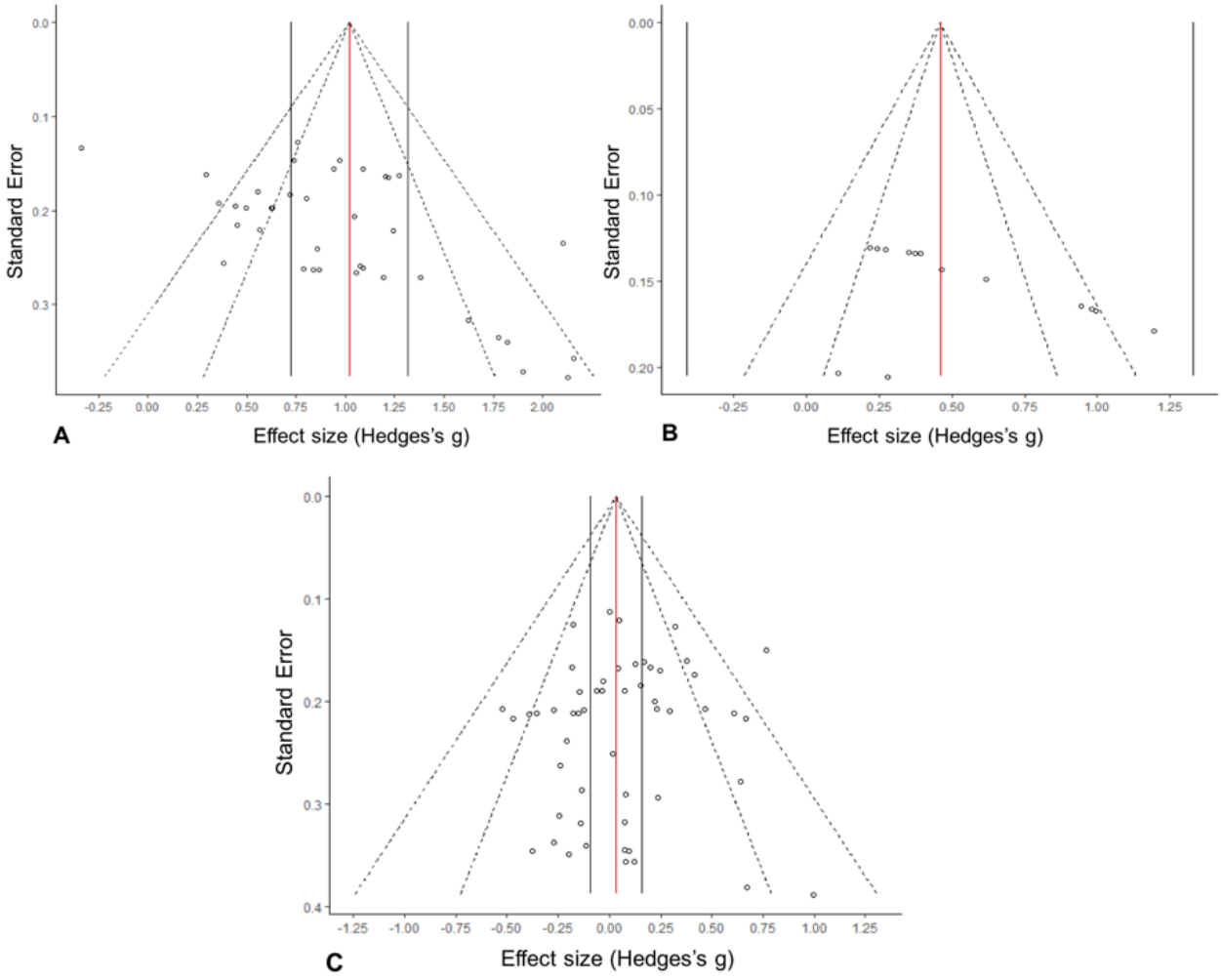


Figure 2. Funnel plots for Subset 1 (A), Subset 2 (B), and Subset 3 (C). Each point represents the effect size of one included comparison. The X axis represents Hedges's g for each comparison, and the Y axis is the corresponding standard error. Red solid line: mean effect size; black solid lines: CI for mean effect size; dashed lines: lower-limit and upper limit values for the 95% CI and 99% CI regions.

Subset 2: Spaced retrieval practice vs. spaced restudying

Weighted mean effect size -The dataset included 16 effect size estimates from three unique studies, with between two and eight effect sizes per study (min = 2, median = 5.33, max = 8). The overall weighted mean effect size across all 16 effect size estimates was g = 0.46 (95% CI = [-0.41, 1.33], p = 0.148) with an estimated between-study SD of 0.20 (Table 4). Nevertheless, this effect size was not significantly different from zero, probably due to limited degrees of freedom (df = 1.93). Higgins test suggested no heterogeneity (I^2 = 0 %).

Publication bias analysis– We estimated publication bias for this subset using a funnel plot representation (Figure 2, B). Egger's regression test was not significant, suggesting that the funnel plot is symmetrical and therefore that there is no publication bias (t = 1.6286, df = 14, p = 0.1257).

Subset 3: expanding vs. 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 = 0.032 (95% CI = [-0.10, 0.17], p = 0.62) with an estimated between-study SD of 0.06 (Table 4). Varying the assumed correlation between within-study conditions had no impact at all on the mean effect size or standard error and no impact on the estimated between study variance (T^2) when $\rho = 0$ to when $\rho = 1$ (Appendix 1). Higgins test suggested no heterogeneity (I^2 = 0 %).

Publication bias analyses for group designs – We estimated publication bias for this subset with a funnel plot representation (Figure 2, C). Egger's regression test was not significant, suggesting that the funnel plot was symmetrical and therefore that there was no publication bias (t = -0.4376, df = 52, p = 0.6635).

Potential moderator effects

We carried out moderator analyses for the two subsets for which there was a sufficient number of comparisons (subsets 1 and 3, Table 5). 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 addition, we computed univariate analyses only because degrees of freedom were insufficient for multivariate analyses

Table 5. Moderator analyses for each subset of comparisons

Moderators	β	SE	95% CI		df	*p*
			LL	UL		
Retention Interval (log minutes)	0.02	0.06	-0.13	0.16	5.38	0.79
Design (between vs. **within subjects**)	0.38	0.50	-1.48	2.25	2.35	0.51
Feedback (no vs. **yes**)	0.38	0.23	-0.18	0.95	5.51	0.15
Stimuli (pairs vs. **others**)	-0.30	0.24	-0.95	0.36	4.20	0.28
Number of exposures for a given item (2- 4 vs. **more than 4**)	-0.41	0.16	-1.05	0.23	2.19	0.12
Training test type (cued-recall vs. **free recall**)	-0.15	0.34	-1.31	1.00	2.64	0.68
Spacing schedule (expanding vs. **uniform**)	-0.05	0.11	-0.39	0.29	3.23	0.69
Retention interval (log minutes)	0.03	0.04	-0.07	0.13	7.45	0.54
Design (between vs. **within subjects**)	0.11	0.17	-0.29	0.50	8.14	0.55
Feedback (no vs. **yes**)	-0.13	0.09	-0.35	0.10	8.23	0.23
Stimuli (pairs vs. **others**)	0.24	0.17	-0.23	0.71	4.29	0.24
Number of exposures for a given item (4 exposures vs. **more than 4**)	0.22	0.11	-0.04	0.48	8.51	0.09 *
Training test type (cued-recall vs. **free recall**)	-0.09	0.10	-0.43	0.24	2.93	0.43
Final assessment type (cued-recall vs. **free recall**)	-0.17	0.10	-0.47	0.13	3.18	0.18
Placement of the first retrieval attempt (different vs. **same**)	0.25	0.23	-0.43	0.93	3.53	0.35

Notes. For discrete moderators, categories are indicated in brackets with the reference that is not in bold. SE: standard error; CI: confidence interval; LL: Lower limit for the interval of confidence; UL: Upper limit for the interval of confidence; df: adjusted degrees of freedom. Results are not reliable when df < 4. Signif. codes: < .01 *** < .05 ** < .10 *.

For Subset 1, no moderator came close to statistical significance at $\alpha = 0.1$. This analysis was limited by the limited degrees of freedom available for many moderators (it is estimated that at least 4 dfs are necessary for a reliable analysis (Fisher & Tipton, 2015)).

For Subset 3, comparisons including more than 4 exposures for each item to learn (n = 25) were associated with increased effect sizes compared to those including 4 exposures (n = 26) (p = 0.09). This suggests that more than 4 exposures might be needed to differentiate expanding from uniform schedules. Indeed, when each item was presented more than 4 times (initial exposure + retrieval practice attempts) during the training phase, g = 0.2 (95% CI = [-0.07, 0.46], df = 4.65, p = 0.12), compared to g = -0.04 (95% CI = [-0.17, 0.094], df = 9.91, p = 0.52) when it was presented 4 times or less. No other moderator was statistically significant at α =

0.1. This analysis was also limited by the limited degrees of freedom available for many moderators.

2.6 Discussion

Our analysis of Subset 1 (11 studies, 39 comparisons) using Robust Variance Estimation indicated a large advantage of spaced retrieval practice over massed retrieval practice (g = 1.02, 95% CI = [0.68, 1.36]). There was evidence for publication bias, however, even after correcting for publication bias with the trim-and-fill method, the effect remained substantial (g = 0.74, 95% CI = [0.55, 0.92]). This overall mean effect size is consistent with previous meta-analyses of spaced vs. massed learning (Donovan & Radosevich, 1999; Hattie, 2008a; 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. Unfortunately, moderator analyses failed to illuminate this heterogeneity.

Our analysis of Subset 2 (3 studies, 16 comparisons) indicated a non-significant moderate effect of spaced retrieval practice compared to spaced restudying (g = 0.46, 95% CI = [-0.41, 1.33], p = 0.148). Moderator analyses were not possible due to the low number of comparisons. Although the effect size of g = 0.46 is associated with a considerable margin of error, it is consistent with previous estimations of retrieval practice effect, which ranged from g = 0.5 (Rowland, 2014) to g = 0.61 (Adesope et al., 2017a). Thus, the results of the Subset 2 meta-analysis, although non-significant, can be minimally interpreted as showing that retrieval practice seems to have similar effects in spaced than in other learning schedules (i.e., massed fashion).

A tentative compilation of Subset 1 and 2 meta-analyses together with previous estimates of both retrieval practice and spacing effects is shown in Figure 3. Overall, this summary illustrates the well-established effects of both retrieval practice and spacing, and suggests that the spacing effect (g = 0.71) may be larger than the retrieval practice effect (g = 0.5 to 0.61), and is consistent with the hypothesis that their effects are simply additive. The available data collected for the present review do not allow us to test directly the hypothesis of an interaction

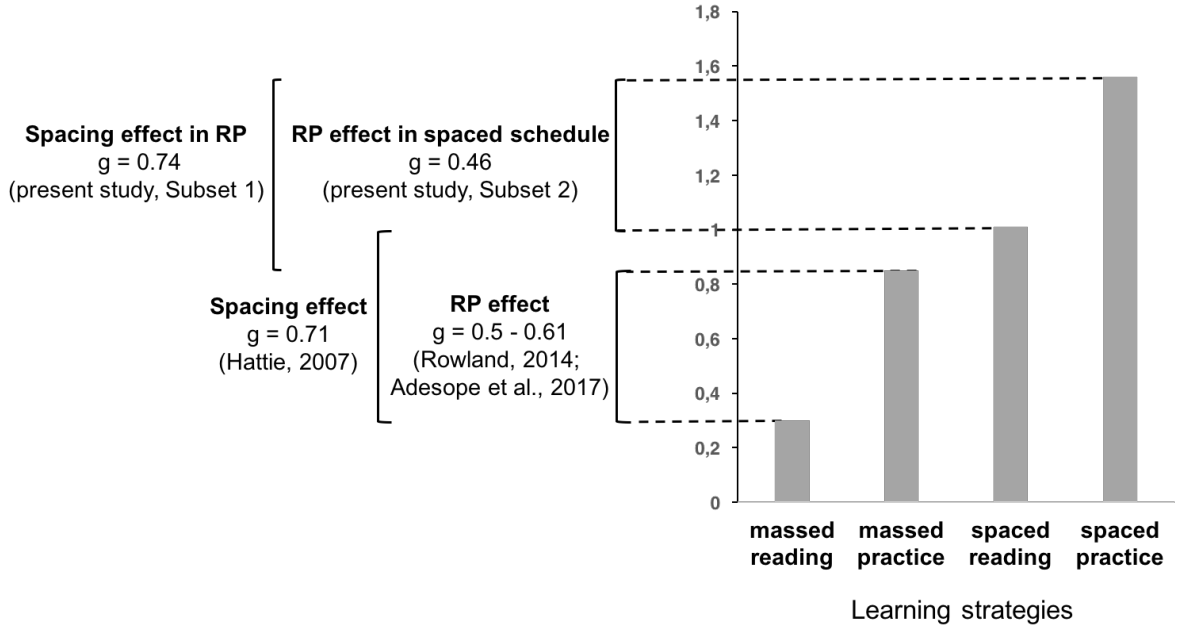


Figure 3. Synthesis of present and previous meta-analyses with the comparison between four learning strategies. The Y-axis is expressed in terms of mean effect size (Hedges's g), with an arbitrary value for the massed reading strategy (g = 0.3). "RP" is for retrieval practice.

Finally, our analysis of Subset 3 (16 studies, 54 comparisons) indicated a non-significant effect of spacing schedule (g = 0.032 (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) whereas 43% were negative (i.e., equal-uniform schedule superiority), with one equal to zero. Thus, 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. Indeed, there was an advantage of the expanding over the uniform schedule when there were many exposures (more than 4) to each unit of knowledge than when there were fewer, leading to a disadvantage of the uniform schedule at 4 exposures. The role of this moderator (p = 0.09) and the advantage of the expanding schedule at high repetition rates (g = 0.2 (95% CI = [-0.07, 0.46]) being non-significant, this conclusion remains tentative. Other potential moderators of the spacing schedule effect have been proposed. Firstly, the advantage of the expanding schedule might depend on the retention interval before a final test, as suggested by several findings (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 (β = 0.02, 95% CI = [-0.13,

0.16]). Secondly, 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 3 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. It could also be argued that the best learning schedule is an adaptive one, taking into account each learner's performance on each item, thus making comparisons between generic uniform vs expanding schedules less relevant (Lindsey et al., 2014; Tabibian et al., 2019).

The present results of the 3 meta-analyses are complementary to those that were reported in previously published meta-analyses. Indeed, the intervals between the training 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 massed reading and spaced testing, and 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 between 0.1 and 0.3, see Figure 5).

The results from Subset 3 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 or contrary well-known results (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 better than 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 initial presentation of the item.

Obviously, a number of factors limits our conclusions. 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 favour 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. Concerning Subset 3, our finding of no difference between spacing schedules is unlikely to result from a publication bias, since this would probably have favoured the expanding schedule. Second, there were too few studies and comparisons in Subset 2 to reliably estimate the difference between spaced testing and spaced reading. Third, there was significant heterogeneity between studies (in Subset 1), which our moderator analyses failed to explain. Finally, diversity in some 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.

2.7 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., 2007a). 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 favour of spaced retrieval practice are warranted (Horvath, Lodge, & Hattie, 2016; Kang, 2016; Weinstein et al., 2018). 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 more evidence on the potential interest of combining both practices. This would also allow one

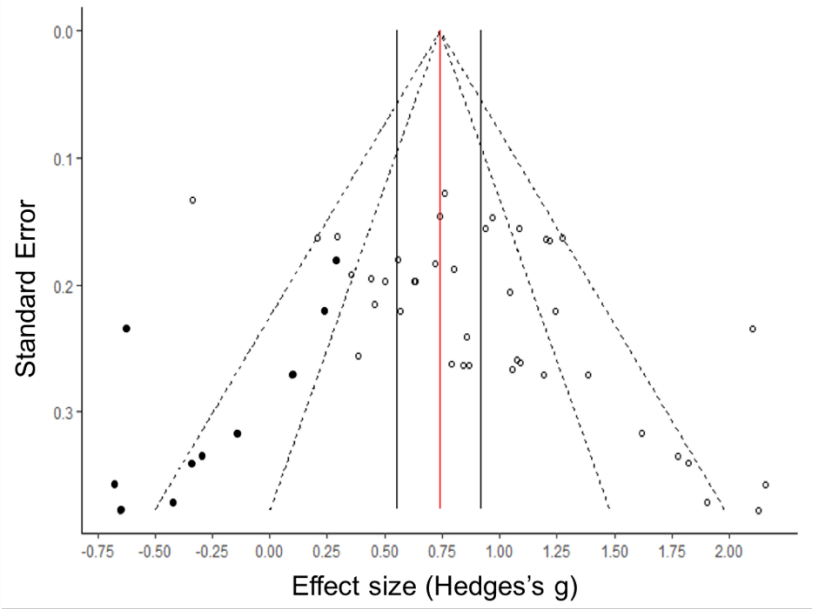
to quantify the effects whose estimation was unreliable (retrieval practice in spaced schedules) or impossible (spaced reading vs. massed practice), thus providing a more direct comparison of the two practices (in case one has to choose). Finally, we call for new studies comparing different spacing schedules, both in restudying and in retrieval practice conditions, since the currently available evidence seems inconclusive.

2.8 Appendices

Appendix 1. Sensitivity analysis for Subset 1 and Subset 3. 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	1.020	1.020	1.020	1.021	1.021	1.021
Std. Error	0.152	0.152	0.152	0.152	0.152	0.152
Tau^2	0.219	0.221	0.223	0.225	0.228	0.230

Subset 3	Rho = 0	Rho = 0.2	Rho = 0.4	Rho = 0.6	Rho = 0.8	Rho = 1
Mean effect size	0.0318	0.0318	0.0318	0.0318	0.0318	0.0318
Std. Error	0.0632	0.0632	0.0632	0.0632	0.0632	0.0632
Tau^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000



Appendix 2. Trim-and-fill funnel plot for Subset 1 publication bias correction. Each point represents the effect size of one included comparison. The X-axis is Hedges's g for each comparison, and the Y-axis the corresponding standard error. White dots are effects sizes from the included comparisons, while black dots are those added by the trim-and-fill procedure (10 new effect sizes). Red solid line: mean effect size; black solid line: CI for mean effect size; dashed lines: lower-limit and upper limit values for the 95% CI and 99% CI regions.

2.9 References

Adipose, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, 34654316689306. 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., & 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* (pp. 83–105). New York, NY, US: Psychology Press.

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

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of Frequent Classroom Testing. *The Journal of Educational Research*, *85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818

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* (New York: Worth Publishers, pp. 56–64).

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. https://doi.org/10.1002/9780470743386

Brown, P. C., Roediger (III), H. L., & McDaniel, M. A. (2014). *Make It Stick*. Harvard University Press.

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

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., & Toftness, A. R. (2017). The Effect of Prequestions on Learning from Video Presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

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. https://doi.org/10.1080/02701367.2004.10609146

Dobson, J. L., Perez, J., & Linderholm, T. (2016). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education*, n/a-n/a. 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]*. Retrieved from http://arxiv.org/abs/1503.02220

Franks, J. J., Bilbrey, C. W., Lien, K. G., & McNamara, T. P. (2000). Transfer-appropriate processing (TAP). *Memory & Cognition*, *28*(7), 1140–1151. https://doi.org/10.3758/BF03211815

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. https://doi.org/10.1080/17470218.2011.600806

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

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 (Hove, England)*, *22*(5), 462–469. https://doi.org/10.1080/09658211.2013.798416

Greving, C. E., & Richter, T. (2019). Distributed Learning in the Classroom: Effects of Rereading Schedules Depend on Time of Test. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02517

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

Hattie, J., Rogers, H. J., & Swaminathan, H. (2014). The Role of Meta-analysis in Educational Research. In *A Companion to Research in Education* (pp. 197–207). Springer.

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

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

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.

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. (n.d.). *Spaced and Expanded Practice: An Investigation of Methods to Enhance Retention*. Retrieved from http://cornerstone.lib.mnsu.edu/cgi/viewcontent.cgi?article=1205&context=jur

Kang, S. H. K. (2016). Spaced Repetition Promotes Efficient and Effective Learning Policy Implications for Instruction. *Policy Insights from the Behavioral and Brain Sciences*, 2372732215624708. 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*. https://doi.org/10.3389/fpsyg.2016.00350

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*(Supplement C), 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. 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. 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. (2013). 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. (2013). 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

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

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five Popular Study Strategies: Their Pitfalls and Optimal Implementations. *Perspectives on Psychological Science*, *13*(3), 390–407. https://doi.org/10.1177/1745691617710510

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). https://doi.org/10.1371/journal.pmed.1000097

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval Practice in Classroom Settings: A Review of Applied Research. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00005

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., & Others. (1994). Influence of Abilities on Performance during Practice: Effects of Massed and Distributed Practice. *Journal of Educational Psychology*, *86*(1), 134–144.
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*. Retrieved from http://eric.ed.gov/?id=ED498555
Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory and Cognition*, *35*(8), 1917–1927. Retrieved from Scopus.
Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496
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). New York, NY, US: Psychology Press.
Rohrer, D., & Taylor, K. (2006). *The Effects of Overlearning and Distributed Practice on the Retention of Mathematics Knowledge*. Retrieved from https://eric.ed.gov/?q=(spaced+OR+distributed+OR+retrieval)+AND+(practice+OR+testing)&pg=64&id=ED505642
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
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). 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., 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.

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). Retrieved from https://search.proquest.com/docview/1297338409/citation/99A7F419CA124520PQ/1

3 Chapitre 3 : “Does pre-testing promote better retention than post-testing?”

This section is a reprint of the following paper : Latimier, A., Riegert, A., Peyre, H., Ly, S. T., Casati, R., & Ramus, F. (2019). Does pre-testing promote better retention than post testing? *Npj Science of Learning*, *4*(1), 1–7. https://doi.org/10.1038/s41539-019-0053-1

The preregistration for this experiment can be accessed at

http://aspredicted.org/blind.php?x=wg7nj6

Learning materials[24], de-identified data, and data analysis scripts are posted on Dataverse website with the following link: https://doi.org/10.7910/DVN/XPYPMF

3.1 Abstract

Compared to other learning strategies, retrieval practice seems to promote superior long-term retention. This has been found mostly in conditions where learners take tests after being exposed to learning content. However, a pre-testing effect has also been demonstrated, with promising results. This raises the question, for a given amount of time dedicated to retrieval practice, whether learners should be tested before or after an initial exposure to learning content. Our experiment directly compares the benefits of post-testing and pre-testing relative to an extended reading condition, on a retention test 7 days later. We replicated both post-testing (d = 0.74) and pre-testing effects (d = 0.35), with significantly better retention in the former condition. Post-testing also promoted knowledge transfer to previously untested questions, whereas pre-testing did not. Our results thus suggest that it may be more fruitful to test students after than before exposure to learning content.

[24] L’intégralité du cours sur l’ADN ainsi que des exemples de questions d’apprentissage sont fournis en dans la section « Annexes » (Annexe 1)

npj | Science of Learning

www.nature.com/npjscilearn

ARTICLE OPEN

Does pre-testing promote better retention than post-testing?

Alice Latimier[1,2], Arnaud Riegert[3], Hugo Peyre[1,4,5], Son Thierry Ly[3], Roberto Casati[2] and Franck Ramus[1]

Compared with other learning strategies, retrieval practice seems to promote superior long-term retention. This has been found mostly in conditions where learners take tests after being exposed to learning content. However, a pre-testing effect has also been demonstrated, with promising results. This raises the question, for a given amount of time dedicated to retrieval practice, whether learners should be tested before or after an initial exposure to learning content. Our experiment directly compares the benefits of post-testing and pre-testing relative to an extended reading condition, on a retention test 7 days later. We replicated both post-testing ($d = 0.74$) and pre-testing effects ($d = 0.35$), with significantly better retention in the former condition. Post-testing also promoted knowledge transfer to previously untested questions, whereas pre-testing did not. Our results thus suggest that it may be more fruitful to test students after than before exposure to learning content.

npj Science of Learning (2019)4:15 ; https://doi.org/10.1038/s41539-019-0053-1

INTRODUCTION

The testing effect is a strong and well demonstrated effect.[1–5] As opposed to common learning practices such as reading, taking tests, and more generally retrieval practice during the learning phase contribute to better long term retention[6] by reducing the forgetting rate of information across time.[7] The benefits of retrieval practice have been demonstrated in both laboratory and classroom settings[8,9] and for both simple (e.g. word lists) and complex (e.g. prose passages) material.[10] In a meta-analysis, Rowland[11] reported a mean effect size of $g = 0.50$ [IC-95%: 0.42, 0.58] from 61 studies comparing the effects of testing vs. restudying on the ability to learn new information after a first exposure to learning contents. In another meta-analysis, Adesope et al.[12] found a mean effect size of $g = 0.61$ [IC-95%: 0.58 and 0.65] by comparing retrieval practice to other practices.

The testing effect may also lead to better retention of previously untested information and to greater knowledge transfer than restudying.[13–15] A recent meta-analysis[16] on the transfer of retrieval practice effects found that retrieval practice yielded transferrable learning relative to a restudying control condition ($d = 0.40$, 95%CI [0.31 and 0.50]). However, transfer does not necessarily occur in all circumstances and with all types of content.[14] Pan and Rickard[16] made a distinction between untested application and inference questions and untested information seen during the first exposure to material. Interestingly, they found a transfer effect of retrieval practice on application and inference questions, but not on untested information seen during initial study.

The testing effect has mainly been shown in the context of tests given after exposure to learning contents. However, a pre-testing effect has also been shown in laboratory settings with promising results. Indeed, taking a test before being exposed to learning content enhances retention compared with no retrieval practice.[17–23] Although the pre-testing effect was demonstrated on written materials (prose passages as well as paired words), a recent study also found it on video-based learning as well.[17] Moreover, Hartley's results indicated that the harder the pre-tested questions, the larger the improvement on retention. Even when the rate of success obtained in the pre-test was very low (in Richland et al.'s study, participants got as many as 95% of the pre-test answers wrong), learning with a pre-test was better than just studying twice the content.[22,24] Some of these studies reported that pre-testing also facilitated the learning of untested information,[17,19,25] while others found an effect only for pre-tested information, suggesting a lack of transfer.[21,22] In two studies, there was actually a decrease in final performance for the untested information compared with a control group with no pre-test, suggesting a detrimental effect on untested material.[20,23] This was also found in a literature review on question position when learning prose materials.[26] Thus, while the post-testing effect seems to transfer to untested material under certain conditions,[16] the evidence is more ambivalent for the pre-testing effect. One might predict that pre-testing might narrow attentional focus to tested information only, thus harming the learning of untested information.

In terms of putative mechanisms, the experiments of Pressley et al.[21] and Richland et al.[22] showed that just reading the pre-test is insufficient to enhance final retention, it is the process of attempting to find relevant answers that accounts for the pre-testing effect. Generating an answer, even an incorrect one, may strengthen the retrieval routes between the question and the correct answer and encourage deep processing.[24] Furthermore, being exposed to questions ahead of time may also help focus one's attention to the most relevant parts of the learning content. However, these studies showed an effect of pre-testing on retention only when compared with additional study, or to a passive-learning condition.

Given prior knowledge on the testing effect, and in a context of trying to optimise the time allocated to learning, a more relevant question would be: Should teachers test their students before or

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France; [2]Institut Jean Nicod, ENS, EHESS, CNRS, PSL University, Paris, France; [3]Didask, Paris, France; [4]Université Paris Diderot, Sorbonne Paris Cité, UMRS 1141, Paris, France and [5]Assistance Publique-Hôpitaux de Paris, Robert Debré Hospital, Child and Adolescent Psychiatry Department, Paris, France
Correspondence: Alice Latimier (alice.latimier@ens.fr)

Received: 29 January 2019 Accepted: 9 July 2019
Published online: 24 September 2019

after the lecture? Existing studies have only tangentially addressed the question. Studies conducted by Frase[27,28] were not exactly a comparison of pre- and post-testing, but compared placing adjunct questions within prose passages either before or after target information, and found better retention in the latter case. Similarly, in Sagaria and Di Vesta's[23] study, participants read pairs of paragraphs and questions, and the authors compared the effects of reading and copying each question just before or just after reading the paragraph. Thus this experimental situation does not really emulate testing before or after a lecture. Two quantitative reviews of such studies on adjunct questions during learning of prose materials reported similar effects of pre- and post-target information questions on tested material, but an advantage of post-questions on untested material.[26,29] However, the studies did not directly compare the effect of questions before and after reading a prose passage. Rather, pre- and post-testing effects measured in different studies were compared. Finally, McDaniel et al.'s[30] Experiments 2A and 2B compared the effects of pre- and post-lesson quizzes, but in fact both quizzes and the lecture were preceded by an initial reading of the chapter, so this did not strictly speaking constitute pre-testing but rather interim testing. Both experiments reported similar results, i.e., lower final performance in the pre-test than in the post-test condition.

In 2007, the US Department of Education published a summary report with several recommendations for improving teaching to reinforce learning.[31] One of them was to "use pre-questions to introduce a new topic". Yet the level of evidence associated with this recommendation was indicated to be low, because of the scarcity of relevant studies on the pre-testing effect. However, in the 12 years since the publication of the practice guide, the potential of pre-testing to promote better learning in educational settings still remains little explored. It therefore remains important to measure in the same experiment whether it is truly beneficial to spend time testing students before the first exposure to learning content, rather than afterwards. Furthermore, it is also uncertain whether pre-testing promotes as much transfer to untested material as post-testing does.

In this paper, we report an experiment that directly compares the learning effects of pre-testing (quiz-reading condition), post-testing (reading-quiz condition), and re-reading (reading-reading condition) on long term memory. Specifically, we aimed to determine (i) the size of testing effects, and how they compare between pre-testing and post-testing; (ii) to what extent pre- and post-testing benefits transfer to questions that were not tested (trained vs. untrained questions).

RESULTS

Participants' data

From a total of 334 recruited participants, 44 participants (13%) gave up participation between the learning phase and the final test, four participants were excluded because they did not complete the learning phase entirely, and one was excluded for completing the learning phase twice in a day, so we had a total of 285 participants that were included in the analysis. Demographic data are indicated in Table 1. An independent-samples *t*-test revealed no significant differences between the three groups in terms of age and education level (ps > 0.05). The sex ratio did not differ either (χ^2 (2) = 0.32 and $p = 0.85$). We also asked participants to estimate their degree of knowledge on DNA on a scale from 1 ("I do not have any knowledge on DNA") to 5 ("I have extensive knowledge on DNA"). Most of the participants were unfamiliar with DNA ($M = 1.59$; S.D = 0.69), and this did not differ between the groups (ps > 0.5).

Table 1. Summary characteristics of the three groups of participants (reading-reading, quiz-reading, and reading-quiz)

	Learning conditions			
	Reading-reading	Quiz-reading	Reading-quiz	Total
N	95	95	95	285
Gender (male/female)	40/55	39/56	36/59	115/170
Age (years)	34.8 (9)	35.9 (9)	37.7 (10.9)	36.1 (9.7)
Education (years)	14.4 (1.7)	14.6 (1.7)	14.5 (1.9)	14.6 (1.9)
Degree of knowledge on DNA (from 1 to 5)	1.6 (0.7)	1.6 (0.7)	1.6 (0.7)	1.6 (0.7)

Standard deviations are in parentheses

Training phase

Quiz performance. Even though we instructed participants to complete each training module only once, a few participants returned to some of the modules a second time. The mean number of module studied (for a total of 7) during the training phase was 7.24 ($SD = 0.86$) in the quiz-reading group and 7.56 ($SD = 1.13$) in the reading-quiz group ($t(183) = 2.14$, $p = 0.033$, and $d = 0.31$). This variable was used as a covariate for the final test analyses. When taking into account only the first attempt for each question, the reading-quiz group had a better total score ($M =$ 68.09%, $SD = 16.99\%$) than the quiz-reading group ($M = 47.15\%$, $SD = 12.92\%$), $t(188) = 9.56$, $p < 0.0001$, and $d = 1.39$.

Training times. We also computed the total time spent in the training phase, i.e. the cumulative time spent on learning contents and on quizzes, including quizzes that were taken twice or more than once. For technical reason, training time could not be calculated for eight participants.

Because times are not normally distributed, statistics were computed on the natural logarithm of the total training time. There was a main effect of the learning condition on the time spent on training contents, $F(2,274) = 125.52$, $p < 0.0001$, and $\eta^2 =$ 0.48. The amount of time spent on the training session was longer for the quiz-reading group than for the reading-quiz group. Both the groups with quizzes spent significantly more time on the training session than the group that was not assigned to learn with quizzes (Table 2).

Final test phase

A two-way ANOVA showed a main effect of the learning condition, $F(2,564) = 13.56$, $p < 0.0001$, and $\eta^2 = 0.035$; with the reading-quiz group showing the best final performance, then the quiz-reading group, and finally the reading-reading group (Table 3). There was also a main effect of question type ($F(1,564) = 187.78$, $p < 0.0001$, and $\eta^2 = 0.24$). The questions trained during the learning phase were more successfully answered by participants ($M = 59.51\%$ and $SD = 19.57\%$) than the untrained questions ($M = 39.06\%$ and SD $= 16.84\%$). The correlation between the two types of questions was $r = 0.71$ and $p < 0.0001$. There was a trend for an interaction between the learning condition and question type variables ($F(2,564) = 2.70$, $p = 0.068$, and $\eta^2 = 0.007$).

To answer our research questions, we conducted separate one-way ANOVA analyses for trained, new, and generalisation questions as dependent variable, and the three learning conditions as factor. Furthermore, considering the relatively low number of new and generalisation questions, and the absence of significant difference between them, we also grouped them in a post-hoc analysis of the category "untrained questions". Thus, the

Table 2. Time spent on the training in log(min) and effect sizes for the planned *t*-test comparisons between the different groups of participants

Learning conditions			Planned *t*-test comparisons		
Reading-reading	Reading-quiz	Quiz-reading	Reading-quiz vs. Reading-reading	Quiz-reading vs. Reading-quiz	Quiz-reading vs. Reading-reading
2.87	3.59	3.76	$d = 1.94$ [1.59, 2.29]	$d = 0.35$ [0.05, 0.64]	$d = 1.89$ [1.54, 2.24]
(0.35)	(0.40)	(0.47)	$p < 0.0001$	$p = 0.011$	$p = 0.013$

Standard deviations are in parentheses. Effect sizes (Cohen's *d*) are given with 95% confidence intervals

Table 3. Final performance for each learning condition (reading-reading, reading-quiz, and quiz-reading) to the trained and untrained questions (new, generalisation, and total untrained) and across questions

Learning conditions			
	Reading-reading	Reading-quiz	Quiz-reading
Trained questions	52.87% (17.94%)	66.15% (19.94%)	59.51% (18.79%)
New questions	39.04% (17.44%)	43.73% (20.64%)	37.13% (16.86%)
Generalisation questions	35.37% (18.44%)	41.68% (21.72%)	37.16% (20.19%)
Untrained questions (total)	37.29% (15.2%)	42.76% (18.61%)	37.14% (16.08%)
All questions	45.91% (15.82%)	55.70% (17.88%)	49.52% (16.03%)

Standard deviations are in parentheses

Table 4. Effect sizes for the planned *t*-test comparisons between the different conditions (the post-testing effect, the effect of quiz position, and the pre-testing effect) to the trained and untrained questions (new, generalisation, and total untrained) and across questions

	Planned *t*-test comparisons		
	Reading-quiz vs. Reading-reading	Reading-quiz vs. Quiz-reading	Quiz-reading vs. Reading-reading
Trained questions	$d = 0.74$ [0.44, 1.04]	$d = 0.34$ [0.05, 0.63]	$d = 0.35$ [0.07, 0.64]
	$p < 0.0001$	$p = 0.019$	$p = 0.013$
New questions	$d = 0.27$ [−0.02, 0.56]	$d = 0.35$ [0.06, 0.64]	$d = -0.11$ [−0.40, 0.17]
	$p = 0.092$	$p = 0.017$	$p = 0.44$
Generalisation questions	$d = 0.34$ [0.05, 0.63]	$d = 0.22$ [−0.07, 0.50]	$d = 0.09$ [−0.20, 0.37]
	$p = 0.032$	$p = 0.14$	$p = 0.52$
Untrained questions (total)	$d = 0.36$ [0.07, 0.66]	$d = 0.32$ [0.03, 0.61]	$d = -0.01$ [−0.29, 0.27]
	$p = 0.028$	$p = 0.027$	$p = 0.95$
All questions	$d = 0.62$ [0.32, 0.92]	$d = 0.36$ [0.07, 0.65]	$d = 0.22$ [−0.06, 0.51]
	$p < 0.0001$	$p = 0.013$	$p = 0.12$

Effect sizes (Cohen's *d*) are given with 95% confidence intervals

two dependent variables—trained and untrained question scores—were analysed as the within-subject variable Question Type. Tables 3 and 4 and Fig. 1 summarise final recall for each question type (trained and untrained) and learning condition, and post-hoc analysis for each one-way ANOVA.

Trained questions. Consistent with predictions, there was a main effect of the learning condition, $F(2,282) = 11.76$, $p < 0.0001$, and $\eta^2 = 0.077$. As shown in Table 4, planned *t*-test comparisons showed a post-testing effect, a pre-testing effect, and most interestingly, a position effect: retention was significantly higher for the reading-quiz group compared with the quiz-reading group (Table 3).

Untrained "new" questions. There was a main effect of the learning condition, $F(2,282) = 3.24$, $p = 0.041$, and $\eta p^2 = 0.022$. Planned *t*-test comparisons did not show a significant post-testing effect or a pre-testing effect (Table 4). However, there was a significant position effect: the reading-quiz group had better final performance than the quiz-reading group (Table 3).

Untrained "generalisation" questions. No main effect of the learning condition was found, $F(2,282) = 2,48$, $p = 0.086$, and $\eta p^2 = 0.017$. For information, effect sizes for planned *t*-test comparisons are presented in Table 4.

Untrained questions (new and generalisation). There was a main effect of the learning condition, $F(2,282) = 3.49$, $p = 0.032$, and $\eta^2 = 0.024$. As shown in Tables 3 and 4, comparison of final test performance on this question type revealed a significant advantage for the reading-quiz group on both reading-reading (post-testing effect) and quiz-reading groups (position effect).

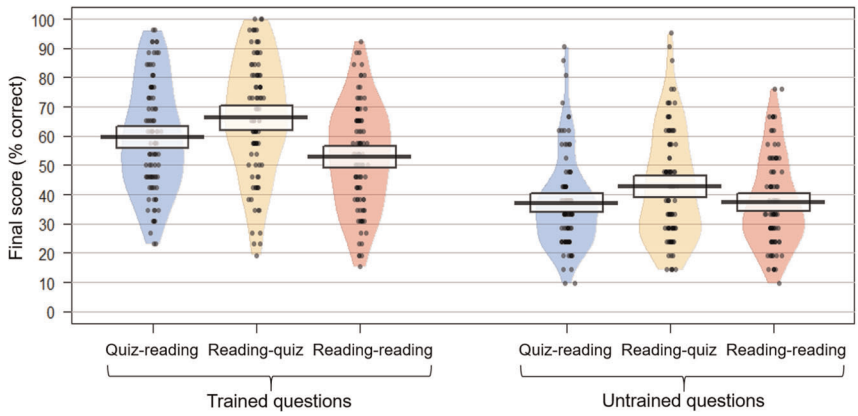


Fig. 1 Final test performance to the trained and untrained questions (percentage correct answers) and for each group of participants. Each point is a participant. The thick horizontal line represents the median, colour-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Interval

However, final test performance for the untrained information in the quiz-reading group was not different from the performance in the reading-reading group (no pre-testing effect).

Covariates and exploratory analyses

In the multiple linear regression model analysis restricted to the two learning conditions with quizzes, we added the total number of attempts to take the quizzes and the log of the time spent on training as covariates. The total number of attempts to take the quizzes had an influence on the final test score, $F(1,352) = 4.82$, $p = 0.03$, and $\eta^2 = 0.013$. However, the influence of the log (training time) was not significant, $F(1,352) = 2.01$, $p = 0.157372$, and $\eta^2 < 0.01$. Main effects of the learning condition (reading-quiz vs. quiz-reading) and of the question type (trained vs. untrained) were still significant ($F(1,352) = 8.86$, $p < 0.01$, and $\eta^2 = 0.014$; and $F(1,352) = 140.83$, $p < 0.0001$, and $\eta^2 = 0.28$), but not the interaction $F(1,352) = 0.10$, $p = 0.75$, and $\eta^2 < 0.001$.

Then, in the full multiple linear regression model (three learning conditions × two question types), adding log(training time) had no significant effect ($F(1,547) = 0.33$, $p = 0.56$, and $\eta^2 < 0.001$). The main effects of the learning condition and of the question type (trained vs. untrained) remained significant ($F(2,547) = 13.63$, $p < 0.0001$, and $\eta^2 = 0.02$; and $F(1,547) = 180.43$, $p < 0.0001$, and $\eta^2 = 0.24$). There was a trend for a significant interaction ($F(2.547) = 2.76$, $p = 0.06$, and $\eta p^2 < 0.01$).

Finally, and as an exploratory analysis, we added the participants' age and degree of knowledge about DNA as covariates in the full linear model too. Age had a significant influence on final test score such as when age increased, final test performances tended to decrease ($F(1,460) = 8.94$, $p < 0.01$, and $\eta^2 < 0.01$); as expected, the degree of knowledge about DNA had a significant positive effect ($F(1,460) = 28.05$, $p < 0.01$, and $\eta^2 = 0.04$). Like in the initial analyses, the learning condition and question type factors remained significant ($F(2,460) = 13.21$, $p < 0.0001$, and $\eta^2 = 0.05$; and $F(1,460) = 153.37$, $p < 0.0001$, and $\eta^2 = 0.23$). The interaction was not significant, $F(2,460) = 2.21$, $p = 0.11$, and $\eta^2 < 0.01$.

DISCUSSION

The results from this experiment provide insights into the effective placement of retrieval practice relative to studying the content as well as replicate results from previous research with robust data. First, significant testing effects (in comparison to reading-reading time) were found when quizzes were placed either after (post-testing effect) or before reading the same contents (pre-testing effect), in line with previous findings.[26,29] Furthermore, the post-testing effect was significantly larger than the pre-testing effect. Finally, whereas post-testing increased the retention of related but untrained material, pre-testing did not. These results were obtained for the learning of prose passages of scientific content, in a digital learning environment, and with a retention interval of 7 days.

Regarding the post-testing effect, we replicated the same large effect size ($d = 0.62$ across question types) as found by Rowland[11] in his meta-analysis of between-subject experiments ($g = 0.69$). This effect was particularly strong for the trained questions ($d = 0.74$), and smaller but significant for the untrained questions ($d = 0.32$). Thus, initial testing led to enhanced recall for untrained but related questions. This latter result is consistent with that of Chan et al.,[14] although with a smaller effect size (in their Experiment 1, they obtained $d = 0.69$ for the testing group on untested questions relative to the reading-reading group, and $d = 0.56$ relative to the control group). The benefit of post-testing practice on untrained questions seems to be driven by generalisation rather than new questions (whose answer was present in the learning material). This is in line with a recent meta-analysis on the transfer of learning:[16] the authors found no evidence for transfer of the testing effect to untested materials seen during initial study (similar to our new question type), but they found an overall positive transfer on application and inference questions (similar to our generalisation question type).

Regarding the pre-testing effect, we found a smaller but significant effect on trained questions ($d = 0.35$) but not on untrained questions ($d = -0.01$). This is rather lower than in previous recent studies. In Richland et al.'s study,[22] participants had a very low success rate in their pre-test sessions (around 5%), which provided a huge margin of improvement. In contrast, in our experiments, the mean percentage correct during the learning phase was 40% in the quiz-reading condition, which may indicate a certain level of prior knowledge, and provides less margin of improvement (although there was no hint of a ceiling effect in the final test).

Our study replicates previous results with educational contents and huge sample size. These results are consistent not only with those of Frase[27,28] showing an advantage of post-test over pre-test location but also with previous studies that compared post-testing and combined pre- and post-testing and found little difference.[30] Similarly, a new study by Geller et al.[32] found that asking questions before having a lecture did not enhance the learning of the pre-tested information more than the learning of other information which was not pre-tested. Moreover, doing a pre-test did not boost the benefits of the post-testing effect.

What may account for the superiority of post-testing over pre-testing? First, unlike pre-testing, post-testing enables the

consolidation of previously read information. Second, in the quiz-reading condition, participants answered incorrectly and therefore received negative feedback much more frequently than in the reading-quiz condition. This decreased reward may have affected their motivation and learning, although the notion of "desirable difficulties" would have predicted the opposite.[33] Third, and in relation to this second point, participants in the quiz-reading condition had on average lower initial performance during the training phase in comparison to the reading-quiz participants. Rowland[11] found that higher initial performance may increase the magnitude of the testing effect. Thus, the difference in initial training performance between reading-quiz and quiz-reading groups might explain all or part of the performance difference in the final test. However, exploratory covariance analyses adjusting for initial test performance showed that the main group effect remained the same irrespective of initial test performance.

Interestingly, we observed no difference between the quiz-reading and the reading-reading condition on untrained questions (neither for new nor for generalisation questions), consistently with previous research.[21,22] Moreover, performance on new questions was significantly greater for post- than for pre-testing. One possible explanation is that, during the reading phase, participants in the post-testing condition were not induced to prioritise among the available information: all pieces of information were a priori equally relevant and therefore perhaps equally attended. In the pre-testing condition, however, participants may have been incited to focus their attention on the answers to pre-tested questions, to the detriment of other material that was used to create the new untrained questions for the final test. By contrast, Carpenter and Toftness[17] found a benefit of pre-testing on both tested and untested information but in the context of video-based learning. Thus, the effect of pre-testing on the learning of untested information might be dependent of the format of the learning content, and may be harmful to learning by encouraging selective attention when material is read but not listened.

A potential limitation of the present study might be that trained and untrained questions were not counterbalanced across participants. Thus, the greater performance on trained questions may be due both to training, and to the fact that untrained questions were intrinsically more difficult (as suggested by the lower performance on untrained questions even in the reading-reading condition). This would be a problem if our main goal was to compare absolute performance between trained and untrained questions. However, our interest here was rather to investigate the interaction between question familiarity and test location.

Our conclusions are also a priori limited to a certain type of material (scientific text), mode of presentation (an e-learning platform), and to a certain type of participants, i.e., workers doing a paid task, rather than students. It will be important to address the same question using other learning contents and populations, especially school children in a more ecological context. Nevertheless, our replication of the well-established post-testing, pre-testing, and transfer effects suggests that our experimental conditions produce similar results as other studies in different conditions, so we see no reason to suspect that our new result (the superiority of post- over pre-testing) would not generalise.

Contrary to previous experiments where training time was closely matched across conditions,[2,22] in the present experiment participants were free to spend as much time as they felt necessary on the contents (prose passages, quizzes, and feedback). This induced important differences in training time between conditions, with participants in the reading-reading condition spending on average less than half the time in training than participants in the two quiz conditions, and participants in the quiz-reading condition spending 18% more time on training than those in the reading-quiz condition. However, we found no correlation whatsoever between training time and final performance, and covariance analyses adjusting for training time obtained exactly the same results. Thus, our pre- and post-testing effects cannot be attributed to the extra time spent on training, and the extra time spent on pre- vs. post-testing would predict the opposite from the observed advantage of post-testing.

Finally, we were unable to exclude or control potential extra study time between the training phase and the final test. However, participants were paid for participation regardless of performance, thus they had no incentive to spend more time studying than the minimum imposed. Furthermore, there is no reason to think that participants in different groups might have invested differently in extra study. Therefore, there is no reason to think that this may have changed the pattern of results.

To conclude, our results may help refine the recommendation of the US Department of Education[31] about the use of pre-questions to foster learning in classroom settings. It is important to note that pre-questions may be used in different ways and serve different functions. Even though pre-testing did not enhance retention as much as post-testing in our experiment, there may be other benefits of asking questions before a lecture. For instance, pre-questions may be used to test whether prerequisite knowledge is acquired, and to refresh it right before new content is exposed. In that case, those so-called pre-questions actually implement post-testing of previously learned material. However, our results do not support the specific idea that pre-questions on the content of the subsequent lecture improve learning more than post-questions. Thus, current evidence suggests that testing time dedicated to enhancing retention is better spent after than before the initial exposure to learning content.

METHODS

Participants

We calculated that at least 64 participants per group were necessary in order to detect a testing effect of size $d = 0.50$[11] in a between-subjects design with a statistical power of 0.8 (alpha = 0.05, bilateral *t*-test). Because we anticipated that the pre-testing effect might be smaller than the post-testing effect, and because we wanted to compare the pre- and post-testing effects, we aimed at including 100 participants per group, thus giving us 80% power to detect an effect size of 0.4. We therefore recruited a total of 334 adult participants via a French online work platform (Foule Factory https://www.foulefactory.com/). Inclusion criteria were that participants are native French speakers and without any reported neurological or psychiatric disorders. Furthermore, participation of students and professionals in medicine or biology was discouraged in the call for volunteers and in the information letter. All participants provided written informed consent on the online platform. The study was approved by the local research ethics committee (Comité d'Ethique de la Recherche of Paris Descartes University) and the participants were paid for their participation.

Material

The experiment was entirely run online on the Didask digital learning platform (https://www.didask.com/). On Didask, each course consists of a set of modules organised in a logical order. A module is an elementary learning unit, including both learning material (text, videos, pictures, …) and a corresponding training quiz with at least five questions (multiple choice, pairwise matching, ordering, sorting into categories, …) and can last between 5 and 15 min. We used a modified version of Didask designed for experimental testing, allowing us to specify several learning conditions and to better control the learning environment, in particular the order in which the content and the quiz were presented.

Study material consisted of a course on DNA from the French curriculum for 12th grade science tracks (age range: 18 years old). Specific contents were borrowed from material provided by CNED (Centre National d'Enseignement à Distance). Texts were adapted and illustrations were added so as to create seven short texts (length between 85 and 227 words), forming a logical progression like in a textbook, and constituting seven modules in Didask. For each of the seven short texts, five or six (depending on the length of the text) multiple choice questions were created for the training phase by selecting five or six main facts from the

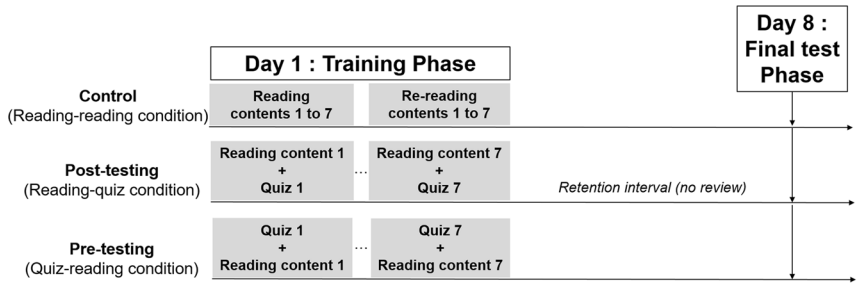


Fig. 2 Schema of the experimental procedure used in the three learning conditions

corresponding text. Thus, all the facts questioned in the training quizzes were covered in the corresponding text; but some facts included in the text passages remained unquestioned. An elaborative feedback was given immediately after each answer, indicating the correctness of the answer and providing explanations directly extracted from the corresponding text passage (feedback explanations did not exceed one or two sentences). A total of 38 multiple choice questions were created for the training session. These multiple choice questions were either single answer questions ($n = 30$) or multiple answers questions ($n = 8$) and we proposed between two and four different choices.

We also created 21 additional questions included in the final exam only, not in the training quizzes. This was to assess the transfer of the retrieval practice effects to untested questions. These untested questions were of two categories: 10 new questions that were directly related to the written information presented in the readings, and 11 generalisation questions that were more difficult, cutting across several learning modules and requiring inferences. Thus, 10 new multiple choice questions were created in the same way as the question included in the training quizzes (information present in the text passages of the learning contents). Furthermore, 11 generalisation questions were also created to measure indirect benefit of the retrieval practice. The answers were not literally included in the learning material, but required inferring or synthesising information across several texts (e.g., placing cell constituents correctly on a picture of a cell that was not included in the learning material). Finally, nine "catch" questions were interspersed throughout the training quizzes and the final test to ensure participants were paying proper attention and were not answering randomly (e.g., What is the colour of Henri IVth's white horse?), and to be used as exclusion criteria (data from participants who correctly answered fewer than seven of the nine "catch" questions during the training phase were excluded).

The final test therefore included a total of 52 questions: 26 randomly selected from the set of the training phase's questions (i.e., trained), 10 new, 11 generalisation, and 5 catch questions. Thus, the final test contained the same number of trained and untrained questions, and each module was equivalently represented.

Design and procedure

In the present study we used a mixed factorial design. This type of design usually includes at least one between subject variable and at least one within subject variable,[34] which in the case of this study were three between-subject Learning Conditions (quiz-reading, reading-quiz and reading-reading) for the acquisition phase and two within-subject Question Types (trained and untrained) for the final test.

After giving their consent and filling in a demographical questionnaire, participants were allocated alternatingly to one of the three conditions based on order of registration, and provided with an individual link to the testing platform on Didask. All learning and testing took place online, from participants' personal computers (uncontrolled settings, such as their homes or offices). The experiment consisted in a training phase and a test phase, separated by 7 days (Fig. 2).

Training phase. In the quiz-reading condition, clicking on a module first sent participants to the corresponding quiz. No time limit was imposed to answer. After answering each question, feedback immediately appeared and was displayed for at least 10 s, and until they decided to go on to the next question. After completing the last question of the quiz, participants were asked to study the learning content associated to that module, which was displayed for at least 50 s, and until they decided to go on to the next module. This sequence repeated itself from module 1 to 7 (Fig. 2).

In the reading-quiz condition, the procedure was exactly the same except that when clicking a module, the learning content first appeared for at least 50 s, then when clicking to continue the quiz with feedback started. This was repeated from module 1 to 7. These two conditions included exactly the same 38 training multiple choice questions; the only difference was the relative placement of quizzes and reading. For both conditions with testing, the training phase for all seven modules lasted about 30 min in total. No additional testing or reading of the material took place between this training phase and the final test phase.

In the reading-reading condition, only the learning content was presented in each module. The participants had to go from the first to the seventh module to study each content for at least 50 s. Once they had finished the 7th module, they were sent again to the 1st module for a second study phase under the same conditions. This learning condition lasted about 15 min in total. Because durations differed between the three conditions, they were recorded and taken into account for data analysis (i.e. time on training phase did not influence the main effects).

Participants were paid for the training phase only after successful completion of the training procedure. The instructions required the participants to go through each module only once (or twice in the reading-reading condition), however for technical reasons we were not able to block extra uses of each module. We therefore recorded the total number of module visits in order to take them into account for data analysis (see Covariates and exploratory analyses in "Results" section). The learning phase was open to new participants for about 28 h, from the time we published the advertisement to the time we obtained 334 participants and decided to close the task on Foule Factory. The access to the training space on Didask was closed after the last participant completed the learning phase. Participants were instructed not to study more about DNA before the test phase. This was of course impossible to control, but there was no incentive for further study, given that payment was a flat rate for participation, regardless of performance in the final test.

Final test phase. Participants who completed the learning phase correctly were asked to participate in the final test phase for an additional payment and 7 days after the learning phase (mean retention interval = 6.91 days (SD = 0.20; range = 6.27–7.59 days)). Questions that were presented in this final test were the same and in the same order for all participants. There was no time limit to complete the final test. It took about 15 min. No feedback was given after any of the 52 questions, but at the end of the test participants were given their total score.

In both learning and test phases, each question was scored 1 if the answer was entirely correct and 0 when any error was made. Percentage correct answers across all seven modules were then analysed.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The preregistration for this experiment can be accessed at http://aspredicted.org/blind.php?x=wg7nj6. An initial version of the preregistration planned final test at both 7 and 28 days post training. However, during the preparation of the experiment, we decided to drop the 28-day test, hence the new version of the preregistration. De-identified data, and materials are posted on Dataverse website with the link:[35] https://doi.org/10.7910/DVN/XPYPMF.

CODE AVAILABILITY

Data analysis scripts for this experiment along with a code-book are posted on Dataverse website with the following link:[35] https://doi.org/10.7910/DVN/XPYPMF.

ACKNOWLEDGEMENTS

We acknowledge funding from Programme d'investissements d'avenir (Efran programme), Agence National de la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*).

AUTHOR CONTRIBUTIONS

A.L. and F.R developed the study concept and design in collaboration with S.T.L. The experimental conditions were programmed on Didask by A.R. Testing and data collection were performed by A.L. and A.R. under the supervision of F.R. Data analysis and interpretation were performed by A.L., H.P. and F.R. A.L. drafted the paper; F.R., H. P. and R.C. provided critical revisions. All authors approved the final version of the paper for submission.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Science of Learning* website (https://doi.org/10.1038/s41539-019-0053-1).

Competing interests: A.R. and S.T.L. are shareholders of Didask. However, they agreed to preregister the study and to publish the results regardless of the outcome. They intervene neither on the analysis of the results nor on the conclusions of the study.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Bangert-Drowns, R. L., Kulik, J. A. & Kulik, C.-L. C. Effects of frequent classroom testing. *J. Educ. Res.* **85**, 89–99 (1991).
2. Roediger, H. L. & Karpicke, J. D. Test-enhanced learning taking memory tests improves long-term retention. *Psychol. Sci.* **17**, 249–255 (2006).
3. Spitzer, H. F. Studies in retention. *J. Educ. Psychol.* **30**, 641–656 (1939).
4. Zaromb, F. M. & Roediger, H. L. The testing effect in free recall is associated with enhanced organizational processes. *Mem. Cogn.* **38**, 995–1008 (2010).
5. McDaniel, M. A., Fadler, C. L. & Pashler, H. Effects of spaced versus massed training in function learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **39**, 1417–1432 (2013).
6. Vaughn, K. E., Rawson, K. A. & Pyc, M. A. Repeated retrieval practice and item difficulty: does criterion learning eliminate item difficulty effects? *Psychon. Bull. Rev.* **20**, 1239–1245 (2013).
7. Carpenter, S. K., Pashler, H., Wixted, J. T. & Vul, E. The effects of tests on learning and forgetting. *Mem. Cogn.* **36**, 438–448 (2008).
8. McDaniel, M. A., Wildman, K. M. & Anderson, J. L. Using quizzes to enhance summative-assessment performance in a web-based class: an experimental study. *J. Appl. Res. Mem. Cogn.* **1**, 18–26 (2012).
9. Roediger, H. L. III, Agarwal, P. K., McDaniel, M. A. & McDermott, K. B. Test-enhanced learning in the classroom: Long-term improvements from quizzing. *J. Exp. Psychol. Appl.* **17**, 382–395 (2011).
10. Karpicke, J. D. & Aue, W. R. The testing effect is alive and well with complex materials. *Educ. Psychol. Rev.* **27**, 317–326 (2015).
11. Rowland, C. A. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* **140**, 1432–1463 (2014).
12. Adesope, O. O., Trevisan, D. A. & Sundararajan, N. Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* https://doi.org/10.3102/0034654316689306 (2017).
13. Butler, A. C. Repeated testing produces superior transfer of learning relative to repeated studying. *J. Exp. Psychol. Learn. Mem. Cogn.* **36**, 1118–1133 (2010).
14. Chan, J. C. K., McDermott, K. B. & Roediger, H. L. Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *J. Exp. Psychol. Gen.* **135**, 553–571 (2006).
15. Karpicke, J. D. & Blunt, J. R. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* **331**, 772–775 (2011).
16. Pan, S. C. & Rickard, T. C. Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychol. Bull.* **144**, 710–756 (2018).
17. Carpenter, S. K. & Toftness, A. R. The effect of prequestions on learning from video presentations. *J. Appl. Res. Mem. Cogn.* **6**, 104–109 (2017).
18. Hartley, J. The effect of pre-testing on post-test performance. *Instr. Sci.* **2**, 193–214 (1971).
19. Little, J. L. & Bjork, E. L. Multiple-choice pretesting potentiates learning of related information. *Mem. Cogn.* **44**, 1085–1101 (2016).
20. Peeck, J. Effect of prequestions on delayed retention of prose material. *J. Educ. Psychol.* **61**, 241–246 (1970).
21. Pressley, M., Tanenbaum, R., McDaniel, M. A. & Wood, E. What happens when university students try to answer prequestions that accompany textbook material? *Contemp. Educ. Psychol.* **15**, 27–35 (1990).
22. Richland, L. E., Kornell, N. & Kao, L. S. The pretesting effect: do unsuccessful retrieval attempts enhance learning? *J. Exp. Psychol. Appl.* **15**, 243–257 (2009).
23. Sagaria, S. D. & Di Vesta, F. J. Learner expectations induced by adjunct questions and the retrieval of intentional and incidental information. *J. Educ. Psychol.* **70**, 280–288 (1978).
24. Grimaldi, P. J. & Karpicke, J. D. When and why do retrieval attempts enhance subsequent encoding? *Mem. Cogn.* **40**, 505–513 (2012).
25. Warr, P. B., Bird, M. & Rackham, N. *Evaluation of Management Training: A Practical Framework, with Cases, for Evaluating Training Needs and Results* (Gower P, London, 1970).
26. Anderson, R. C. & Biddle, W. B. in *Psychology of Learning and Motivation* (ed. Bower, G. H.) On asking people questions about what they are reading. 9, 89–132. (Academic Press, New York, 1975).
27. Frase, L. T. Learning from prose material: length of passage, knowledge of results, and position of questions. *J. Educ. Psychol.* **58**, 266–272 (1967).
28. Frase, L. T. Effect of question location, pacing, and mode upon retention of prose material. *J. Educ. Psychol.* **59**, 244–249 (1968).
29. Hamaker, C. The effects of adjunct questions on prose learning. *Rev. Educ. Res.* **56**, 212–242 (1986).
30. McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B. & Roediger, H. L. Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *J. Educ. Psychol.* **103**, 399–414 (2011).
31. Pashler, H. et al. Organizing instruction and study to improve student learning. IES Practice Guide. NCER 2007–2004. *Natl Cent. Educ. Res.* (2007).
32. Geller, J. et al. Prequestions do not enhance the benefits of retrieval in a STEM classroom. *Cogn. Res. Princ. Implic.* **2**, 42 (2017).
33. Bjork, E. L. & Bjork, R. A. in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (eds Gernsbacher, M. A., Pew, R. W., Hough, L. M. & Pomerantz, J. R.) Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. 56–64. (Worth Publishers, New York, 2011).
34. Howell, D. C. *Statistical Methods for Psychology*. (Cengage Learning, 2012).
35. Latimier, A. et al. Does pre-testing promote better retention than post-testing? https://doi.org/10.7910/DVN/XPYPMF (2019).

4 Chapitre 4 : “Retrieval practice promotes long-term retention irrespective of grain size”

This chapter constitutes the following manuscript: Latimier, A., Riegert, A., Ly, S.T., Ramus, F. Retrieval practice promotes long-term retention irrespective of grain size (in prep).

The preregistration for this experiment can be accessed at:
http://aspredicted.org/blind.php?x=zp2486

Learning materials[25], de-identified data, and data analysis scripts are posted on OSF website with the following link:
https://osf.io/vya9u/?view_only=9db6896452cb42be83107beddbade3ff

We aknowledge the start-up company Cog’X for providing the full access to their training contents to create the learning material for this experiment.

4.1 Abstract

Using retrieval practice instead of re-reading the same information during learning enhances long-term retention. There is a growing literature on interpolated retrieval practice that consists in interspersing retrieval practice throughout the learning contents instead of concentrating it at the end of the reading. While interspersed testing promotes better performance during the training phase, results in laboratory and classroom settings suggested that it has little impact on retention, regardless of the retention interval. The present study aimed at comparing the effect of different grain sizes of learning contents on memory retention, depending on whether retrieval practice or re-reading was used, and on retention interval (one-week vs one month). Our experiment was entirely run online on a digital learning platform and used professional training contents. We replicated the testing effect on retention at both short and long intervals, but contrary to predictions, we did not find that overall performance was different between the different grain sizes of learning periods. However, a significant interaction between learning strategy and grain size suggested that grain size matters particularly when learning with successive readings: small reading grains led to better retention than longer readings. However, when learning with retrieval practice, grain size did not affect long-term retention: interspersed quizzes throughout small readings is equivalent to postponed quizzes after a single long reading.

[25] Une partie du contenu à lire ainsi que des exemples de questions d’apprentissage sont fournis en dans la section « Annexes » (Annexe 2)

4.2 Introduction

Research on learning strategies demonstrated robust benefits of retrieval practice (RP). For instance, taking a test during the process of learning new information promotes superior long-term retention compared to passively reading: this is also known as the *testing effect* (Adesope, Trevisan, & Sundararajan, 2017; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Karpicke, 2006; Rowland, 2014). The benefits of RP has been found mostly in conditions where learners take tests after being exposed to learning contents (e.g., word list, paired words, prose passages, and more complex text materials); but a pre-testing effect has been demonstrated too, although it may be weaker (Carpenter, Rahman, & Perkins, 2018; Carpenter & Toftness, 2017; Latimier et al., 2019; Richland, Kornell, & Kao, 2009). The current way of investigating the RP effect is to give learners some content to study, and offer them the opportunity to retrieve part of this content as a reviewing opportunity in order to do a final retention test (either immediate or delayed).

There is an abundant literature on interpolated RP, which consists in interleaving training questions with the study material. Several studies showed that interpolated questions appeared to promote significantly higher learning performance than text without any (i.e., reading only or restudying; Reynolds, Standiford, & Anderson, 1979; Watts & Anderson, 1971). Apart from this typical *backward* RP effect, some studies have also shown a strong *forward* testing effect. Namely, doing RP on previously studied content facilitates learning of subsequent content with word lists as well as text materials (for a review see Yang, Potts, & Shanks, 2018). Using interpolated RP brings the benefits associated with retrieval practice relative to restudying, but may also have more specific effects, such as avoiding proactive interference during the study of the learning content, reducing the occurrence of mind wandering, while increasing the frequency of note taking and maintaining learners' attention (Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013).

The literature on interpolated RP versus ready only learning leads to the question of the optimal placement of RP episodes relative to the reading episodes to promote long-term retention. It is especially relevant with lengthy and complex learning content. So far, two different placements of RP episodes have been compared. The first placement is interpolated RP as described above, in which RP questions are interspersed throughout the learning contents; while the second placement is postponed RP questions in which RP questions were concentrated in second time after reading the whole learning contents. Researchers who investigated the question of the optimal placement of RP hypothesized that learners' final retention should be better when learning small chunk of material interpolated with RP because this enables a sustainable

attentional focus and engagement on the content (Duchastel & Nungester, 1984; Healy, Jones, Lalchandani, & Tack, 2017; Uner & Roediger, 2017; Weinstein, Nunes, & Karpicke, 2016; Wissman & Rawson, 2015). It has also been argued that retrieval success was higher when learning small amount of text with small RP episodes thus retrieval success should persist at a retention test too (Wissman, Rawson, & Pyc, 2011). Moreover, retrieving information at the end of each small reading segment may consolidate retention of the previously learned information while it boosts the encoding of subsequent information (i.e., *forward* RP effect).

Postponed RP might also have benefits. First, this placement emulates in a sense spaced practice which enhances long-term retention (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Wissman & Rawson, 2015). While interpolated RP emulates massed retrieval strategy, postponed RP increases the space between the study of key concepts and the moment retrieving them with a certain lag depending on the length and complexity of the learning content. Moreover, postponed RP might enable a better apprehension of the contents as a whole without any interruption of the study, which could be advantageous to link the main ideas with each other through the reading. By contrast, frequent switching between RP events and encoding phase of new related content might impair the retention of this new learning and induces a sort of effort due to "switch cost" (Pashler et al., 2000).

Given that each type of RP placement may have different benefits on retention, it particularly is interesting to compare both placements in the same experiment. Duchastel and Nungster (1984) directly compared interpolated versus postponed questions and found that retention of short text's items were equally promoted by the two types of question placements at a delayed final test 2 weeks later. In the context of learning brief text passages, Wissman & Rawson (2015) replicated these results at two different retention intervals (15 or 20 min versus 2 days) and different formats of final tests (free recall, cued recall, or recognition). In formats in seven experiment, the authors did not observe a superiority of the interpolated RP over postponed RP, although there was a benefit of the interpolated RP on the training performances. Interestingly, recall for main ideas was greater during the training phase in the interpolated RP condition, but this effect was washed away by the retrieval of unimportant details at the final test. Weinstein et al.'s study (2016) led to the same conclusion. Namely, one week after the training phase, there was no benefit of interpolated RP in a small scale laboratory study ($d=0.06$, $p=.85$) nor for a large scale study done outside the laboratory ($d=0.14$, $p= 0.44$). Finally, the authors mimicked a real word learning context by using live lectures and assessing participants' memory after a very long retention interval (41 days on average). Similar to their two other

experiments, no difference in long term retention was found between the two RP placements (d=0.06, p = 0.69). Uner and Roediger (2017) also compared the two placements of RP with educational contents (i.e., a biology chapter 40 pages long). Their experiment consisted in self-paced learning. Participants in the RP conditions trained on short answer (SA) questions with feedback following three different placements: some SA were interpolated, some SA were postponed, and some SA were presented on both occasions (these SA were thus trained twice). They designed an equivalent condition with restudy episodes following the same placements manipulated as a within subject variable, and they added a single reading condition as a baseline. Obviously, they found a RP effect, with the single reading condition yielding the worst final retention 2 days later, and the restudy conditions yielding intermediate final performance. Results on final retention restricted to the RP conditions showed again no difference between the interpolated and the postponed questions, while questions presented on both occasions showed the best final recall.

To summarize, these laboratory and classroom experiments, including some with large samples and long retention intervals, did not find any difference between interpolated and postponed placements of RP episodes. However, to our knowledge, one recent study found a significant advantage of the interpolated RP placement on final retention (Healy et al., 2017). In the context of artificial facts learning, the authors demonstrated that participants recalled more information when they read a text with interpolated RP, even when the final test interval was expanded to one week (Experiment 1). They also found a benefit of interpolated RP, not only on trained but also on untrained items, at short term and long term retention intervals (but this benefit was greater at the immediate final test than at the delayed one). They explained this benefit by the fact that interpolated RP enhanced the mean engagement ratings of the learners during the initial phase (i.e., more effort and motivation but less boredom, fatigue, and mind wandering, Experiment 2). One big difference between this study in which there was a significant superiority of the interpolated RP condition and the previous studies that found no benefit is the relation between the different sections of the learning content. In Healy et al's experiments (2017), participants studied eight facts about a plant category followed by a MCQ quiz associated to the previously studied facts ant then moved to the next facts about a new plant category. Thus, the different sections of the learning content were not connected to each other in contrast with the four similar studies. In this learning situation that emulated blocked practice of independent facts (by contrast with interleaved practice, e.g.; Carpenter & Mueller, 2013), participants did not need to make connections between the different sections that were not

related to each other and thus might have taken advantage of learning with quizzes placed right after the relevant facts.

Overall, these results showed that the simple question of the placement of the RP episodes is quite limited. Nevertheless, this literature raised the more specific question of the optimal grain size of the interpolated RP episodes: at what precise moment of the reading content should the RP questions be interpolated? In their experiments, Duchastel and Nungster (1984) and Weinstein et al. (2016) used factual questions during the learning phase and only one question was asked at the end of each section in the interpolated RP condition. This might not be enough to elicit of benefit of interpolated RP over postponed RP. Wissman & Rawson (2015) used free recall episodes interspersed within four sections of lengthy learning contents while Uner and Roediger (2017) used short answer interspersed within eighteen sections of a course with four SA in the interpolated RP condition. Thus, the difference between the two types of RP placements might be a matter of size of the chunks in the interpolated condition.

4.3 The present study

In line with previous studies, the present study aimed at assessing the effect of the RP placement on final retention, by comparing interpolated versus postponed RP, relative to a control condition that consisted in interpolated versus postponed re-reading. Learning involved lengthy text passages on real-world complex knowledge, studied on an online platform. In an attempt to understand further the optimal way of interpolating RP**,** we investigated two distinct sizes of grains between which RP could be interpolated. Thus, this study was also an investigation of the optimal grain size of RP. We compared three grain size conditions: small, medium, and large (with the largest grain size being equivalent to postponed placement). Finally, we assessed long-term retention of both trained and untrained material, in order to investigate to what extent the RP grain sizes might affect learning transfer; and at two retention intervals (one weeks and four weeks).

First, we hypothesized that retrieval practice conditions promote better retention that restudy conditions (i.e., main effect of RP). Second, we hypothesized that final retention might differ depending on grain size (main effect of grain size). Third, we hypothesized that the RP effect might depend on grain size, as manifested by an interaction between grain size and retrieval practice. In the latter two cases, given the previous state of the literature, we made no specific prediction regarding the direction of the difference. Finally, for exploratory purposes, we

compared training performance between the three grain sizes in RP conditions; and we investigated the influence of various covariates (i.e., training times, demographical data).

4.4 Methods

Participants

Based on previous studies, we calculated that at least 64 participants per group were necessary in order to detect a testing effect of size d = 0.50 (Rowland, 2014) with a statistical power of 0.8 (alpha = 0.05, two-tailed t-test). Because we anticipated that the grain size effect might be more subtle than the testing effect, we wanted to include around 100 participants per group (six in total). We therefore be able to recruit a total of 380 adult participants via a French online work platform (Foule Factory https://www.foulefactory.com/). Inclusion criteria were that participants be native French speakers and without any reported neurological or psychiatric disorders. Furthermore, participation of students and professionals in cognitive sciences, ergonomics or in occupational medicine was discouraged in the call for volunteers and in the information letter. All participants provided written informed consent on the online platform. The study was approved by the local research ethics committee (Conseil d'Évaluation Éthique pour les Recherches En Santé of Paris Descartes University) and the participants were paid for their participation.

Material

The experiment was entirely run online on the Didask digital learning platform (https://www.didask.com/), which allows to present both learning material and corresponding quizzes. We used a modified version of Didask designed for experimental testing, allowing us to specify several learning conditions and to better control the learning environment, in particular the grain size of the learning contents and associated quizzes.

Study material consisted of a course on cognitive balance at work provided by the company Cog'X. This training addresses three sub-themes: cognitive overload, mental fatigue, and attentional issues at work. The learning material consisted of short texts and comic-like illustrations with corresponding quizzes. These were adapted for this experiment, progressing logically like in a textbook. From the three sub-themes of the training six small chapters of learning content were created for the training phase (we divided into two the contents about each sub-themes). Overall, the learning contents consisted in 42 slides made of comic-like illustrations and short text passages (between 80 and 150 words). When divided into six small

chapters, the number of slides per chapter varied from 5 to 10 slides, and the number of multiple choice questions (MCQ) for the six associated quizzes varied from 6 to 8 MCQ.

The quizzes could take the form of multiple choice, pairwise matching, ordering, or sorting into categories. Thus, all the facts questioned in the training quizzes were covered in the corresponding text and comics. A total of 45 MCQ were created for the training session : 39 trained MCQ and 6 "catch" questions that were interspersed throughout the training quizzes to ensure participants were paying proper attention and were not answering randomly, and to be used as exclusion criteria (e.g., "What is the color of Henri IVth's white horse ?"). An elaborative feedback was given immediately after each answer. Each answer was scored 1 if it was entirely correct (thus, when multiple answers were expected, all of them had to be answered correctly), and 0 when any error was made. In addition, the feedback provided explanations directly extracted from the corresponding text materials (feedback explanations did not exceed 4 or 5 sentences). Appendix 1 shows the 6 topics covered by the learning phase called "*Learn how to listen to your brain at work*".

For the final test phase, 24 new additional MCQ were created in the same way, thus asking about information that had been present in the learning content. These questions were not included in the training quizzes and thus were untrained questions. The final tests therefore included a total of 69 MCQ: 39 trained, 24 untrained, and 6 catch questions.

Design and Procedure

We used a mixed factorial design that included, for the training phase, 2 between-subject Learning Conditions (reading-quiz, reading-reading) and 3 between-subject Grain Sizes (small, medium, large) for the training phase. For the final test phase, we manipulated 2 within-subject Question Types (trained, untrained) and 2 within-subject Retention intervals (7 days, 28 days). After giving their consent and filling in a demographic questionnaire, participants were allocated alternatingly to one of the six training conditions based on order of registration, and provided with an individual link to the testing platform on Didask. All learning and testing took place online, from participants' personal computers. The experiment consisted of a training phase and two test phases, one separated by 7 days, and one separated by 28 days from the training phase (Figure 1).

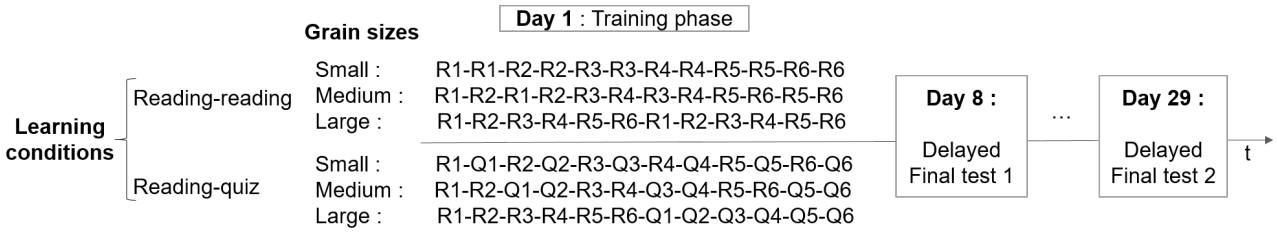


Figure 1. Schema of the experimental procedure, R is for "reading" and Q is for "quiz".

Training phase - In the reading-quiz conditions, participants were first sent to learning contents (texts and illustrations) and then to the corresponding quizzes. In the small grain size condition, participants had to study the content and to do the quiz for one skill at a time, thus leading to an alternation of six study-quiz periods. In the medium grain size condition, the skills and corresponding quizzes were merged by two (i.e.; chapters 1+2 and quizzes 1+2, then chapters 3+4 and quizzes 3+4, and so on). Thus, chapter 1 consisted in 15 slides, chapter 2 in 14 slides, and chapter 3 in 13 slides. Each of the three associated quizzes consisted in ten to fourteen MCQ. Finally, in the large grain size condition, the six learning contents were merged together as a whole skill called « course » and the quizzes were also merged as a whole quiz. The participants included in the condition read the 42 slides of the course and then practiced with the 39 MCQ.

No time limit was imposed to answer the questions. After answering each question, feedback immediately appeared and was displayed for at least 10 secs, and until the participants decided to go on to the next question. In the reading-reading conditions, only the learning contents were presented, with a re-studying episode with exactly the same learning content instead of a quiz. In the small grain size condition, participants had to study the content and to study it again immediately for one skill one at a time, thus leading to an alternation of six study-restudy periods. In the medium grain size condition, the learning contents were merged by two (i.e.; chapters 1+2, chapters 3+4, chapters 5+6) thus leading to an alternation of three study-restudy periods. In the large grain size condition, the six learning contents were merged together as a whole skill called « course » and the participants had to study this sizeable course twice.

When a skill was finished, it was locked in a way so it was not possible for the participant to access it again (a tree that grew after completion of each skill visualized progress). The exposure to the contents thus was identical for all participants across conditions. We expected that the training durations differed between conditions and lasted between 15 minutes and 1 hour in total (*a posteriori* analysis on training times are indicated in results sections). Therefore, these differences were taken into account for data analysis.

Participants were paid for the training session only after successful completion of the training procedure. The training phase was open to new participants for about 28 hours, from the time we published the ad to the time we obtained enough participants and decided to close the task on Foule Factory. Because it is technically not possible to recruit more than 600 participants in one shot, we had to recruit participants in successive waves of 150 participants until we had a satisfactory number of participants per condition.

Participants were instructed not to study more before the test phase. This was of course impossible to control, but there was no incentive for further study, given that payment was a flat rate for participation, regardless of performance in the final test.

Final test phase - Participants who completed the training phase properly were asked to participate in the two final tests for an additional payment 7 days and 27 days after the training phase (mean retention interval final test 1= 7.10 days with SD = 0.75 and range = 4.95-9.02 days; mean retention interval final test 2= 27.61 days with SD = 1.11 and range = 22.26-31.92 days). There was no time limit to fill in the final test. It took about 30 min to fill in entirely each final test. No feedback was given after any of the 69 MCQ, but at the end of the test participants were given their total score. Percentage of correct answers were then analyzed.

4.5 Results

Participants' data

From a total of 380 participants who undertook the initial learning phase, 30 participants were excluded because: 1) they did not complete the learning phase, or 2) they did not complete final test 1, or 3) they did not take the learning phase and the final tests in the prescribed timeframe, or 4) they did not answer correctly the catch questions (participants who correctly answer fewer than 10 of the 12 "catch" questions). Out of those 350 participants, 46 (13.14%) gave up participation between the learning phase and final test 1, and 43 participants (14.63%) gave up participation between final test 1 and final test 2 (total attrition: 99 participants, 28.29% of the original sample). Thus, we had a total of 294 participants included in the analysis for final test 1 and 251 for final test 2. Demographic data are indicated in Table 1. The 6 groups did not differ in age ($F(5,284) = 1.37$, $p = 0.24$ for final test 1, $F(5,243) = 0.90$, $p = 0.48$ for final test 2), in education level ($F(5,288) = 0.76$, $p = 0.58$ for final test 1; and $F(5,245) = 0.44$, $p = 0.82$ for final test 2), and in sex ratio ($\chi^2(5) = 5.86$, $p = 0.43$ for final test 1, $\chi^2(5) = 7.87$, $p = 0.16$ for final test 2).

Table 1: Summary characteristics of the six groups of participants included in final test 1 and in final test 2.

	Learning groups						
FINAL TEST 1	**R-R small grain**	**R-R medium grain**	**R-R large grain**	**R-Q small grain**	**R-Q medium grain**	**R-Q large grain**	**Total**
N	48	52	56	51	51	38	294
Sex (M/F)	17/31	15/37	20/34	15/36	10/41	11/27	88/205
Age (years)	41.92 (11.75)	38.92 (10.87)	38.89 (11.58)	36.17 (10.60)	40.34 (11.43)	39.32 (12.02)	39.23 (11.39)
Education (years)	13.85 (2.62)	14.35 (2.47)	14.30 (2.47)	13.92 (2.48)	14.60 (2.27)	14.47 (2.03)	14.24 (2.41)
FINAL TEST 2							
N	41	45	42	44	46	33	251
Sex (M/F)	15/26	12/33	17/25	11/33	8/38	11/22	74/177
Age (years)	41.76 (12.16)	39.46 (11.06)	39.26 (12.30)	36.66 (10.65)	40.28 (11.93)	40.04 (11.00)	39.53 (11.53)
Education (years)	14.07 (2.67)	14.27 (2.43)	14.49 (2.62)	14.00 (2.18)	14.61 (2.28)	14.30 (2.00)	14.29 (2.36)

Note. Standard deviations are in parentheses.

Initial learning phase

Training time - We computed the total time spent in the training phase (i.e.; the cumulative time spent on reading the contents and on training with quizzes in the case of the three RP conditions). Training times for 17 participants could not be calculated because of abnormal times spent on contents on their screen (i.e., several hours) and thus we decided to exclude them from these analyses.

There was a main effect of the learning condition on time spent on contents ($F(1,271) = 143.86$, $p < 0.001$, $\eta^2 = 0.35$). Indeed, the amount of time spent on the training phase was longer for the RP conditions ($M = 58.34$ min, $SD = 20.08$ min) than for the reading-reading conditions ($M = 30.74$ min, $SD = 18.26$ min). However, we found no main effect of the grain size ($F(2,271) = 1.26$, $p = 0.29$, $\eta^2 < 0.01$). Thus, time spent on contents were equivalent across the three grain sizes ($M_{Large} = 44.47$ min, $SD_{Large} = 25.62$ min; $M_{Medium} = 42.78$ min, $SD_{Medium} = 24.28$ min; and $M_{Small} = 43.38$ min, $SD_{Small} = 20.99$ min). Finally, we did not find a significant interaction between the learning condition and the grain size ($F(2,271) = 1.06$, $p = 0.35$, $\eta^2 < 0.01$; Figure 2, Table 2).

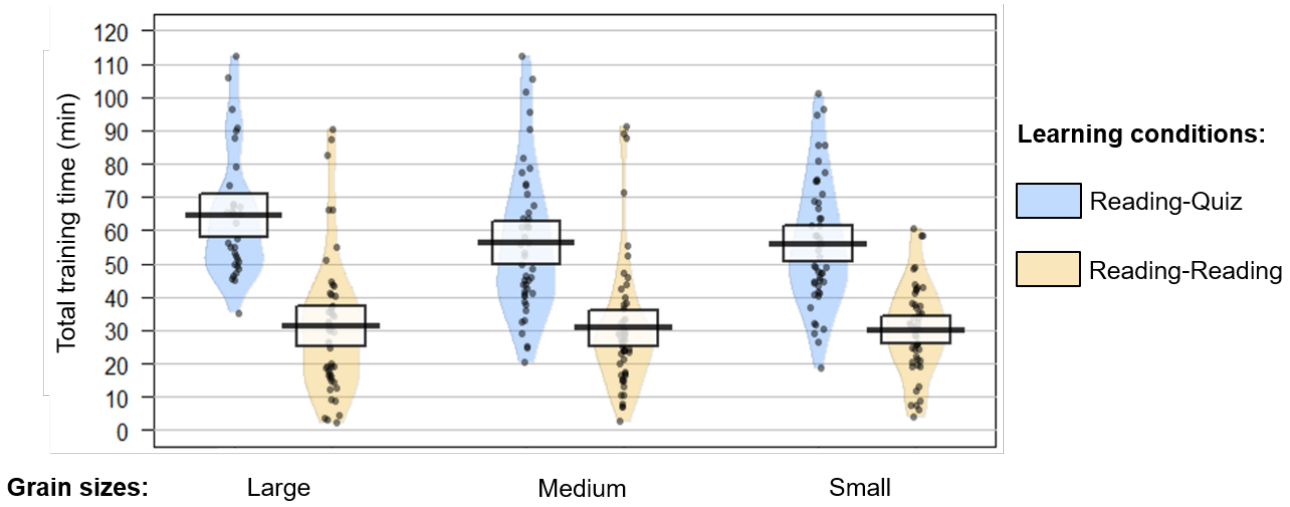


Figure 2. Cumulative time spent in the training phase (in minutes) for the two learning conditions as a function of the grain size. Each point is a participant. The thick horizontal line represents the median, color-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Intervals.

Table 2. Time spent in the training phase (in minutes) for the different groups of participants (learning conditions X grain sizes).

Learning groups						
R-R small grain	**R-R medium grain**	**R-R large grain**	**R-Q small grain**	**R-Q medium grain**	**R-Q large grain**	**Total**
30.17 min (13.60 min)	30.74 min (19.55 min)	31.28 min (20.63 min)	56.05 min (18.89 min)	56.40 min (21.77 min)	64.44 min (18.60 min)	43.50 min (23.54 min)

Note. Standard deviations are in parentheses.

Training performance in the retrieval practice conditions – One variable that was analyzed in previous studies was the score obtained in the training quizzes. We aimed at measuring if the training performance differed between the three grain sizes. Thus, we computed a one-way ANOVA with grain size as the independent variable and training score as the dependent variable. We did not find a main effect of the grain size during initial learning with quizzes, despite a subtle trend in favor of participants in the medium grain size condition ($F(1,137) = 2.43$, $p = 0.09$, $\eta^2 = 0.03$; Table 3, Figure 3).

Table 3. Training performance for each grain size in the retrieval practice conditions and planned t-test comparisons (effects sizes and p values).

Grain sizes			planned t-test comparisons		
Small	Medium	Large	Small vs Medium	Medium vs Large	Small vs Large
51.38 % (11.19 %)	55.10 % (10.08 %)	50.27 (12.20 %)	d = -0.37 [-0.65, -0.09] p = 0.08	d = 0.43 [-0.005, 0.87] p = 0.044	d = 0.10 [-0.18,0.38] p = 0.66

Note. Standard deviations are in parentheses. Effect sizes (Cohen's d) are given with 95% confidence intervals.

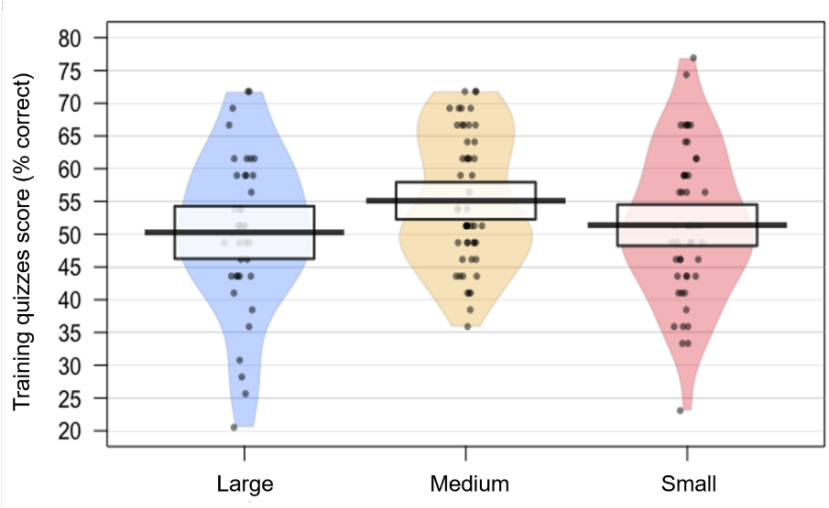


Figure 3. Initial performance in the training quizzes (percentage of correct answers) at each grain size in the RP conditions. Each point is a participant. The thick horizontal line represents the median, color-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Intervals.

Final tests' performance

Main analyses - Tables 4 summarizes post-hoc analyses. Table 5 summarizes the final recall performance at each final test, by learning condition and grain size, and for the types of final test questions (trained and untrained). Following our preregistered analysis plan[26], we ran a four way ANOVA (learning condition x grain size x retention interval x question type). First, we found a main effect of the learning condition ($F(1,1066) = 40.37$, $p < 0.001$, $\eta^2 = 0.031$); with the reading-quiz groups showing better final performance than the reading-reading groups (Table 4). The size of the overall RP effect across grain size was d = 0.37 [0.2, 0.49], $p < 0.001$.

[26] http://aspredicted.org/blind.php?x=zp2486

This did not interact with retention interval ($F(1,1066) = 1.69$, $p = 0.19$, $\eta^2 < 0.01$). Second, there was no main effect of the grain size ($F(2, 1066) = 0.45$, $p = 0.64$, $\eta^2 < 0.001$) and the grain size did not interact with the retention interval ($F(2,1066) = 0.15$, $p = 0.85$, $\eta^2 < 0.001$). Finally, there was a significant interaction between the learning condition and the grain size ($F(2,1066) = 7.21$, $p < 0.001$, $\eta^2 = 0.012$, Figure 4). This interaction did not differ according to the retention interval, as shown by the absence of a triple interaction ($F(2,1066) = 0.04$, $p = 0.96$, $\eta^2 < 0.001$). Post-hoc comparisons showed that at large and medium grain sizes the RP effect was stronger (d = 0.53 and d = 0.56 respectively) than at the small grain size (d = -0.063). In other words, there was a significant grain size effect restricted to the reading-reading groups ($F(2,558) = 6.28$, $p < 0.01$, $\eta^2 = 0.026$), with the smaller grain size leading to better overall final retention than medium one (d= -0.30) and the large grain one (d = -0.34). However, in the reading-quiz group, there was no grain size effect ($F(2,520) = 1.70$, $p = 0.18$, $\eta^2 =< 0.01$)).

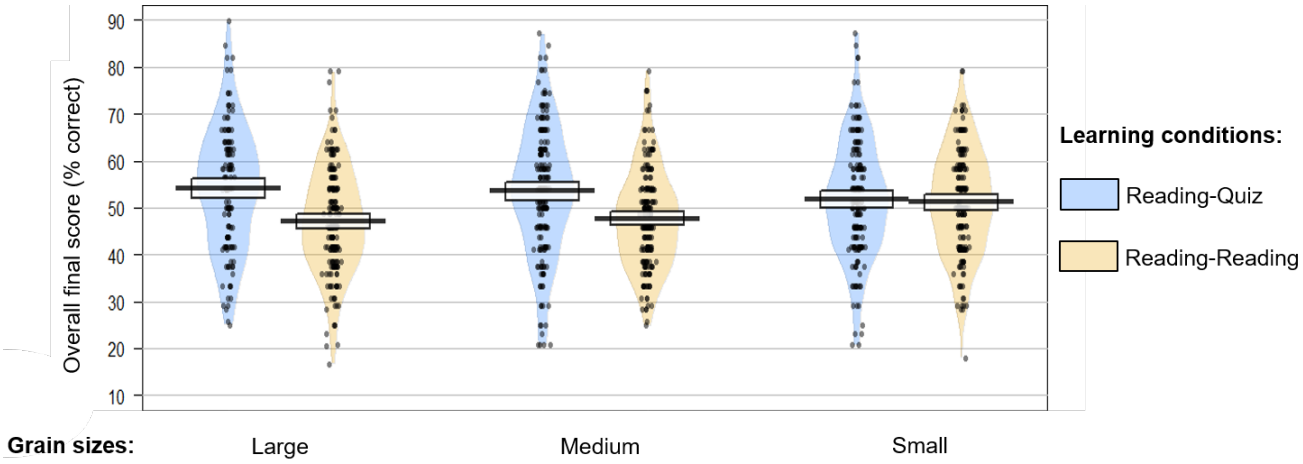


Figure 4. Overall final performance (percentage of correct answers) as a function of learning condition and grain size. Each point is a participant. The thick horizontal line represents the median, color-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Interval.

Table 4. Effect sizes and *p*-values for the planned t-test comparisons between the different conditions. The sign of each effect size is given by the comparison of the conditions listed in columns to those listed in rows.

	R-R small grain	R-R medium grain	R-R large grain	R-Q small grain	R-Q medium grain	R-Q large grain
R-R small grain		**d= -0.30 [-0.50, -0.10] *p* < 0.01**	**d= -0.34 [-0.53, -0.14] *p* < 0.01**	d = 0.063 [-0.13, 0.26] *p* = 0.55	NR	NR
R-R medium grain			d= -0.045 [-0.24, 0.15] *p* = 0.66	NR	**d = 0.56 [0.36, 0.76] *p* < 0.001**	NR
R-R large grain				NR	NR	**d = 0.53 [0.31, 0.74] *p* < 0.001**
R-Q small grain					d= 0.14 [-0.05, 0.34] *p* = 0.14	d= 0.20 [-0.01, 0.41] *p* = 0.087
R-Q medium grain						d= 0.036 [-0.17; 0.25] *p* = 0.75
R-Q large grain						

Note. Effect sizes (Cohen's d) are given with 95% confidence intervals in brackets. NR: Not relevant comparisons with *a priori* hypothesis.

Secondary analyses – The analyses indicated that there was a main effect of the question type on overall final score ($F(1, 1066) = 7.74$, $p < 0.01$, $\eta^2 = 0.02$). Questions encountered during the training phase were more successfully answered during the final test phase (overall performances: $M = 51.80$ %, $SD = 13.00$ %) than untrained ones ($M = 49.83$ %, $SD = 11.38$ %). Moreover, the scores obtained for the two types of questions were positively correlated and at both final tests ($r = 0.46$, $p < 0.001$ for final test 1; and $r = 0.49$, $p < 0.001$ for final test 2).

There was no main effect of the retention interval ($F(1,1066) = 0.08$, $p = 0.78$, $\eta^2 < 0.001$). The correlation between scores at final test 1 and scores at final test 2 was positive and strong ($r = 0.73$, $p < 0.001$). Moreover, we found a significant interaction between the learning condition and the question type on overall performance ($F(1,1066) = 56.82$, $p < 0.001$, $\eta^2 = 0.05$). Namely, final scores were higher for the trained questions in the reading-quiz conditions relative to reading-reading conditions (for this group all final test questions consisted of untrained

questions). This result partly explained the overall important RP effect that we found. But the participants had lower performances for the untrained than for the trained questions, and especially in the reading-quiz group (see Table 5). No other significant interaction was found.

Table 5. Final performance for each, learning condition and grain size, to the trained and untrained questions, and for each final test as well as overall performances.

		Learning groups						
		R-R small grain	**R-R medium grain**	**R-R large grain**	**R-Q small grain**	**R-Q medium grain**	**R-Q large grain**	**Total**
Final test 1	Trained	47.70 % (11.84 %)	45.36 % (9.70 %)	45.06 % (9.61 %)	54.30 % (13.85 %)	58.47 % (13.31 %)	57.56 % (13.75 %)	51.09 % (13.17 %)
	Untrained	52.86 % (9.26 %)	49.76 % (10.71 %)	48.53 % (11.79 %)	49.51 % (9.92 %)	49.84 % (11.02 %)	51.97 % (10.20 %)	50.30 % (10.57 %)
	All questions	49.67 % (9.24 %)	47.03 % (8.52 %)	46.38 % (9.15 %)	52.47 % (11.21 %)	55.18 % (11.25 %)	55.43 % (11.01 %)	50.79 % (10.62 %)
Final test 2	Trained	50.41 % (9.99 %)	46.26 % (9.58 %)	48.53 % (11.20 %)	56.30 % (11.92 %)	58.08 % (13.96 %)	56.80 % (14.99 %)	52.60 % (12.71 %)
	Untrained	53.45 % (12.67 %)	49.81 % (11.98 %)	47.32 % (12.09 %)	49.97 % (10.98 %)	48.37 % (13.19 %)	50.25 % (11.96 %)	49.39 % (12.25 %)
	All questions	51.57 % (9.99 %)	47.62 % (8.68 %)	48.07 % (10.31 %)	52.74 % (9.87 %)	54.38 % (12.55 %)	54.30 % (12.62 %)	51.36 % (10.92 %)
Both final tests	Trained	48.94 % (11.05%)	45.78 % (9.60%)	46.58 % (10.43 %)	55.22 % (12.97 %)	58.29 % (13.56 %)	57.20 % (14.24 %)	51.80 % (12.98 %)
	Untrained	53.14 % (10.90 %)	49.70 % (11.26 %)	48.00 % (11.88 %)	48.33 % (10.45 %)	49.14 % (12.06 %)	51.17 % (11.00 %)	49.83 % (11.38 %)
	All questions	51.04 % (11.14 %)	47.78 % (10.63 %)	47.29 % (11.17 %)	51.78 % (12.24 %)	53.71 % (13.60 %)	54.19 % (13.04 %)	50.82 (12.24 %)
Learning Conditions	Trained		47.05 % (10.40 %)			56.89 % (13.55 %)		
	Untrained		50.24 % (11.52 %)			49.40 % (11.23 %)		
	All questions		48.65 % (11.08 %)			53.14 % (12.98 %)		

Note. Standard deviations are in parentheses. R-R: Reading-Reading, R-Q: Reading-Quiz. Descriptive data for the Learning conditions lines are those for the three R-Q conditions versus the three R-R conditions.

Finally, the linear regression restricted to the untrained questions indicated no significant effect of the learning condition ($F(1,533) = 0.75$, $p = 0.39$, $\eta^2 < 0.001$), neither of the grain size ($F(2,533) = 0.80$, $p = 0.45$, $\eta^2 < 0.001$) on the score obtained on this question type. In addition, the scores was equivalent to final test 1 and final tes2 (no effect of the retention interval, $F(1,533) = 1.06$, $p = 0.30$, $\eta^2 < 0.001$). Interestingly, there was a significant interaction between the learning condition and the grain size ($F(2,533) = 5.41$, $p < 0.01$, $\eta^2 = 0.02$).

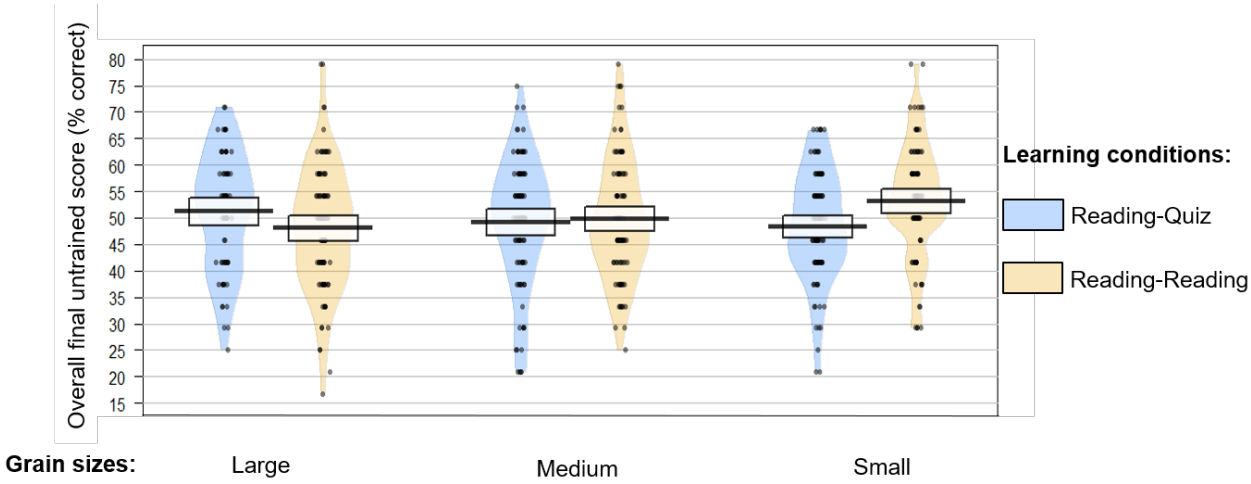


Figure 5. Final performances to the untrained questions (percentage of correct answers) as a function of the learning condition and the grain size. Each point is a participant. The thick horizontal line represents the median, color-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Interval.

Covariates and exploratory analyses

Initial training covariates - Although there was no difference between conditions in the duration of retention interval between initial learning and final test 1 (R1), and initial learning and final test 2 (R2); we added retention intervals as a covariate in the model. None of the two intervals had an influence on the final performance ($F(1,978) = 0.43$, $p = 0.51$, $\eta^2 < .001$ for R1, and $F(1,978) = 0.81$, $p = 0.37$, $\eta^2 < 0.001$ for R2), neither on the main effects and interactions. Given that we found differences between the groups of learning condition, it was relevant to use this variable as a covariate. First, the simple correlation between the total training times and the overall final test score was positive and significant ($r = 0.27$, $p < 0.001$). When included in the four way ANOVA as a covariate, the total training time had an effect on final test scores ($F(1, 1001) = 47.98$, $p < 0.001$, $\eta^2 = 0.04$). However, this did not alter any of the main results. Namely, the main effect of the learning condition was ($F(1, 1001) = 37.36$, $p < 0.001$, $\eta^2 < 0.001$). There was no main effect of the grain size ($F(2, 1001) = 0.82$, $p = 0.44$, $\eta^2 < 0.01$), and

the interaction the learning condition and the grain size stayed significant ($F(2, 1001) = 6.38$, $p < 0.01$, $\eta^2 = 0.011$).

Other covariates – Age and years of education were included together as covariates in the four-way ANOVA. While education level had an influence on overall final performances ($F(1, 1064) = 44.07$, $p < 0.001$, $\eta^2 = 0.033$), age did not ($F(1, 1052) = 2.69$, $p = 0.11$, $\eta^2 < 0.001$). However, this did not change any of the main results. Namely, the main effect of the learning condition was ($F(1, 1064) = 43.33$, $p < .001$, $\eta^2 < .001$). There was no main effect of the grain size ($F(2, 1064) = 0.55$, $p = .58$, $\eta^2 < .01$), and the interaction the learning condition and the grain size stayed significant ($F(2, 1064) = 6.87$, $p < .01$, $\eta^2 = .011$).

Taken together, these covariates seemed to have a relative importance to explain a proportion of the total variation in our results. Table 6 summarizes the adjusted R^2 of the model without and with the socio-demographic and training times covariates. Adding the covariates in the linear regression model increases the value of the adjusted R^2, thus increases the proportion of the total variation explained by the model. This value did not change when adding more covariates (i.e., retention interval at final test 1 and retention interval at final test 2).

Table 6. Adjusted R^2 we obtained

Linear regression models	Adjusted R^2 of the model
Without covariates Predictors : Learning condition X Grain size X Question type X Retention	0.11
+ Education level + Age	0.15
+ Education level + Age Total + Total training time	0.19

4.6 Discussion

The results from this experiment provide insights into the effective placement of RP relative to content study with lengthy and complex material. This study also shed light on new results on the grain size of interpolated RP to promote long-term retention. These results were obtained for the learning of prose passages of scientific content, in a digital learning environment, and with retention intervals of 7 days and 27 days. First, we found a significant benefit of learning with RP in comparison to extra reading on overall performance (i.e. RP effect). Across grain sizes and retention intervals, the reading-quiz groups had better final retention relative to the reading-reading groups. Second, we did not find an overall grain size effect across learning conditions and retention intervals. Overall, the three grain size conditions led to equivalent final

retention. Finally, our results extend beyond those of the studies that explored the question of the optimal placement of RP because we found a significant interaction between the learning condition and the grain size. Namely, the effect of the grain size variable was not the same depending on the learning condition. In the reading-quiz conditions, readings and quizzes alternated. Thus, grain size referred to the size of the chunks to be read, and then to be tested on. In the reading-reading conditions, each content had to be read twice. Grain size referred to the size of the chunks to be read, before having to read them again. When learning with RP, whatever the grain size final retention is the same. However, learning with small grain size of interpolated extra reading promoted better final retention than learning with medium and large grain sizes of extra reading. This significant interaction persisted regardless of retention interval and question type. Therefore we replicated previous findings on the equivalent benefit of interpolated RP vs postponed RP as demonstrated in previous studies (Duchastel & Nungester, 1984; Uner & Roediger, 2017; Weinstein et al., 2016; Wissman & Rawson, 2015). However, in the reading-reading conditions, the large and medium grain sizes led to equivalent and lower final performances than the small grain size. Consequently, the amplitude of the RP effect was modulated by the grain size. The RP effect was larger at large ($d = 0.53$) and medium ($d = 0.56$) than at the small grain sizes ($d = -0.063$). In the extra reading condition, retention was better at small than at medium ($d = 0.30$) and large grain sizes ($d = 0.34$), thus cancelling the retrieval practice effect specifically at small grain size.

At large grain size in the extra reading condition, participants had to read the entire material once, then had to read it entirely again or took a quiz on the entire content while at small grain size, participants read much shorter contents and read them a second time immediately after the first time. In the context of educational settings, it applied that students should read each page twice before moving on to the next page when reading a textbook, and so on for the entire chapter. This is certainly not a conventional restudy technique, compared to the standard practice of reading the entire chapter twice (comparable to our large grain size condition). Yet, the present results suggest that the extra reading with small grain size produced the best retention. One hypothesis for the finding is that the first exposure consists in discovering the topic and new key concepts while the second and immediate re-exposure really enables the leaners to understand and consolidate the information in long-term memory before moving the next chapter. This strategy of learning might enable the consolidation by the immediate repetition of the reading while gradually increasing the number of new information, and therefore might indirectly, preventing from retroactive interference. In the large and medium grain reading-reading conditions, learners might have felt more fatigue associated with less

attentional focus because there was too many new information to encode. In their study Szpunar, McDermott, & Roediger (2008) used a restudy condition more or less equivalent to our large grain size condition. Participants had to restudy word lists at the end of the whole first study session (i.e., postponed restudying). They found that restudying all the lists at the end of the first exposure led to the same final performances as learning with one single study session; as if additional restudying gave no benefit at all. In the context of multimedia learning, Mayer & Pilegard (2014) stated that splitting the learning material enables people to learn more efficiently and attentively while avoiding important overload. The authors summarized ten comparisons between groups that received a multimedia presentation broken into small segments with the pace controlled by the learner and groups that were presented the same multimedia content but as a unique continuous presentation Overall, there was a positive effect size for segmented learning based on transfer test score ($d = 0.79$).

Within the RP conditions, the large grain size produced the best retention relative to the small grain size, albeit not significantly ($d = 0.2$, $p = 0.087$). Thus, RP may be more beneficial as applied to the equivalent of an entire textbook chapter, than to single pages or small sections.

One interpretation of this trend is that, the larger the grain size, the more distant the quiz from the exposure to the contents. Larger grain sizes may be more beneficial because they impose greater spacing between exposure to key knowledge and the associated RP episode, and thus induce more effort during the retrieval process. A second interpretation is that the small grain size might hinder the connections between the main ideas of the different chapters of the course and thus hinder the construction of a coherent representation of the key knowledge, as suggested by Wissman & Rawson (2015).

Finally, results on the training performance in the RP conditions did not replicate previous results from the related literature that showed an obvious and significant advantage of the interpolated RP during the training session. Interestingly, the three RP groups improved their performance on the trained MCQ between the end of the training session and the final tests session. Descriptive results showed that the best improvement was for the largest grain size of RP, equivalent to a postponed RP placement, thus giving some argument in favour of this placement. Nevertheless, the fact that we found no difference at all between the large and the medium grain sizes suggested that increasing twice the amount of information to read and the number of MCQ in the quizzes does not impact the learning process. What is important is the act of being active during the step of storing the knowledge, whatever the placement of RP relative to reading the content.

As one would expect, there was a significant interaction between the learning condition and the question type: in the reading-quiz learning conditions, final test scores were higher for trained questions than for untrained questions (d = 0.60). In the extra reading in which there was no distinction between trained and untrained questions, final tests score on the trained questions were surprisingly lower than on untrained questions (d = -0.29). Therefore, any difference in performance must reflect intrinsic difficulty differences between the two sets of questions: it appears that the questions selected for the training quizzes were on average more difficult than those selected for the final test only. This implies that the advantage of trained over untrained questions in the RP conditions is probably under-estimated. Moreover, we did not find a difference between RP and extra reading groups on the untrained questions (d = 0.07, p = .39). Thus, we did not replicate some earlier results showing that the testing effect may also lead to better retention of previously untested information and to greater knowledge transfer than restudying (Butler, 2010; Chan, McDermott, & Roediger, 2006; Karpicke & Blunt, 2011). However, a recent meta-analysis on the transfer benefits of the RP effect found mixed results going both ways, and depending on the untrained question type (Pan & Rickard, 2018). The authors made a distinction between untested application and inference questions and untested information seen during the first exposure to material. Interestingly, they found a transfer effect of RP on application and inference questions, but not on untested information seen during initial study. In the present study, the untrained MCQ were very similar to the trained MCQ, including questions related to untrained information seen during the learning phase, but they may not have involved many inferences.

Interestingly, analyses restricted to the untrained questions scores at both final tests showed that the grain size effect was emphasised in the extra reading conditions compared to the grain size effect measured on overall performances. Thus, it seems that the benefit on these specific questions was stronger when reading twice with small grain size relative to RP. Some studies found that interpolated RP could impair subsequent learning. Thus, the act of retrieving some information can undermine the ability to learn new information immediately thereafter and frequent switching between RP of previously studied material and encoding of new related information might reverse the RP effect into a negative effect on the recall of new items (Davis, Chan, & Wilford, 2017; Finn & Roediger, 2013). Finn and Roediger (2013) argued that intermixing RP and new learning might encourage learners to prioritize reviewing the initial learning items over the new items. Thus, frequent switching between RP and reading new related content can impair the retention of this subsequent information (Pashler et al., 2000).

Our covariates analyses showed that the time spent on the training phase had a non-negligible effect final tests scores. Uner and Roediger (2017) also found that time on learning and final performances were strongly correlated (r = 0.56). Because of the self-paced nature of the experiment that better simulates how students learn in real world context, Training time was only partly controlled. Minimum reading times were imposed but it was not possible to equalize study time between reading and quizzes. As a result, total training time differed considerably between reading-reading and reading-quiz conditions, and this had an impact on final tests scores. Nevertheless, all the main results reported remained the same once training time was adjusted. In addition, this type of exploratory analyses also emphasizes that it is appropriate to take into account for new predictors *a posteriori* to explain some proportion of the huge variance in final scores.

The present results may be subject to several limitations that are very similar to the ones we exposed in one of our previous study (Latimier et al., 2019). Namely, our conclusions are limited to a certain type of material and mode of presentation as well as to certain type of participants (i.e., workers doing a paid task, rather than students). We tried to design a controlled laboratory experiment that mimicked reviewing situation close to real word learning contexts. Our large grain size condition was the equivalent of a textbook chapter, whereas our small grain size was one sixth a large grain, thus the equivalent of a textbook chapter section. But it would be important to address the same question using other learning contents and populations, especially school children in a more ecological context. Finally, we were unable to exclude or control potential extra study time between the training phase and the final test. However, participants were paid for participation regardless of performance, thus they had no incentive to spend more time studying than the minimum imposed. Furthermore, there is no reason to think that participants in different groups might have invested differently in extra study. Therefore, there is no reason to think that this may have changed the pattern of results.

In terms of practical Implications, we can provide some modest recommendations. The answer to the question of the optimal moment to interpolate learning questions while delivering new study contents on a specific topic is that the grain size does not really matter for long-term retention. However, we hypothesize that bigger grain sizes should be preferred because they impose greater spacing between exposure and first retrieval. By contrast, the grain size of the learning contents is especially of interest when learning with repeated reading, for which smaller grain size should be preferred. Nevertheless, much more research is needed to systematically investigate the effects of various grain sizes and learning contents, to ultimately be able to recommend an optimal grain size for a given type of content.

4.7 References

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. https://doi.org/10.3102/0034654316689306

Butler, A. C. (2010). Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

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

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671–682. https://doi.org/10.3758/s13421-012-0291-4

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology. Applied*, *24*(1), 34–42. https://doi.org/10.1037/xap0000145

Carpenter, S. K., & Toftness, A. R. (2017). The Effect of Prequestions on Learning from Video Presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

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

Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The Dark Side of Interpolated Testing: Frequent Switching Between Retrieval and Encoding Impairs New Learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 434–441. https://doi.org/10.1016/j.jarmac.2017.07.002

Duchastel, P. C., & Nungester, R. J. (1984). Adjunct Question Effects with Review. *Contemporary Educational Psychology*, *9*(2), 97–103.

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

Finn, B., & Roediger, H. L. (2013). Interfering Effects of Retrieval in Learning New Information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1665–1681. https://doi.org/10.1037/a0032377

Healy, A. F., Jones, M., Lalchandani, L. A., & Tack, L. A. (2017). Timing of Quizzes During Learning: Effects on Motivation and Retention. *Journal of Experimental Psychology. Applied*. https://doi.org/10.1037/xap0000123

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305–318. https://doi.org/10.1037/xap0000087

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Latimier, A., Riegert, A., Peyre, H., Ly, S. T., Casati, R., & Ramus, F. (2019). Does pre-testing promote better retention than post-testing? *Npj Science of Learning*, (15). https://doi.org/10.7910/DVN/XPYPMF

Mayer, R. E., & Pilegard, C. (2014). Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In *Cambridge Handbooks in Psychology*. *The Cambridge handbook of multimedia learning, 2nd ed* (pp. 316–344). https://doi.org/10.1017/CBO9781139547369.016

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., Jolicœur, P., Dell'Acqua, R., Crebolder, J., Goschke, T., De Jong, R., … Hazeltine, E. (2000). Task switching and multitask performance. In *Control of cognitive processes: Attention and performance XVIII* (pp. 275–423). Cambridge, MA, US: The MIT Press.

Reynolds, R. E., Standiford, S. N., & Anderson, R. C. (1979). Distribution of reading time when questions are asked about a restricted category of text information. *Journal of Educational Psychology*, *71*(2), 183–190. https://doi.org/10.1037/0022-0663.71.2.183

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

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

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

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during Study Insulates against the Buildup of Proactive Interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392–1399.

Uner, O., & Roediger, H. L. (2017). The Effect of Question Placement on Learning from Textbook Chapters. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2017.09.002

Watts, G. H., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology*, *62*(5), 387–394. https://doi.org/10.1037/h0031633

Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology. Applied*, *22*(1), 72–84. https://doi.org/10.1037/xap0000071

Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(2), 439–455. https://doi.org/10.1037/xlm0000047

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140–1147. https://doi.org/10.3758/s13423-011-0140-7

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *Npj Science of Learning*, *3*(1), 8. https://doi.org/10.1038/s41539-018-0024-y

5 Chapitre 5 : “Do spacing and retrieval practice effects interact?”

This chapter constitutes the following manuscript: Latimier, A., Riegert, A., Ly, S.T., Ramus, F. Do spacing and retrieval practice effects interact? (in prep).

The preregistration for this experiment can be accessed at:

http://aspredicted.org/blind.php?x=e29eu7

Learning contents and examples of training questions are provided in the Appendices section (see Annexe 3).

5.1 Abstract

As opposed to commonly used learning strategies such as reading and cramming (massed learning), retrieval practice and spaced learning have been shown to contribute to better long-term retention among a great variety of populations and content. Thus, spaced retrieval practice seems an optimal learning strategy. However, it is not known whether the effects of spaced learning and retrieval practice are simply additive, or whether their interaction produces even greater retention. This study investigated the benefits of retrieval practice, spaced learning, and the combination of the two learning strategies on long- term retention relative to massed reading. Furthermore, we measured judgments of learning at different steps of the experiment in order to assess to what extent either learning practice and their combination produces illusions of competence. We ran the experiment in an online environment with educationally relevant contents. First, we replicated the retrieval practice effect but we did not find a significant spacing effect. Second, we found no interaction between retrieval practice and spacing. Finally, participants in the retrieval practice groups tended to be less overconfident than the participants in the reading groups, and they felt more able to recall the new knowledge at the exams. Overall, our results suggested that being tested during learning had a stronger benefit than spacing reviewing sessions, at least at 1- and 7-day intervals. We found no evidence that combining both strategies was more advantageous than retrieval practice alone. This study paves the way for further research to better understand how combining different efficient learning strategies may potentiate their respective benefits on memory.

5.2 Introduction

Scope of the testing and spacing effects

Research on the efficacy of learning practices typically consists in studying the potential benefit of a practice during a training phase over another one, on final performance in a test phase. During the training phase, participants are exposed to learning contents (i.e., reading or listening), and have to review the information under various modalities and time scales. Tests for memorization of the initial learning contents occur after a certain retention interval (i.e., immediate or delayed). At least two major results, named "desirable difficulties" (Bjork & Bjork, 2011, Weinstein et al. 2018) have been described in the literature.

- Testing (or retrieval practice) effect

First, the effect of testing one's knowledge with retrieval practice (e.g., free or cued recall, multiple choice questions, or application exercises) during the study phase leads to better retention than just re-reading or re-exposure to the material (see two meta analyses; Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014). To briefly sum up, the testing effect is strong and positive with a mean effect size between g = 0.50 [0.42, 0.58] and g = 0.61 [0.58, 0.65]. Moreover, one retrieval event is enough to elicit better retention than no testing at all, and the retention interval before a final test should be long enough to promote long term benefits without being too distant from the end of the learning phase (*i.e.,* 1-30 days). Rowland (2014) also reported that retrieval practice was stronger when the learning contents were more difficult, when the type of retrieval practice was more effortful, and when feedback was given during practice. This learning strategy also provides benefits that are transferred to untested information during the final test phase depending on the format of the question (see Pan & Rickard, 2018 for a recent meta-analysis on the transfer effect of retrieval practice). Finally, this "testing effect" is well-established for a great diversity of learning material (Karpicke & Aue, 2015) and in laboratory or classroom settings (Karpicke & Grimaldi, 2012; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011).

- Spacing effect

Inserting time intervals between study episodes promotes better retention than massed practice (i.e., one session without inter-study intervals). Spaced (or distributed) learning is thought to optimally counteract forgetting and improve the consolidation of newly learned information (Cepeda et al., 2009; McDaniel, Fadler, & Pashler, 2013). In contrast, it seems that reading a

text once again right after the first reading even provides negligible gains in retention relative to reading the text only once (Callender & McDaniel, 2009).

Similar to the testing effect, the spacing effect has been shown in different learning contexts and populations (Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Shana K. Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Gluckman, Vlach, & Sandhofer, 2014; Larsen, 2018; Vlach & Sandhofer, 2012). Spacing effects have been shown from very short inter-study intervals a few seconds or several minutes, e.g., Glover & Corkill, 1987; Whitten & Bjork, 1977), to much larger intervals such as days or weeks (e.g., Dobson, Perez, & Linderholm, 2016; Hopkins, Lyle, Hieb, & Ralston, 2016).

The mean effect size of the spacing effect is less clear than the testing effect but the meta-analyses from Hattie (2008) yielded an estimation of d= 0.71. Since then, a large literature on spaced practice of verbal/educational content has emerged and these estimations encompassed a wide array of learning tasks, with a predominance of motor tasks, and relatively few verbal learning tasks. Cepeda et al. (2006) did not provide an effect size but their meta-analysis suggested that the optimal spacing effect is obtained when the inter-study interval increases together with the retention interval. Several studies have suggested 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; Gerbier, Toppino, & Koenig, 2015; Godbole, Delaney, & Verkoeijen, 2014; Kapler, Weston, & Wiseheart, 2015). However, the amplitude of spaced learning may depend on the retention interval (Greving & Richter, 2019; Rawson & Kintsch, 2005). Indeed, these findings suggested that spaced learning might produce better retention on delayed assessments, while massed learning might promote better retention on immediate assessments. Very–recently, Greving and Richter (2019) replicated this interaction between the schedule of repeated readings and the retention interval before the final test.

Combination of testing and spacing: spaced retrieval practice

Even though both learning strategies have been primarily investigated separately from each other (Miyatsu, Nguyen, & McDaniel, 2018), a number of studies have investigated them in combination. Whitten and Bjork (1977) were among the first to study the combination of retrieval practice and spacing and found a "spacing-of-tests effect" in both paired-associate learning and free-recall tasks. At a final test, they found a testing effect and a spacing effect and found a benefit of spaced (41% correct) over massed retrieval practice (34% correct). Since this pioneer experiment, many studies have replicated the benefits of spaced retrieval practice

compared with massed practice: introducing spacing between repeated retrieval events enhanced significantly retention in laboratory as a well as in educational settings with long retention intervals (e.g., Lyle, Bego, Hopkins, Hieb, & Ralston, 2019; Maddox & Balota, 2015; Shaughnessy & Zechmeister, 1992).

Overall results suggest that spaced retrieval practice is a robust and significant learning strategy that should be easily implemented in curricular activities across various settings, materials, tasks, and retention intervals (Karpicke, Blunt, & Smith, 2016; Larsen, 2018). Finally, a recent meta-analysis conducted by our team provides reveals a strong and significant benefit of spaced over massed testing g = 0.74 (95% CI = [0.55, 0.92]; Latimier, Peyre, & Ramus, unpublished). However, the available studies did not allow us to assess whether spacing and retrieval practice had additive or interactive effects. Thus it seems important to more directly investigate this question: this is the primary goal of the present study.

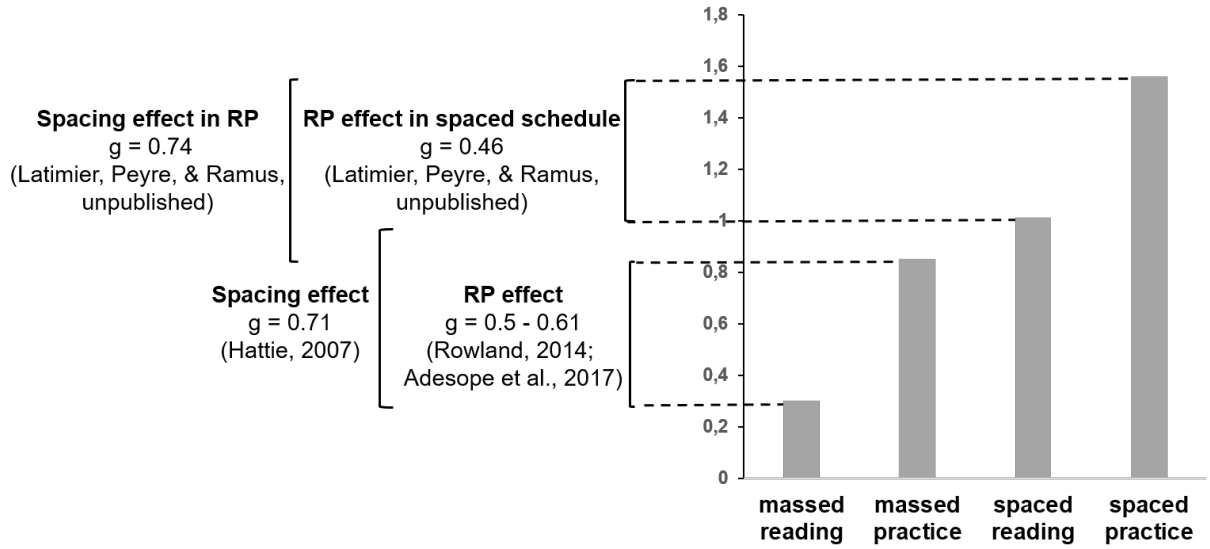


Figure 1. Synthesis of previous meta-analyses and the ones conducted by our team. This represents the comparison between four learning strategies which are investigated in research on the retrieval and spacing effects (from Latimier, Peyre, & Ramus, unpublished). The Y axis is expressed in terms of mean effect size (Hedges's g), with an arbitrary value for the massed reading strategy (g = 0.3). "RP" is for retrieval practice. The effect size of the benefit of spaced RP over massed reading is unknown, as well as the difference between massed practice and spaced reading.

Different schedules can be used to space the retrieval practice events. Several laboratory experiments found a superiority of the expanding schedule (William L. Cull, Shaughnessy, &

Zechmeister, 1996; Kang, Lindsey, Mozer, & Pashler, 2014; Maddox, Balota, Coane, & Duchek, 2011; Storm, Bjork, & Storm, 2010), but other studies found the opposite result (W. L. Cull, 2000; Karpicke & Roediger, 2007; Logan & Balota, 2008, Toppino, Phelan, & Gerbier, 2018 in a high level initial training condition). In addition, a non-negligible part of the literature reports no difference between the two schedules (Carpenter & DeLosh, 2005; W. L. Cull, 2000; William L. Cull et al., 1996; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2010; Logan & Balota, 2008; Pyc & Rawson, 2007; Storm et al., 2010; Terenyi, Anksorus, & Persky, 2018, Petersen-Brown, 2019). We compiled results from these different studies in our recent meta-analysis and our results did not support the widely held idea that retrieval intervals should be progressively increased to enhance long term retention; g = 0.032 (95% CI = [-0.10, 0.17], Latimier, Peyre, & Ramus, unpublished).

Judgments of learning and effective learning strategies

Judgments of learning (JOLs) are usually used to evaluate metamemory (i.e., self-awareness of memory) of learning contents (Rhodes, 2016). It consists in asking learners to predict how well they would remember the content on a retention test that is more or less delayed. Studies who investigated the benefits of testing and spacing on Judgments of learning accuracy found that learners overestimated their ability to recall correctly when items were i) restudied with reading and ii) learned in a massed schedule (Kornell & Son, 2009; Rhodes & Castel, 2008; Son & Simon, 2012). Such overconfidence was in turn associated with lower learning success. The fact that less efficient learning strategies (i.e., repeated reading, cramming, blocked practice) create a feeling of knowing well is a phenomenon also called « illusion of mastery » (Bjork, Dunlosky, & Kornell, 2013). By contrast, learning strategies that promote long-term retention often yield metacognitive judgments that inappropriately reflect the difficulty of immediate acquisition, rather than the robustness of long-term benefits. Indeed, in experiments investigating the effectiveness with which learners monitor the effects of testing, participants judged restudied items as more likely to be remembered, but they actually remembered previously tested items better (e.g., Roediger & Karpicke, 2006). This inaccurate appreciation of the benefits of retrieval practice seems particularly true when learning generates errors (Huelser & Metcalfe, 2012). Learners fail to realize that generating errors greatly facilitates recall under this condition, even after having just experienced enhanced test performance. It is quite plausible that learners rely on these types of global retrospective judgments when deciding what learning strategy to use, thus explaining the widespread use of inefficient strategies (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). A recent study showed that learners

can appreciate that retrieval practice is more effective than restudying when retrieval practice is successful during the learning phase (Toppino, LaVan, & Iaconelli, 2018). Similar failures to appreciate the benefits of spacing repetitions or interleaved practice have also been reported. (Baddeley & Longman, 1978; Zechmeister & Shaughnessy, 1980).

Recently, Fernandez and Jamet (2017) investigated the effect of testing on monitoring accuracy: participants in the retrieval practice group were significantly less overconfident in their ability to recall recently learned information than the reading only group. Greving and Richter (2019) reproduced the same result in the context of learning with spaced reading versus massed reading. Thus, using desirable difficulties might promote better self-regulated learning than less efficient learning strategies under certain circumstances (McCabe, 2011). To our knowledge, the effects of spaced retrieval practice on metacognitive accuracy with JOLs have not been investigated before: does spaced retrieval practice induce more perceivable difficulty than other learning strategies that do not combine spacing and testing?

Our study

The present study aimed at answering the following questions: i) What is the relative effect size of retrieval practice and spacing effects? ii) Are retrieval practice and spacing effects additive or do they interact? Thus, we used a crossed design with retrieval practice (reading vs testing) and spacing (massed vs spaced) as factors in order to provide a more direct comparison of the two practices and provide more evidence on the potential interest of combining both practices. Overall, this study aimed at providing the missing and unclear effect size values on Figure 1. We also assessed the impact of the different learning strategies on monitoring accuracy through Judgment Of Learning questions. In other words, we asked participants to make predictions on their performance on the upcoming final tests, after study, but prior to taking these tests.

5.3 Methods

Participants

We calculated that at least 64 participants per group were necessary in order to detect a testing effect of size d = 0.50 (Rowland, 2014) in a between-subjects design with a statistical power of 0.8 (alpha = 0.05, bilateral t-test); and at least 33 participants per group were necessary in order to detect a spacing effect of size d = 0.71 (Hattie, 2007). As we tested an interaction between testing and spacing whose possible effect size is unknown thus we aimed at including 100 participants per group. We therefore recruited a total of 395 adult participants at the first phase

of our experiment via the American online work platform Amazon Mechanical Turk (https://www.mturk.com/). Inclusion criteria were that participants are native English speakers and without any self-reported neurological or psychiatric disorders. Furthermore, participation of students and professionals in medicine or biology was discouraged in the call for volunteers and in the information letter. All participants provided written informed consent on the online platform. They were paid a total of 20US$ for their participation, in 4 instalments: after the first exposure phase, at the end of the training phase, at the end of both Exams.The study was approved by the local research ethics committee (Comité d'Ethique de la Recherche of Paris Descartes University) and all the participants were paid for their participation.

Material

The experiment was entirely run online on the digital learning platform Didask (https://www.didask.com/). In this platform, each course consists of a set of modules organized in a logical order. A module is an elementary learning unit, including both learning material (text, videos, pictures…) and a corresponding training quiz with at least 5 questions (multiple choice, pairwise matching, ordering, and sorting into categories…). We used a modified version of Didask designed for experimental testing, allowing us to specify several learning conditions and to better control the learning environment (e.g., the order in which the content and the quiz were presented).

Study material consisted of a course on DNA from the French curriculum for 12th grade science tracks (age range: 18 years old). Specific contents were borrowed from material provided by CNED (Centre National d'Enseignement à Distance). Texts were adapted and illustrations were added so as to create seven short chapters (their lengths varied between 100 and 172 words) forming a logical progression like in a textbook. For each of the seven short texts, five or six multiple choice questions (MCQ) were created for the training phase by selecting five or six main facts from the corresponding text. The MCQ were either directly related to the written information presented in the chapters, or generalized questions that were more inference-based questions that were more difficult and could cover several main concepts of the readings. Thus, all the facts questioned in the training quizzes were covered in the corresponding text. However, not all the facts covered in the chapters were used to create item quizzes, about half of them remained unquestioned.

An elaborative feedback was created to be given immediately after each question of the quizzes. This feedback indicated the correctness of the answer and provided explanations. These

explanations did not exceed three or four sentences and were directly extracted from the corresponding chapter.

Thus, a total of 40 MCQ were created for the reviewing content phase. Between 2 and 4 different choices were offered for each MCQ. Then we divided this set of MCQ into two lists of 20 (i.e., List 1 and List 2) constituting a training and a test set. Half of the participants were randomly assigned to and trained with list 1 during the initial learning phase, and was tested with both lists during the final exam (trained list 1 and untrained list 2), while the other half were trained with list 2.

Finally, five "catch" questions were interspersed throughout the set of training MCQ and the final tests to ensure participants were paying proper attention and were not answering randomly (e.g., "What is the colour of Henri IVth's white horse?"). They were also used as exclusion criteria (data from participants who correctly answer fewer than 4 of the 5 "catch" questions during the training phase were excluded).

For the learning conditions with quizzes, the training quizzes therefore included a total of 25 questions: 20 MCQ and 5 catch questions. For all the learning conditions, the final test quiz therefore included a total of 45 questions: the 40 MCQ from the two lists and the 5 catch questions. Thus, the final test quiz contained the same number of trained and untrained questions, and each chapter of the course was equivalently represented.

Design and Procedure

We used a mixed factorial design that included: 2 between-subject Learning Practices (testing versus reading) and 2 between-subject Learning Schedules (massed versus spaced) for the learning phase, thus leading to four learning conditions and groups of participants (massed reading, massed testing, spaced reading, spaced testing). Moreover, the design included 2 within-subject Retention Intervals for the final test (at Day 5 and at Day 12). For the two groups with retrieval practice, there was a further "Trained MCQ List" (List 1 versus List 2) between-subject factor corresponding to the training list.

After giving their consent and filling in a socio-demographic questionnaire, participants were allocated alternatingly to one of the 4 groups of learning conditions (2 Learning Practices X 2 Learning Schedules) based on order of registration, then were provided with an individual link to the testing platform on Didask. All reading and testing took place online, from participants' personal computers (uncontrolled settings such as their homes or offices).

The experiment consisted in three phases (see Figure 2). There was a first exposure to the learning contents, then a review phase, and finally a final assessment phase at two different intervals. We wanted to recreate a learning context as naturalistic as possible, that mimicked the situation in which a student attends a new lecture and has to review the key concepts to prepare a final exam.

The recruitment phase was open to new participants for 20 hours, until we obtained a sufficient number of participants and decided to close the task on Amazon MTurk.

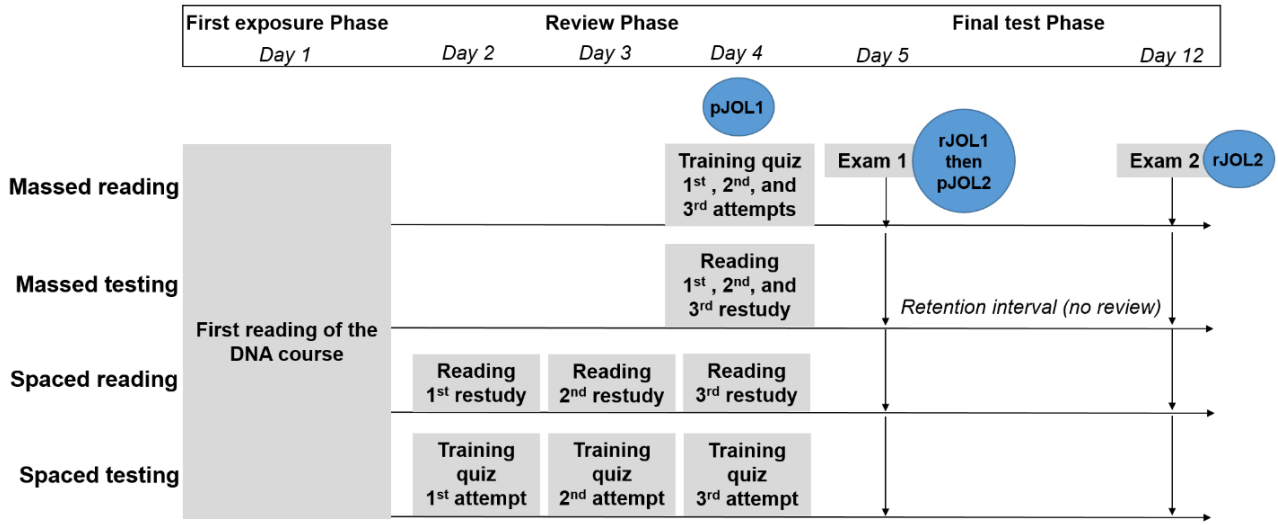


Figure 2. Schema of the experimental procedure used in the four learning conditions. pJOL: prospective Judgment Of Learning, rJOL : retrospective JOL, 1 and 2 for exam 1 and exam 2.

First exposure phase – Immediately after recruitment, each participant went on with the first exposure phase, reading the seven chapters of the course on DNA. The course took the form of a pdf (one page for each chapter) and was displayed on screen for at least 6 min in total and no upper time limit. This phase was identical for all participants. After the last recruited participant completed this first exposure phase, access to the course was closed.

Review phase – At the end of the first phase, participants were informed of instructions for the next phases depending on their group. Participants in the spaced schedule groups had to review the contents on Day 2, 3 and 4, while the massed schedule groups had to do the review phase in one shot on Day 4. The massed reviewing phase consisted in either doing three times the same training quiz (testing groups) or reading three times the same learning content (the seven short chapters on DNA presented in the first exposure phase).

Once a reading or an attempt to do the quiz was completed, it became unavailable until the next day. Thus, the two learning schedules emulated and a spaced reviewing schedule (once per day until the exam) vs. a crammed schedule (same amount of reviewing the day before the exam). No time limit was imposed to restudy the course or to do the training quiz. After answering each question of the quiz, feedback immediately appeared and was displayed for at least 7 secs before participants could go on to the next question. Reviewing times were recorded in order to be taken into account for data analysis. At the end of the first exposure and reviewing phases, participants were instructed not to study anymore on DNA. This was of course impossible to control, but there was no incentive for further study, given that payment was a flat rate for participation, regardless of performance in the final test.

Final test phase - Participants who completed the reviewing phase correctly until the end were asked to participate in the final test phase one day (Exam 1) and seven days (Exam 2) after the reviewing phase. The MCQ that were presented in the two exams were the same and were presented in the same order for all participants. There was no time limit to complete them and no feedback was given.
In both reviewing and final test phases, each question was scored 1 if the answer was entirely correct and 0 when any error was made.

Judgment of learning assessment – We used a similar method as Fernandez and Jamet (2017) to ascertain whether a specific learning strategy had an effect on participants' judgement of learning. At the end of the reviewing phase, each participant had to provide a prospective Judgment Of Learning (pJOL1 in Figure 1) by answering the following questions: "*On a scale of 0-100, what percentage of the information you've learned on DNA do you think you will recall at exam 1 tomorrow?*"; "*On the 40 multiple choice question exam tomorrow, what performance do you expect to reach?* (from 0 to 40/40 correct answers)". Immediately at the end of Exam1, they provided a retrospective JOL (rJOL1 in Figure X) by answering the following questions "*At this first exam, what performance do you expect to reach? (from 0 to 40/40 correct answers)?*"; and a prospective JOL about Exam 2 (pJOL2) success "*On the 40 multiple choice question exam in one week, what performance do you expect to reach?* (from *0 to 40/40 correct answers)?"*. The same rJOL was asked immediately at the end of Exam 2 (rJOL2).

5.4 Analyses

The present study was pre-registered (http://aspredicted.org/blind.php?x=e29eu7). We used generalized linear mixed-effect models (GLMM) with the R package lme4 (Bates et al., 2015). The correctness of each question in the two exams was used as the dependent variable. After a data cleaning step in line with the *a priori* exclusion criteria, we performed a logistic mixed effects regression model with participant and question as random effects. Namely, we excluded data points from final test analyses when : 1) a participant did not finish the learning session (first exposure phase or review phase); 2) a participant did not finish the first final test; 3) a participant correctly answered fewer than 4 of the 5 "catch" questions (within the training quizzes or the exams); 4) a participant did complete the different sessions during the allocated time (e.g., several days after the D-Day). Thus, models were conducted on the accuracy (0 or 1) for each question of the final test quiz. We computed a first model with the three independent variables that were common to all the participants (i.e., Learning Practice, Learning Schedule and Retention Interval) and their interactions (two by two and the triple interaction). Moreover, we were interested in studying the effect of different covariates on the main effects and their interactions. In a secondary analysis, we thus included as covariates socio-demographic data such as age, level of education, and subjective rating of knowledge on DNA. Training times were also included *a priori* because the time spent on the learning contents might moderate the testing and the spacing effects. An analysis restricted to the two retrieval practice groups added the Trained MCQ list factor and the Question Type factor. For these analyses, all available data points were analyzed; no outliers were excluded.

We used a linear regression analysis (ANOVA) for training times, training performance (massed and spaced testing conditions) and JOL analyses with the R package lme4. For the training times and JOL dependent variables, we used Learning Practice, Schedule Type, and Retention Interval as predictors; for the training performance dependent variable, we used Schedule Type and Trained MCQ list as predictors.

Finally, analyses on other interesting predictors that were identified *a posteriori* are presented in the Exploratory Results section (i.e.; inherent difficulty of the training questions).

5.5 Results

Participants' data

From a total of 389 participants who undertook the session at Day 1 (i.e., first exposure phase), 151 participants were excluded in total because: 1) they did not complete the first reading and/or reviewing content phases, or 2) they did not complete the final tests entirely, or 3) they did not take the different phases in the prescribed timeframes, or 4) they did not answer correctly the catch questions (participants who correctly answer fewer than 4 of the 5 "catch" questions in the training quizzes and at both final exams).

Because some participants were not able to do Exam1 at Day 5 for different reasons but wanted to go on with the experiment, we let them do Exam 2. Thus, 5 participants who completed the learning session did only Exam 2 at Day 12.

Thus, we had a total of 238 participants included in the analysis for Exam 1 and 243 participants for Exam 2 (a total of 233 participants did both exams). Demographic data are indicated in Table 1. The 4 groups did not differ in age ($F(3,230) = 1.82$, $p = 0.14$ for Exam 1, $F(3,234) = 1.36$, $p = 0.26$ for Exam 2), nor in education level ($F(3,233) = 1.46$, $p = 0.87$ and $F(3,237) = 1.10$, $p = 0.66$ for Exam 1 and 2 respectively), nor in sex ratio (χ^2 (3) = 5.87, p = 0.12 and χ^2 (3) = 4.64, p = 0.20 for Exam 1 and 2 respectively). Finally, the 4 groups did not differ in their subjective rating of DNA knowledge ($F(3,233) = 1.17$, $p = 0.32$ and $F(3,237) = 1.10$, $p = 0.35$ for Exam 1 and 2 respectively). For the record, the unbalanced sample sizes were partly due to a problem at the beginning of the recruitment at Day1 in the assignment by order of arrival (i.e., at the end of the socio-demographic questionnaire, when a participant gave his/her email address, Didask assigned most of the participant to the two reading learning conditions). But the analyses showed that the four groups were similar in terms of socio-demographic profile.

Table 1. Summary characteristics of the 4 groups of participants included in final test 1 (top) and final test 2 (bottom).

Exam 1	**Massed reading**	**Massed testing**	**Spaced reading**	**Spaced testing**	**Total**
N	84	47	64	43	238
Sex (M/F)	48/36	23/24	25/39	25/18	121/117
Age (years)	37.94 (11.09)	36.89 (12.09)	39.67 (12.55)	42.23 (11.79)	38.98 (11.89)
Education (years)	14.73 (1.39)	14.67 (1.40)	14.40 (1.37)	14.72 (0.80)	14.63 (1.30)
Degree of knowledge on DNA (1-5)	2.24 (0.77)	2.14 (0.75)	2.00 (0.80)	2.11 (0.73)	2.13 (0.77)
Exam 2					
N	85	46	67	45	243
Sex (M/F)	48/37	24/22	28/39	27/18	127/116
Age (years)	38.06 (10.90)	36.88 (12.47)	39.49 (12.24)	41.55 (11.96)	38.89 (11.80)
Education (years)	14.67 (1.37)	14.68 (1.43)	14.43 (1.37)	14.73 (0.78)	14.61 (1.29)
Degree of knowledge on DNA (1-5)	2.20 (0.74)	2.16 (0.71)	1.99 (0.81)	2.15 (0.77)	2.12 (0.76)

Note. Standard deviations are in parentheses

- Pre-registered analyses

Main hypotheses - We estimated a generalized linear mixed effects model with participants and exam questions as random effects; and Learning Practice (baseline = reading); Learning Schedule (baseline = massed); Retention Interval (baseline = Exam 1); and their interactions as predictors with fixed effects. Accuracy to each question of the final test quiz was the dependent variable. Table 2 summarizes the analyses without additional predictors (i.e. covariates).

First, and as predicted, we found a main effect of the Learning Practice predictor (β=0.61, SE=0.14, $p<0.001$). Overall, a participant in the retrieval practice condition was almost twice as likely to answer correctly a question at the final exam than a participant in the reading condition, as indicated by the odd ratio (OR=1.84). The testing effect remained the same at both retention intervals, as indicated by the lack of interaction.

Contrary to our prediction, we did not find a main effect of the learning schedule (β=0.03, SE=0.12, p=0.80, OR=1.03); this was the same at both exams (the interaction with the retention interval factor was not significant).

Moreover, the interaction between the predictors of interest, Learning Practice and Learning Schedule, was not significant: the testing effect was not modulated by the type of schedule (see Figure 3 for the descriptive data). Interestingly, and contrary to our prediction, final performance at Exam 2 did not significantly differ from performance at Exam 1. Nevertheless, the OR<1 with a small interval of confidence indicate that there was a tendency for a subtle decrease of the performance at Exam 2 relative to exam 1. Additionally, the three-way interaction was not significant. Finally, variability in the dependent measure due to the item questions was more important than the variability due to participants (SD_{item} = 0.98 versus; $SD_{participant}$=0.70).

For the record and as complementary results, we provide descriptive data based on effect sizes (Cohen's d). Table 3 shows the effect sizes obtained for the comparisons between the four learning strategies. First, a direct comparison between the reading and the testing groups showed an overall magnitude of the retrieval practice effect of d= 0.66 [0.49, 0.80] ($M_{reading}$ = 53.07% ± 14.04%, $M_{testing}$=64.62% ± 20.48%). Second, a direct comparison showed that the overall spacing effect was small: the massed and spaced schedule groups performed almost equally well (d= 0.15 [-0.006, 0.30], M_{spaced} =60.89% ± 17.83%; M_{massed} =58.15% ± 19.46%). Finally, the magnitude of the retrieval practice effect was the same in the massed schedule (d= 0.60 [0.38, 0.81]) and in the spaced schedule conditions (d= 0.70 [0.47, 0.94]).

Table 2. Estimated coefficients, standard errors, odd ratios and p values for the generalized linear mixed model with final test performance as dependent variable.

Fixed effect	**Estimated coefficients [CI]**	**SE**	**Odds ratio [CI]**	***p value***
Learning practice	0.61[0.34, 0.89]	0.14	1.84[1.4, 2.43]	<0.001
Learning schedule	0.03[-0.2, 0.27]	0.12	1.03[0.81, 1.31]	0.80
Retention interval	-0.11[-0.23, 0.03]	0.07	0.90[0.78, 1.03]	0.12
Learning practice X Learning Schedule	0.19[-0.21, 0.59]	0.20	1.29[0.81, 1.80]	0.35
Learning practice X Retention interval	0.02[-0.18, 0.23]	0.10	1.00[0.83, 1.29]	0.82
Learning Schedule X Retention interval	0.09[-0.12, 0.29]	0.11	1.06[0.89, 1.34]	0.42
Learning practice X Learning Schedule X Retention interval	-0.20[-0.50, 0.11]	0.15	0.85[0.61, 1.11]	0.20

Note. Learning practice (baseline: reading); Schedule (baseline: massed); Retention interval (baseline: exam 1).

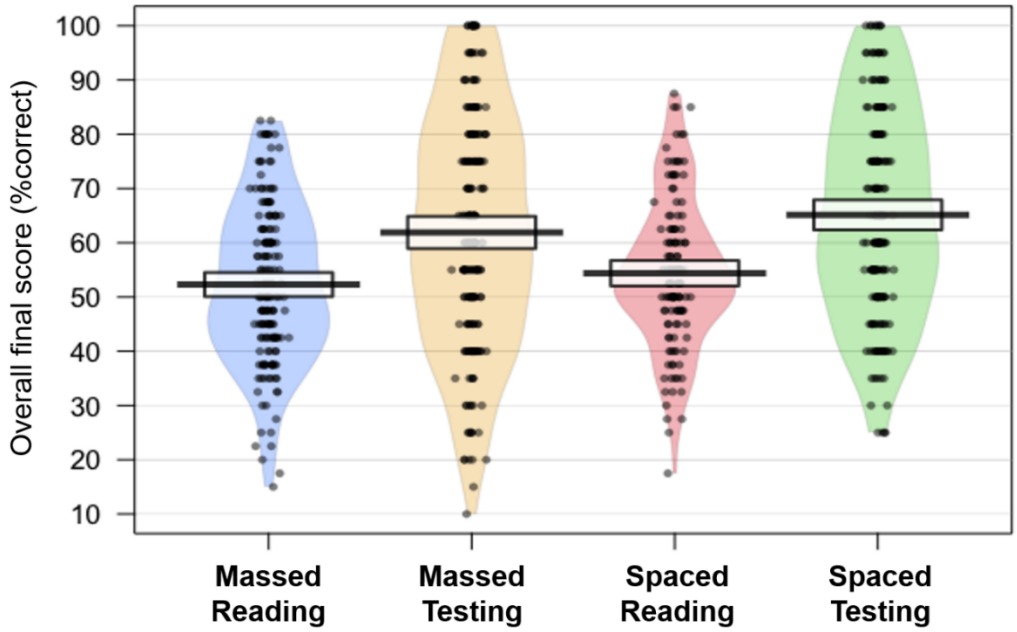


Figure 3. Overall final test performance (percentage of correct answers) for each group of participants (namely: massed reading, massed testing, spaced reading, spaced testing). Each point is a participant. The thick horizontal line represents the median, colour-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Interval.

Table 3. Effect sizes for the planned t-tests comparisons between the different conditions.

Comparisons	Effect sizes
Massed Reading vs Massed Testing	d = 0.60 [0.38, 0.81]
Massed Reading vs Spaced Reading	d = 0.13 [-0.10, 0.36]
Massed Reading vs Spaced Testing	d = 0.81 [0.59, 1.03]
Massed Testing vs Spaced Testing	d = 0.13 [-0.08, 0.33]
Spaced Reading vs Spaced Testing	d = 0.70 [0.47, 0.94]
Massed Testing vs Spaced Reading	d = -0.49 [-0.72, -0.27]

Note. Effect sizes (Cohen's d) are given with 95% confidence intervals in brackets.

As pre-registered, we conducted an analysis restricted to the testing groups, with Learning practice removed, and with the Trained MCQ list and Question type factors added to the same GLMM as presented above. The purpose of the analysis was to assess to what extent different effects were observed for trained vs. untrained questions. Table 4 summarizes the results of this model. There was no significant effect of the Learning Schedule nor of the Trained MCQ list. However, there was a main effect of Question type: participants were less accurate on untrained

questions than on trained questions. Furthermore, there was a main effect of Retention interval: participants were less accurate at Exam 2 than at Exam 1, thus decreasing their final retention. Finally, there was a significant interaction between Retention Interval and Question type in such that being the chance of being accurate on trained questions decreased between Exam 1 and Exam 2 whereas it increased for the untrained questions.

Table 4. Estimated coefficients, standard errors, odd ratios and p values for the generalized linear mixed model restricted to the retrieval practice groups with final tests performance as dependent variable.

Fixed effect	**Estimated coefficients [CI]**	**SE**	**Odd ratio [CI]**	***p value***
Learning Schedule	0.24[-0.03, 0.52]	0.21	1.28[0.96, 1.69]	0.25
Retention interval	-0.35[-0.59, -0.11]	0.11	0.70[0.55, 0.89]	<.001
Question type	-1.49[-1.63, -1.35]	0.10	0.22[0.20, 0.26]	<.0001
Trained MCQ list	-0.01[-0.14, 0.11]	0.19	0.99[0.87, 1.11]	0.94
Learning Schedule X Retention interval	-0.11[-0.51, 0.29]	0.12	0.89[0.60, 1.33]	0.33
Learning Schedule X Question type	-0.007[-0.22, 0.70]	0.12	0.99[0.81, 1.22]	0.95
Retention interval X Question type	0.50[0.29, 0.71]	0.12	1.64[1.34, 2.03]	<.0001

Schedule (baseline: massed); Retention interval (baseline: exam 1), Question type (baseline: trained questions); Trained MCQ list (baseline: list1).

Influence of covariates –The level of education and the degree of knowledge on DNA had a positive and significant effect on the final performance as showed by their respective estimated coefficients (Table 5). For the record, a participant with a master degree is almost twice as likely to have a correct answer to the final test quiz relative to a participant with a high school degree (OR=1.64 [1.11; 2.42]). However, age had a negligible effect on the final performance. Finally, we found no main effect of the total training times (in log(min)) on final performances. Overall, and more interestingly, the socio-demographical information and the time spent on learning the contents did not impact the fixed effects and their interactions. Namely, the main effect of the Learning Practice remained significant whereas the other effects remained not significant (e.g., Learning Schedule and the interaction between Learning Practice and Learning Schedule). When adding these covariates to the GLMM, the OR of the learning practice main effect slightly varied from 1.84 [1.40; 2.43] to 1.57 [1.17; 2.11]. Thus, our results on the final performances are mainly attributed to the manipulated variables, and not to other variables that were not under control in the experimental design.

Table 5. Estimated coefficients, standard errors, and p values for the generalized linear mixed model with final tests performance as dependent variable and the different covariates as predictors.

Fixed Effects & Covariates	Estimated coefficients [CI]	SE	Odd ratio [CI]	*p* value
Learning practice	0.45 [0.15, 0.75]	0.15	1.57 [1.17, 2.11]	<0.01
Learning schedule	0.008 [-0.23, 0.25]	0.12	1.00 [0.80, 1.28]	0.95
Retention interval	-0.10 [-0.25, 0.04]	0.07	0.90 [0.79, 1.04]	0.15
Learning practice X Learning Schedule	0.22 [-0.18, 0.63]	0.21	1.24 [0.83, 1.87]	0.29
Learning practice X Retention interval	0.02 [-0.20, 0.23]	0.11	1.02 [0.82, 1.26]	0.89
Learning Schedule X Retention interval	0.09 [-0.13, 0.30]	0.11	1.09 [0.88, 1.35]	0.43
Learning practice X Learning Schedule X Retention interval	-0.20 [-0.51, 0.11]	0.16	0.82 [0.60, 1.2]	0.21
Age (in years)	0.007 [-0.002, 0.02]	0.004	-	0.12
Level of education - College degree	0.19 [-0.12, 0.51]	0.16	1.21 [0.88, 1.66]	0.23
Level of education - Master degree and more	0.49 [0.10, 0.89]	0.20	1.64 [1.11, 2.42]	0.013
Degree of knowledge on DNA (Likert scale 1-5)	0.15 [0.021, 0.27]	0.06	-	0.022
Total training time (log(min))	0.20 [-0.05, 0.45]	0.13	-	0.12

Note. Level of education (categorical) *Baseline:* High school graduate (GDE)

Secondary analyses - This analysis concerned participants who did exams 1 and 2.

Training times **-** We computed the total time spent doing the first reading phase + reviewing session phase (i.e. the cumulative time spent on learning contents and on quizzes) for each of the four learning conditions. For technical reason, training times could not be calculated for 16 participants. Moreover, some training times were abnormally high, probably reflecting that the participant was doing something else while keeping the window open. We thus excluded outlying data points (mean ± 2.5SD) and we replaced them by the mean of their respective group.

Because distributions of times are skewed, statistics were computed on the natural logarithm of total training time as dependent variable, as is usually recommended. We used Learning Practice and Learning Schedule as predictors. There was a main effect of the learning practice on the time spent on contents, $F(1,218)=55.14$ $p<.001$, $\eta^2 = 0.20$; as well as a main effect of the learning schedule $F(1,218)=4.67$, $p=0.03$, $\eta^2 = 0.02$.

The amount of time spent on training was longer for the groups in the testing than in the reading conditions (Table 6). Moreover, the two groups who were assigned to a spaced schedule spent more time on average on the contents. As the time spent on the first reading did not differ between the 4 groups ($Mean_{MR}$=10.94 ±5.03min, $Mean_{MT}$=10.82± 4.87min, $Mean_{SR}$=9.96 ±3.52min, $Mean_{ST}$=10.20± 3.39min) the main effect of the two variables was due to time spent in the reviewing phase. Taking quizzes took more time than re-reading the same contents, and reviewing once per day took more time than doing the 3 episodes in one single session. Finally, we found no interaction between learning practice and schedule F(1,218)=0.001 p=0.53, η^2 <0.01.

Table 6. Time spent in the training phase (first reading phase + reviewing phase) in minutes for each learning conditions.

Learning conditions			
Massed reading	Massed testing	Spaced reading	Spaced testing
35.97 min (14.99 min)	50.74 min (18.37 min)	41.20 min (22.65 min)	58.44 min (18.70 min)

Note. Standard deviations are in parentheses.

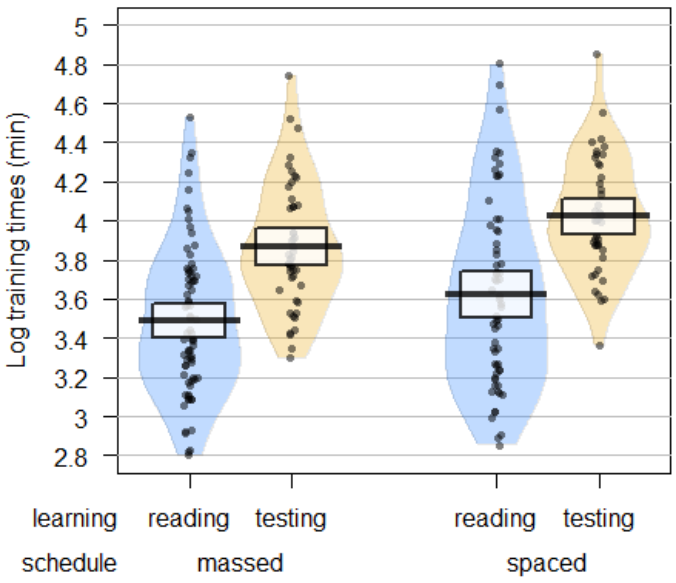


Figure 4. Total training times (time spent doing the first reading phase + reviewing session phase) in log(min) as a function of learning practice and learning schedule. Each point is a participant. The thick horizontal line represents the median, colour-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Interval.

Training performance in the testing conditions - There was a main effect of schedule on training performance, $F(1,82)=5.19$, $p=0.03$, $\eta^2 =0.059$. Participants in the massed testing group had a better overall performance during the training quizzes than those in the spaced testing group ($Mean_{MT}$ = 69.63% ± 14.61%; $Mean_{ST}$ = 63.68% ±13.22%). The Trained MCQ List had no significant effect ($F(1,82)=0.01$, $p=0.92$, $\eta^2 <0.001$): both lists led to the same performance, reflecting a similar level of difficulty ($Mean_{List\ 1}$ = 67.34% ± 14.36%; $Mean_{List\ 2}$ = 66.84% ± 13.89%). There was no significant interaction between the two factors: the massed testing group performed better than the spaced testing group ($F(1,82)=1.43$, $p=0.24$, $\eta^2 = 0.017$. The Table 7 shows that the two groups performed equally well at the first attempt, because it was the first time the participants did this quiz. However, across time, the score of the massed testing group increased faster than the spaced testing group until they reached a ceiling effect (Table 8). Figure 5 shows the improvement of the score for each testing groups across the three attempts of the reviewing session, as well as the final score at both exams for all the learning conditions.

Table 7. Training performance for retrieval practice conditions as a function of the Trained MCQ List.

Massed testing		Spaced testing	
List 1	List 2	List 1	List 2
68.49% (16.70%)	72.27% (11.22%)	65.83% (10.90%)	62.41% (14.46%)

Note. Standard deviations are in parentheses.

Table 8. Training performance for retrieval practice conditions in function of the attempt: the first, second, and third one.

Massed testing			Spaced testing		
Attempt 1	Attempt 2	Attempt 3	Attempt 1	Attempt 2	Attempt 3
50.11% (15.94%)	72.55% (19.32%)	86.22% (15.56%)	51.86% (15.99%)	65.23% (15.70%)	73.95% (15.72%)

Note. Standard deviations are in parentheses.

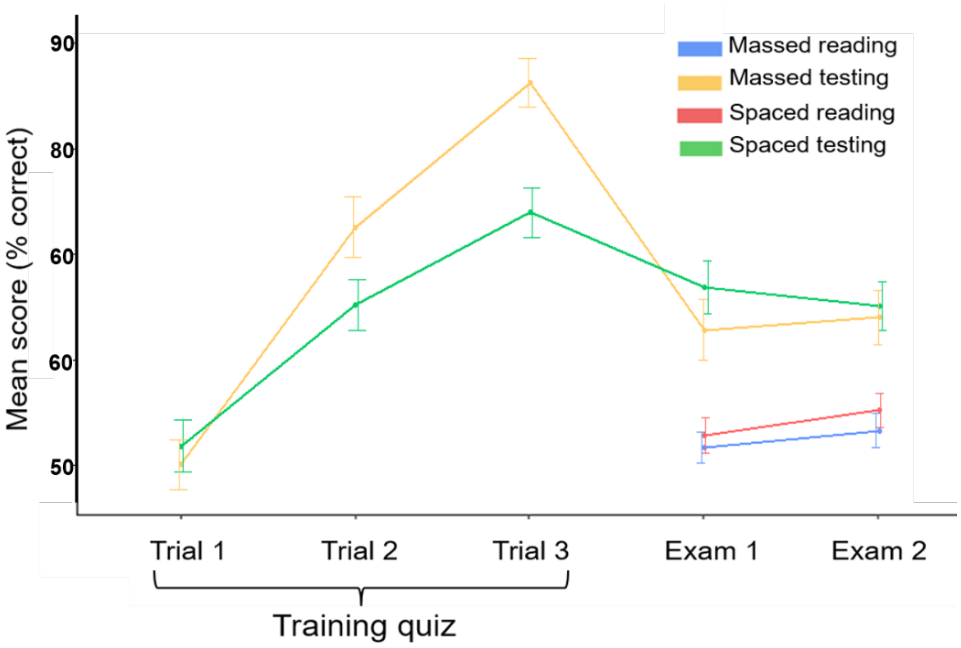


Figure 5. Score (% of correct responses) for each group and experimental phase (review sessions and exams), with error bars representing the standard error of the mean.

Judgment of learning –Each participant gave two JOL per exam: a prospective one (pJOL) and a retrospective one (rJOL). For each measure and each participant, we calculated the difference (Δ) between the estimated and the actual score (out of 40). A positive delta reflects an overconfident estimation while a negative delta reflects an underconfident estimation. Thus, for each participant we had four ΔJOLs (for those who completed the experiment until the end). We ran four separate ANOVAs with each of the four ΔJOL as dependent variable; with Learning Practice and Learning Schedule as independent variables. Table 9 summarizes the descriptive data for each ΔJOL and Table 10 shows the results of the ANOVA for each ΔJOL.

Overall, we found no significant effect of the learning practice or schedule, and no significant interaction, neither for the ΔpJOLs nor for the ΔrJOLs, at neither exam. At the beginning of Exam1, participants tended to overestimate their score in all learning conditions (overall mean: ΔpJOL=5.44 ±7.24). Participants in the two reading groups were numerically more overconfident than the participants in the two testing groups ($M_{reading}$ = 6.17 ± 7.55, $M_{testing}$ = 4.27 ± 6.56) but there only a tendency for a main effect of the Learning Practice, this variable was actually not significant. Second, results on ΔrJOLs indicated that overall overestimation that was measured before Exam 1 vanished after the exam (overall mean ΔrJOL=0.31 ± 7.00).

The same pattern of results was found for JOLs concerning Exam 2 (overall mean ΔpJOL 2 = 3.78 ± 8.36 versus overall mean ΔrJOL2 =-0.78 ± 7.78).

Finally, we analyzed *"On a scale of 0-100, what percentage of the information you've learned on DNA do you think you will recall at exam 1 tomorrow?"* asked at the end of the review session at Day 4. We found a main effect of the Learning Practice, $F(1, 209)=13.45$, $p<.001$, with a subjective of amount of information for the recall at exam being higher in the two testing groups relative to the two reading groups ($M_{testing}$ =74.40 % ±15.80% ; $M_{reading}$= 64.84% ±19.88%). Correlation analysis between this metamemory JOL and final performance at Exam 1 indicated that the judgement made by the two testing groups seemed more strongly and positively related to their final performance than the two reading groups ($r=0.52$, $p<.001$ versus $r=0.36$, $p<.001$ respectively).

However, we found no main effect of the learning schedule, and no significant interaction.

Table 9. Means of the ΔJOL per learning group and per exam.

	Learning Practice X Learning Schedule			
Exam 1	Massed Reading	Massed Testing	Spaced Reading	Spaced Testing
ΔpJOL	6.14 (7.40)	4.76 (7.42)	6.20 (7.83)	3.74 (5.53)
ΔrJOL	0.48 (7.44)	0.28 (6.57)	0.30 (7.61)	0.03 (5.80)
Exam 2				
ΔpJOL	4.12 (8.20)	2.67 (8.31)	3.60 (9.35)	4.55 (7.23)
ΔrJOL	-0.52 (8.23)	-2.29 (6.77)	-0.26 (8.43)	-0.50 (6.89)

Note. Standard deviations are in parentheses.

Table 10. Linear regression models with ΔpJOLs and ΔrJOLs as dependant variables; and the Learning Practice and the Schedule as independent variables for each exam.

	Exam 1			
	Num DF	*Den DF*	*F Value*	*p-value*
ΔpJOL				
Learning Practice	1	209	3.46	0.064
Schedule	1	209	0.12	0.73
Learning Practice X Schedule	1	209	0.28	0.60
ΔrJOL				
Learning Practice	1	220	0.070	0.79
Schedule	1	220	0.054	0.82
Learning Practice X Schedule	1	220	0.001	0.98

	Exam 2			
	Num DF	*Den DF*	*F Value*	*p-value*
ΔpJOL				
Learning Practice	1	226	0.064	0.80
Schedule	1	226	0.11	0.74
Learning Practice X Schedule	1	226	1.09	0.30
ΔrJOL				
Learning Practice	1	233	0.91	0.34
Schedule	1	233	0.69	0.41
Learning Practice X Schedule	1	233	0.53	0.47

- Exploratory analyses

The significant question random effect revealed that questions varied enormously in difficulty. Indeed, across the entire set of 40 MCQ, performance on the first attempt varied from 3.33% to 95% of participants answering correctly. Thus, it seemed interesting to take question difficulty into account in order to explain more variance in performance. Because the % participants correct for each question did not have a normal distribution, we used a median split to categorize each question as "easy" or "difficult". We then computed the same GLMM as presented in the pre-registered section, with Question difficulty as an additional factor. We found a main effect of Question difficulty, with the easiest MCQ that increased by 8.92 times (CI [7.79; 10.20]) the chance of correctly answering relative to the difficult MCQ (β=2.19, SE=0.07, $p<0.001$). All the other main effects and interactions remained the same: only the Learning Practice predictor reached significance (β=0.95, SE=0.11, $p<0.001$) with an odds ratio of 2.59 (CI [2.08; 3.22]). Most interestingly, there was an interaction between Learning practice and Question difficulty (β=-0.50, SE=0.08, $p<0.001$) : the testing effect was larger for the difficult than for the easy MCQ, probably owing to a slight ceiling effect in the latter case (Figure 6).

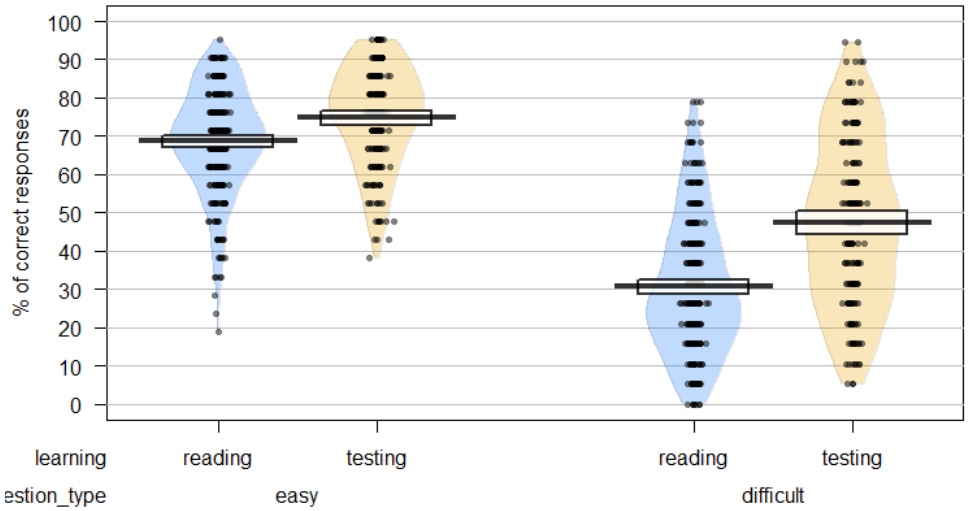


Figure 6. Final exam score (percentage of correct responses) as a function of learning practice (reading versus testing) and of Question difficulty (easy versus difficult). Each point is a participant. The thick horizontal line represents the median, colour-shaded bean plots show the full distribution of the data, and boxes represent 95% Confidence Interval.

5.6 Discussion

Summary of the main results

The present study aimed at investigating two research questions in one single experiment: i) what is the relative effect size of retrieval practice and spacing effects? ii) are retrieval practice and spacing effects additive or do they interact? To do so, we tried to reproduce a learning context as realistic as possible that included an exposure phase similar to the reading of a textbook chapter, then a phase during which the learners had to review the content following a specific learning strategy, with the aim to take an exam 4 days after first exposure. Longer-term retention was also assessed another 7 days later, in order to distinguish short- vs.long-term benefits. First, we estimated the relative benefits of the two well-known testing and spacing effects on memory retention for educational contents. Our results showed that the testing effect was large (d=0.84) and persisted one week later (d=0.70), while the spacing effect was much smaller and did not reach significance at either exam (d=0.17 and 0.14). Furthermore, there was no significant interaction between learning practice and learning schedule: the benefit of learning with testing had the same magnitude whether using a massed or a spaced schedule, and this was the case at the two retention intervals.

Comments on the testing effect

As predicted and in line with previous findings, we found a robust benefit of retrieving new information from memory after a reading session of the learning content in comparison to repeated reading. We replicated the same large effect size (overall testing effect of d= 0.66 [0.49, 0.80]) as found by Rowland in his meta-analysis (2014) of between-subject experiments (g=0.69). In terms of odd ratio, a participant in the retrieval practice condition was almost twice as likely to answer correctly a question in the final exam as a participant in the reading condition.

Moreover, the testing effect found in this experiment remained almost unchanged when we included covariates in the analyses. Contrary to previous experiments where training time was closely matched across conditions, in the present experiment participants were free to spend as much time as they felt necessary on the contents (chapter contents, quizzes, and feedbacks). This induced important differences in training time between conditions, with participants in the repeated reading conditions spending much less time in training than participants in the repeated retrieval practice conditions (39 min versus 55 min respectively). As we suspected this variable to be related to final performance, we computed covariance analyses adjusting for training time. However, this variable did not have a significant effect on final retention and had only a subtle influence on the odds ratio of the testing effect (OR=1.84 without and OR=1.64 with this covariate in the GLMM). Thus, and in line with a previous study, the overall testing effect cannot be attributed to the extra time spent on training (Latimier et al., 2019). Finally, the analyses restricted to the two retrieval practice conditions did not show a main effect of schedule and of retention interval. Nevertheless, there was a main effect of the question type variable: final performance was higher for trained questions relative to untrained questions; which makes sense because the knowledge addressed in the trained questions were seen three times during the reviewing phase, while the knowledge addressed in the untrained questions were seen only during the first exposure phase at Day1. More interestingly, while performance on the trained questions suffered from forgetting between the two exams as found in several studies (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger, Agarwal, McDaniel, & McDermott, 2011; Roediger & Karpicke, 2006); participants performed better on the untrained question at Exam 2 relative to Exam 1. This might be explained by the fact the first exam served as a training session for this question type, thus participants progressed between the two exams.

Comments on the spacing effect

Whereas our replication of the testing effect fits with the literature, we did not replicate the spacing effect on overall performance. Contrary to our predictions, participants who learnt following a spaced schedule (spaced reading and testing) during the reviewing session did not perform better at the final exams than those in the massed conditions (massed reading and testing). A spacing effect was found neither in the repeated reading conditions (d=0.13) nor in the repeated testing conditions (d=0.13); and regardless of the retention interval. Even if our results seem to contradict the literature on the benefits of the spaced learning, they fit with several studies in which the spacing effect was not found. Very recently, in a study conducted in young children with educational contents, Greving and Richter (2019) did not find a benefit of distributed rereading (inter-study interval of 1 week) over the massed schedule at a one week interval. However, contrary to us, they found a difference in favour of the massed schedule at a short-term retention interval. One possible explanation for the discrepancy is that their first exam was immediately after the second reading session, while in our case it was one day later. Thus, some forgetting may have occurred between the reviewing session and Exam 1, cancelling any short-term advantage of the massed schedule.

In their study including three different experiment on spaced rereading, Rawson and Kintsch (2005) did not find a benefit of the spaced conditions (inter-study interval of 1 week) at an immediate final test, but only at a delayed final test 2 days after the learning phase. They suggested that text length may account for this absence of spacing effect at short term retention. They hypothesized that text length of the learning content may influence the relative amount of time spent studying the text depending on the type of schedule to restudy the contents : learning according to a massed pattern decreases the time spent on restudying the contents relative to a spaced pattern. The effect of difference in reading times might appeared only at long term but not at short term interval. This hypothesis cannot explain the results of the present experiment. Indeed, even if the total training time spent on learning the contents was less important in average in the two massed learning conditions than in the two spaced conditions (43.36 min versus 49.82 min respectively), this variable had no significant impact on final performance and on the effect of the learning schedule predictor too (putting the total training times in the GLMM did not enable to reach a significant spacing effect). Another explanation of the absence of a spacing effect may lie in the inter-study interval/retention interval interaction. As stated by Cepeda et al. (2008): “The optimally efficient gap between study sessions is not some absolute quantity that can be recommended, but rather depends dramatically on the RI (a point that was evident in the short-term studies, such as that by Glenberg (1976), and is now shown to extend

to far greater time intervals). To put it simply, if you want to know the optimal distribution of your study time, you need to decide how long you wish to remember something". Indeed, in one experiment Cepeda et al. (2009) showed that this optimum is obtained for a constant fraction for retention interval around 10 to 20% of this interval between the study phase and the final test. In the present experiment, Exam 1 is very close to the end of the reviewing session, the lag between each repeated reviewing session did not correspond to the optimal fraction of this very short retention interval of 24h. Exam 2 was a better retention interval (8 days) to test this hypothesis. However, the first exam may have boosted performance at Exam 2 more for the massed than for the testing groups. Thus better conditions to observe a larger spacing effects might have been 1) to remove Exam 1 to avoid participants in the spacing conditions benefiting from one instance of retrieval practice; 2) increasing even more the retention interval before Exam 2.

To sum up, our results on spaced learning do not fit with the assumption that massed practice may yield greater performance than distributed practice when learning is tested immediately, whereas distributed practice yields greater performance than massed practice on retention tests (Schmidt & Bjork, 1992). Various factors such as the inter-study interval, the retention interval and the type of learning material (factual versus high level knowledge) may moderate this effect.

Interaction of spacing and testing effects

One of the main interest raised by our study was the possible interaction between testing and spacing (i.e., addition, or potentiation). However, we did find any. The benefit of learning with retrieval practice relative to learning with reading had the same amplitude if the retrieval episode were arranged according to a massed (massed testing effect of d=0.60) or a spaced pattern (spaced testing effect of 0.70). In addition, while we found an important benefit of the spaced retrieval practice over the massed one in our recent meta-analysis (d=0.74, Latimier, Peyre, & Ramus, unpublished); the benefit measured here was almost null and not significant (d=0.13). Our results do not fit with the abundant literature on this comparison, as was the case for the spacing effect. For example, Karpicke & Bauernschmidt (2011) found that massed repeated recalling was not better than a single recall condition; but introducing spacing between the three repeated recall enhanced significantly the final retention (i.e., long lags of spacing between repeated recall produced a 200% gain in long-term retention in comparison with the massed condition). Recently, Dosbon & al. (2016) demonstrated a robust « spacing of tests effect » with educational contents. They found that massed retrieval practice produced the

highest immediate recall scores, and there was a striking 81% decline in those scores at delayed retention and the massed practice resulted in the poorest recall one week and one month after the learning phase. Finally, Petersen-Brown et al. (2019) replicated the benefit of spaced over massed retrieval practice with a very young population in mathematics vocabulary learning (d=0.72 with a uniform schedule and d=0.55 with the expanding one).

Several explanations might account for these absence of a superiority of the spaced retrieval schedule over the massed pattern. One major difference between other studies on the topic was the realistic schedule we used during the reviewing session. Usually, experiments included the massed retrieval practice session immediately after the initial exposure phase, this is not the case here because we decided to put the massed review session with a certain lag (2 days) after the initial exposure to contents, but close to the first exam. The space we provided between the initial exposure to the learning content and the first retrieval attempt might have been beneficial in a sense, because the participants had to refresh their memory on what they read 3 days earlier. Thus, they had to produce a stronger retrieval effort than participants in the spaced testing condition whose first retrieval attempt was closer in time to the reading session (Figure 5 showed a slightly better retrieval performance of this group). And this explanation can account for the absence of on overall spacing effect, because the same lag was provided in the massed reading condition. Thus, as the effort to remember the information was more or less equated between the massed and spaced conditions during the first retrieval attempt or reading of the learning phase, this can explain the similar final performances between the two schedules. Numerically, the benefit of spaced testing over massed reading (d=0.81) was slightly larger than the sum of the benefit of massed testing over massed reading and that of spaced testing over massed testing (0.60+0.13=0.73). This suggests that combining spacing and testing might have superadditive effects. However, the small size of the spacing effect in our experiment made our results unsuitable to fully evaluate this possibility. Only experiments with substantial effects of both testing and spacing will be able to properly assess the interaction.

Judgments of learning findings

First, the analyses on the JOLs showed a generalized overconfidence before both exams. The fact that learners overestimate their ability to remember new knowledge is universal (Kornell & Bjork, 2008). Indeed, results on the prospective JOL showed that all the participants overestimated their final recall performances. The results on retrospective JOL were quite different. The overall overestimation that was found before Exam 1 was not measured anymore

after Exam 1. We observed a similar pattern for Exam 2. This result is interesting and reflects the benefit of being tested on metamemory judgment (Robey, Dougherty, & Buttaccio, 2017). Even though analyses showed only a tendency for a main effect of the type of learning practice, the results on prospective JOL before Exam 1 showed that participants in the repeated testing groups tended to be less overconfident in their ability to recall recently learnt information than the repeated reading groups (Fernandez & Jamet, 2017; Huelser & Metcalfe, 2012; Little & McDaniel, 2015).However, the tendency for a smaller illusion of mastery in the retrieval practice conditions did not persist before Exam 2.

Another interesting result was that participants in the repeated testing groups perceived being able to recall significantly more information at the end of the reviewing session than the reading groups. This result suggests that participants who learnt with quizzes may have appreciated the benefits of this learning practice on the amount of information they trained well. One explanation to interpret this result could be that participants who learnt with quizzes benefited from receiving feedback after each questions (Butler & Roediger, 2008; Sitzman, Rhodes, & Kornell, 2016). Thus, they were more aware of their ability to recall a certain amount of information they mastered and used their training performance as a cue to estimate their level of mastery. By contrast, those who reviewed the contents through repeated readings did not have an objective cue of their level of mastery and could only use their own perception to estimate how well they learnt (Bjork et al., 2013; Brown, Roediger (III), & McDaniel, 2014). Finally, contrary to results from some a recent study (Greving and Richter, 2019), the type of learning schedule did not modulate the prospective and retrospective JOLs neither at Exam 1 nor at Exam 2.

Comments on exploratory analyses

The fact that we created a set of training questions with an important variability of difficulty allowed us to answer a question *a posteriori* : how does the inherent difficulty of the training question modulate the benefit of learning with quizzes? To our knowledge, this question has not been raised so far. We found one study in which the item difficulty was manipulated within participants (easy versus difficult word pairs) as well as the learning practice (tested versus studied word pairs; Minear, Coane, Boland, Cooney, & Albat, 2018). Overall, their results did not show a significant interaction between item difficulty and learning practice: they found a testing effect of 8% for easy items and 6% for difficult items. However, they found that the benefit of retrieval practice varies with item difficulty and participant abilities (i.e., “difficulty had differential effects on participants who scored high or low on a measure of gF, with high

gF participants showing larger testing effects for difficult over easy items, whereas low gF participants showed the opposite").

In our, not surprisingly easy MCQ were recalled more than difficult MCQ, for all participants in both learning practice conditions. Even if we only provided exploratory results, we also found the testing effect was not the same whatever the level of difficulty of the trained questions. The present study did not aim at answering the question of the interaction between MCQ difficulty and the type learning practice. Further research is needed to explore this interaction by taking into account individuals differences, and different type of materials and retrieval formats because it is very interesting to explore the optimal level of difficulty that increases the power of retrieval practice (*retrieval effort theory*, e.g., Pyc & Rawson, 2009).

Limitations

One factor that may have hindered the observation of a significant spacing effect is the very large inter-individual variability in test results, within each condition. For instance, in the massed testing condition, final test scores ranged from 10 to 100% correct, probably reflecting a large heterogeneity of both ability and motivation in our participants. For this purpose, we ran the analyses again with covariates such as education level, age, prior knowledge on DNA and training time, which one would expect to partly explain this variability. Those covariates did explain some variance in final test scores, however their introduction had nearly no impact on the effects of learning practice and schedule (see Results section).

Our conclusions might be limited to a certain type of material (scientific text), mode of presentation (an e-learning platform), and to a certain type of participants, i.e., workers doing a paid task, rather than students. It will be important to address the same question using other learning contents and populations, especially school children in a more ecological context on a longer timescale. Further research comparing a cramming situation like ours and a massed review session following immediately the course without any lag would be useful. It might be possible that the retention interval at Exam 1 and Exam 2 were not enough long to detect a subtle spacing effect. A third exam place one month after the reviewing session might have increased the chances of observing a spacing effect.

Nevertheless, we see no reason to suspect that the effects reported here would not generalise.

Practical Implications

Overall, the results of this study provide strong evidence for the systematic use of retrieval practice during the reviewing phase, in line with previous best practices recommendations (e.g.,

Benjamin & Pashler, 2015; Weinstein, Madan, & Sumeracki, 2018). Even if learning outcomes at the longer retention interval were on par for massed and spaced testing in the present experiment, we strongly recommend to use the space schedule because of the high reproducibility of the effect in the literature (Kang, 2016).

Altogether, the results on JOLs indicated the importance of taking into account such JOL in experiments that compared different learning strategies. The participants in the retrieval practice conditions showed a certain sensitivity to the benefits of learning with repeated quizzes (Logan, Castel, Haber, & Viehman, 2012). Thus, we strongly recommend to use repeated quizzes as frequently as possible in a formative assessment perspective but also as a self-assessment tool.

5.7 References

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. https://doi.org/10.3102/0034654316689306

Baddeley, A. D., & Longman, D. J. A. (1978). The Influence of Length and Frequency of Training Session on the Rate of Learning to Type. *Ergonomics*, *21*(8), 627–635. https://doi.org/10.1080/00140137808931764

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

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of Frequent Classroom Testing. *The Journal of Educational Research*, *85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818

Benjamin, A. S., & Pashler, H. (2015). The Value of Standardized Testing: A Perspective From Cognitive Psychology. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 13–23. https://doi.org/10.1177/2372732215601116

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* (New York: Worth Publishers, pp. 56–64).

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Brown, P. C., Roediger (III), H. L., & McDaniel, M. A. (2014). *Make It Stick*. Harvard University Press.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41. https://doi.org/10.1016/j.cedpsych.2008.07.001

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, Shana 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

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

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, William 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.

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

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

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. *Metacognition and Learning*, *12*(2), 131–156. https://doi.org/10.1007/s11409-016-9163-9

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. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 1–16. https://doi.org/10.1016/S0022-5371(76)90002-5

Glover, J. A., & Corkill, A. J. (1987). Influence of Paraphrased Repetitions on the Spacing Effect. *Journal of Educational Psychology*, *79*(2), 198–199.

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 (Hove, England)*, *22*(5), 462–469. https://doi.org/10.1080/09658211.2013.798416

Greving, C. E., & Richter, T. (2019). Distributed Learning in the Classroom: Effects of Rereading Schedules Depend on Time of Test. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02517

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

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

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*(4), 514–527. https://doi.org/10.3758/s13421-011-0167-z

Kalenberg, K. (n.d.). *Spaced and Expanded Practice: An Investigation of Methods to Enhance Retention*. Retrieved from http://cornerstone.lib.mnsu.edu/cgi/viewcontent.cgi?article=1205&context=jur

Kang, S. H. K. (2016). Spaced Repetition Promotes Efficient and Effective Learning Policy Implications for Instruction. *Policy Insights from the Behavioral and Brain Sciences*, 2372732215624708. 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*. https://doi.org/10.3389/fpsyg.2016.00350

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

Kornell, N. (2009). Optimising Learning Using Flashcards: Spacing Is More Effective Than Cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. https://doi.org/10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits - and costs - of dropping flashcards. *Memory (Hove, England)*, *16*(2), 125–136. https://doi.org/10.1080/09658210701763899

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915

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

Latimier, A., Riegert, A., Peyre, H., Ly, S. T., Casati, R., & Ramus, F. (2019). *Does pre-testing promote better retention than post-testing ?* https://doi.org/10.7910/DVN/XPYPMF

Latimier, A., Peyre, H., Ramus, F. A meta-analysis of the effect of spaced retrieval practice on memory retention (*unpublished*).

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition*, *43*(1), 85–98. https://doi.org/10.3758/s13421-014-0453-7

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

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. (2019). How the Amount and Spacing of Retrieval Practice Affect the Short- and Long-Term Retention of Mathematics Knowledge. *Educational Psychology Review*. https://doi.org/10.1007/s10648-019-09489-x

Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty. *Memory & Cognition*, *43*(5), 760–774. https://doi.org/10.3758/s13421-014-0499-6

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

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

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. (2013). 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. (2013). 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

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five Popular Study Strategies: Their Pitfalls and Optimal Implementations. *Perspectives on Psychological Science*, *13*(3), 390–407. https://doi.org/10.1177/1745691617710510

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

Petersen- Brown, S., Lundberg, A. R., Ray, J. E., Paz, I. N. D., Riss, C. L., & Panahon, C. J. (2019). Applying spaced practice in the schools to teach math vocabulary. *Psychology in the Schools*, *56*(6), 977–991. https://doi.org/10.1002/pits.22248

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

Pyc, Mary A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Rawson, K., & Kintsch, W. (2005). Rereading Effects Depend on Time of Test. *Journal of Educational Psychology*, *97*, 70–80. https://doi.org/10.1037/0022-0663.97.1.70

Rhodes, M. G. (2016). Judgments of Learning. *The Oxford Handbook of Metamemory*. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

Rhodes, M. G., & Castel, A. D. (2008). Memory Predictions Are Influenced by Perceptual Information: Evidence for Metacognitive Illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625.

Robey, A. M., Dougherty, M. R., & Buttaccio, D. R. (2017). Making Retrospective Confidence Judgments Improves Learners' Ability to Decide What Not to Study. *Psychological Science*, *28*(11), 1683–1693. https://doi.org/10.1177/0956797617718800

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-Enhanced Learning in the Classroom: Long-Term Improvements From Quizzing. *Journal of Experimental Psychology-Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

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

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

Schmidt, R. A., & Bjork, R. A. (1992). New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science*, *3*(4), 207–217. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society*, *30*(2), 125–128. https://doi.org/10.3758/BF03330416

Sitzman, D. M., Rhodes, M. G., & Kornell, N. (2016). The influence of feedback on predictions of future memory performance. *Memory & Cognition*, *44*(7), 1102–1113. https://doi.org/10.3758/s13421-016-0623-x

Son, L. K., & Simon, D. A. (2012). Distributed Learning: Data, Metacognition, and Educational Implications. *Educational Psychology Review*, *24*(3), 379–399. https://doi.org/10.1007/s10648-012-9206-y

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

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). https://doi.org/10.5688/ajpe6179

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, *46*(7), 1164–1177. https://doi.org/10.3758/s13421-018-0828-2

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). Retrieved from https://search.proquest.com/docview/1297338409/citation/99A7F419CA124520PQ/1

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, *15*(1), 41–44. https://doi.org/10.3758/BF03329756

‘That’s the most important piece of evidence we’ve heard yet,’
said the King, rubbing his hands ;
‘so now let the jury – ’

— Lewis Carroll
Alice's Adventures in Wonderland (1865, 1869), Chapter XII *Alice’s evidence.*

6 Chapitre 6 : Discussion générale

6.1 Résumé et apports des différentes études

6.1.1 Résultats principaux du Chapitre 2 (Méta-analyse)

Quelle est ampleur du bénéfice des révisions avec récupération en mémoire espacées dans le temps sur la rétention de nouvelles connaissances ?

Nous avons répondu en partie à cette première question en quantifiant la taille d'effet du bénéfice sur la rétention à long terme de l'apprentissage par des tests répétés et espacés par rapport à deux autres stratégies d'apprentissage ne combinant pas la récupération en mémoire et l'espacement. Dans la méta-analyse concernant la première compilation d'études, **nous avons trouvé un effet robuste et significatif d'un apprentissage combinant les deux stratégies de révision par rapport à un apprentissage consistant en des épisodes de récupération en mémoire des connaissances massées dans le temps (g = 0.74 [0.55, 0.92], p<0.01,** avec correction pour le biais de publication). Autrement dit : rajouter un intervalle de temps, de l'ordre de la minute ou de la semaine, entre les répétitions d'un même test d'apprentissage améliore le bénéfice des révisions avec récupération en mémoire. Malgré une hétérogénéité importante parmi les comparaisons compilées pour calculer cette taille d'effet, nous n'avons pas fait émerger de variable qui modulerait l'ampleur de ce bénéfice ; à l'inverse de précédentes méta-analyses sur le bénéfice de l'apprentissage par récupération en mémoire (Adesope et al., 2017a; Rowland, 2014). **Par ailleurs, le bénéfice de l'apprentissages avec tests espacés dans le temps par rapport à des relectures espacées n'a pu être démontré (g = 0.46 [-0.41, 1.33], p = 0.15), compte tenu de l'absence de significativité du test à un seuil élevé** (deuxième compilation d'études). Le nombre de comparaisons compilées pour cette analyse était certainement le facteur limitant en plus de la très forte dépendance des effets individuels entre eux (16 comparaisons issues de 3 études seulement). Les conclusions sont donc incertaines pour cette partie des analyses.

Quel est le planning d'espacement préférable pour planifier les épisodes de récupération en mémoire ?

La troisième méta-analyse a permis d'apporter un éclairage quantitatif sur la supériorité supposée du planning expansif de l'apprentissage avec tests répétés relativement au planning uniforme. Alors que les revues de littérature semblaient s'accorder sur une meilleure optimisation de l'apprentissage espacé par un accroissement des intervalles de temps entre

chaque session de révision, notre méta-analyse ne valide pas cette hypothèse pourtant fondée sur des arguments théoriques et empiriques solides (Balota, Duchek, & Logan, 2007; Roediger III & Karpicke, 2011). **Etant donné la robustesse de notre mesure (g = 0.032 [-0.10, 0.17], p = 0.62), ces résultats suggèrent que l'utilisation d'un planning expansif ne permet pas une meilleure rétention finale des nouvelles connaissances que le planning uniforme**, malgré son net bénéfice pour permettre une progression plus rapide lors des sessions d'apprentissage par les tests (e.g., Kang, Lindsey, Mozer, & Pashler, 2014). Enfin, l'absence d'hétérogénéité entre les études incluses montre que pour comparer ces deux plannings les conditions expérimentales manque de diversité, aussi bien dans la population choisie (principalement des étudiants à l'université), que dans le type de contenu à apprendre (en général très simple et artificiel) et la temporalité des différentes phases expérimentales (intervalles de temps courts par rapport à la réalité, entre les sessions de tests répétés, et avant l'examen final).

6.1.2 Résultats principaux du Chapitre 3 (Expérience 1)

De quelle ampleur sont les bénéfices des deux types de positionnement des tests d'apprentissage (pré-test versus post-test) ? Dans quelles conditions sont-ils obtenus ?

Cette première expérience a permis de comparer dans un même design expérimental deux positions des tests d'apprentissage par rapport à la lecture des contenus pédagogiques alors qu'aucune donnée récente n'avait été produite sur la question. Dans cette étude en ligne incluant une population conséquente (285 participants), nous avons mesuré la rétention à long terme de nouvelles connaissances (connaissances scientifiques niveau baccalauréat) via un examen placé une semaine après la session d'apprentissage sur Didask. Restreintes aux scores de mémorisation obtenus pour les questions déjà vues dans les tests d'apprentissage, les analyses montrent effectivement un avantage significatif à apprendre avec des pré-tests placés avant la lecture de chaque module du cours par rapport à la méthode classique de relecture sans aucun test d'apprentissage. **Nous avons donc répliqué le bénéfice des pré-tests d'apprentissage, déjà démontré dans la littérature. Nous avons également retrouvé sur ces mêmes questions le bénéfice significatif bien plus connu de l'apprentissage par les tests placés après les contenus à lire (effet de post-test tel qu'il est appelé dans cette étude).** Les tailles d'effet obtenus des deux effets étaient proches de celles retrouvées dans d'autres études (Adesope et al., 2017; Carpenter & Toftness, 2017; Richland, Kornell, & Kao, 2009;

Rowland, 2014): nous avons trouvé un bénéfice moyen des pré-tests d'une valeur de d=0.35 [0.07, 0.64], et un bénéfice important des post-tests de d =0.74 [0.44, 1.04].

L'analyse des scores obtenus à l'examen final sur de nouvelles questions qui n'étaient pas présentes dans les tests d'apprentissage montre des résultats légèrement différents. Alors que le bénéfice des post-tests perdure (d=0.32 [0.07, 0.66]), celui des pré-tests n'est plus présent (d= -0.01 [-0.29, 0.27]). Autrement dit, l'avantage d'apprendre via des tests placés après la lecture va au-delà des informations testées, à l'inverse de l'apprentissage par les tests placés avant la lecture des modules du cours. Ce bénéfice indirect des post-tests a déjà été retrouvé dans plusieurs études pour certaines conditions d'apprentissage. Ce résultat est également cohérent avec une récente méta-analyse montrant un transfert de l'effet positif de l'apprentissage par récupération en mémoire sur les questions non entraînées (Pan & Rickard, 2018). Un certain nombre d'arguments ont été avancés dans la discussion de l'article présentant cette expérience pour expliquer l'absence de bénéfice indirect de la condition pré-test sur la réussite aux questions nouvelles (Latimier et al., 2019).

Les conditions dans lesquelles sont retrouvés les bénéfices des pré-tests et post-tests d'apprentissage ne sont donc pas les mêmes, et ceux-ci dépendent du type de questions présentées lors de l'examen (déjà vues durant l'apprentissage versus inconnues).

Pour mieux mémoriser, l'apprentissage par des tests doit-il se faire avant ou après la lecture des contenus pédagogiques ?

Les analyses sur l'effet de la position des tests d'apprentissage pré- versus post-test ont montré une nette supériorité des tests placés après la lecture des modules par rapport aux tests placés avant. Cet effet significatif se retrouve aussi bien sur les questions déjà vues durant l'apprentissage (d = 0.34 [0.05, 0.63]) que sur les questions inconnues découvertes lors de l'examen (d = 0.32 [0.03, 0.61]). Ce résultat est cohérent avec ceux décrits ci-dessus qui montrent un effet de post-test plus important que l'effet de pré-test. Pour répondre à cette deuxième question posée dans l'Expérience 1 : **il semble que les tests réalisés dans un second temps après la lecture des ressources pédagogiques est une stratégie plus adaptée pour consolider les nouvelles connaissances et optimiser leur rétention à long terme.**

6.1.3 Résultats principaux du Chapitre 4 (Expérience 2)

Y a-t-il une interaction entre le niveau granularité des contenus (fin, moyen, grossier) et la stratégie d'apprentissage utilisée pour les apprendre (récupération en mémoire sous forme de tests vs relectures) ?

Dans le cadre de l'apprentissage d'un contenu pédagogique dense ayant un certain niveau de complexité, il paraît important de savoir quelle est la meilleure manière de le présenter aux apprenants. Dans cette seconde expérience, il était toujours question d'étudier l'emplacement optimal des épisodes d'apprentissage par les tests par rapport au cours à lire. Cette fois-ci, nous nous sommes intéressés au niveau de granularité des sessions d'acquisition des connaissances : autrement dit, nous avons manipulé le niveau de découpage du contenu à apprendre.

Une nouvelle fois nous avons retrouvé un effet positif global de l'apprentissage par les tests sur la rétention à long terme par rapport à la relecture (effet de test de d = 0.37 [0.2, 0.49]). En revanche, le niveau de granularité des contenus n'avait globalement pas d'effet significatif sur la rétention, c'est à dire que les trois niveaux de découpage du contenu d'apprentissage ont mené à des performances finales équivalentes.

Ensuite, et c'est la question principale posée par cette expérience, nous avons effectivement trouvé une interaction significative entre le niveau de granularité et le type de stratégie d'apprentissage utilisée. Plus précisément : **l'effet du découpage du contenu sur la mémorisation des connaissances n'était pas le même dans la condition d'apprentissage avec tests et dans la condition d'apprentissage avec relectures**. Nous avons comparé la valeur de l'effet de test pour chaque niveau de granularité pour comprendre cet effet d'interaction. Alors que le bénéfice de l'apprentissage par des tests est équivalent pour les découpages grossier (d=0.53) et moyen (d=0.56), il était inexistant dans le cas d'un découpage fin du cours (d=0.063). Cette modulation de l'effet de l'apprentissage par les tests sur la rétention est principalement due à l'absence d'effet de la granularité dans les groupes ayant appris avec des tests par rapport aux groupes ayant appris avec la relecture. Ces résultats sont cohérents avec ceux de la littérature (Duchastel & Nungester, 1984; Uner & Roediger, 2017; Y. Weinstein, Nunes, & Karpicke, 2016; Wissman & Rawson, 2015) : **le découpage fin des contenus lus en alternance avec des petits tests n'apporte pas de bénéfice plus important que la lecture du cours dans son intégralité suivi d'un long test d'apprentissage.**

En revanche, en apprenant avec la relecture successive, un découpage fin par module semble permettre une meilleure mobilisation de l'attention des apprenants et ainsi une meilleure consolidation des connaissances. En effet, **dans les conditions de relecture des contenus,**

les performances finales des participants étaient d'autant plus importantes que la granularité des contenus était fine.

Dans quelles conditions cette interaction a t-elle lieu ?

L'interaction significative entre le niveau de granularité et le type de stratégie d'apprentissage utilisée décrite ci-dessus a été mesurée en analysant les scores finaux dans leur globalité. C'est-à-dire qu'aucune distinction n'a été faite entre l'examen fait une semaine après et l'examen fait un mois après l'apprentissage ; ni entre les deux types de questions incluses dans ces examens. Cette interaction était exactement la même aux deux examens, elle perdure donc sur le long terme.

Même si nous n'avons pas trouvé de triple interaction entre niveau de granularité, type de stratégie d'apprentissage et type de questions présentées aux examens, il semblait intéressant de faire des analyses restreintes sur les scores obtenus aux questions nouvelles. L'allure de l'interaction entre le niveau de granularité et le type de stratégie d'apprentissage est légèrement différente. Le niveau de granularité semble toujours importer dans les conditions d'apprentissage avec relecture par rapport à celles avec des tests (on observe une augmentation des performances plus le niveau de granularité est fin). Alors que l'effet de test reste positif dans le niveau de granularité grossier, il est devenu négatif dans le niveau de granularité fin. Cela confirme que l'utilisation d'un découpage très fin des contenus à lire en alternance avec la réalisation de test n'est pas systématiquement bénéfique. En effet, pour ce niveau de granularité, il n'y avait de bénéfice de l'apprentissage par les tests ni sur les questions déjà vues lors de l'apprentissage ni sur les questions nouvelles de transfert de connaissances.

6.1.4 Résultats principaux du Chapitre 5 (Expérience 3)

De quelle ampleur est le bénéfice l'apprentissage par des tests massées dans le temps et des relectures espacées sur la rétention de nouvelles connaissances ?

Dans cette dernière expérience nous avons souhaité créer des situations de révision de nouvelles connaissances réalistes tout en utilisant un vrai contenu éducatif. Pour cela, nous avons comparé quatre conditions d'apprentissage très étudiées dans la littérature dans le même design expérimental : 1) l'apprentissage avec relectures selon un planning massé vs 2) espacé dans le temps vs 3) l'apprentissage par les tests selon un planning massé vs 4) espacé. Autrement dit, nous avons croisé la méthode d'apprentissage utilisée avec le type de planning d'apprentissage. Le contenu de l'Expérience 1 a été réutilisé et adapté pour des participants anglophones cette

fois-ci. Après une phase de lecture du cours dans son intégralité, les participants devaient s'engager dans une période de révision dépendante de leur condition expérimentale dans le but de réaliser deux examens : un jour puis une semaine après la fin des révisions.

L'analyse des performances finales a une nouvelle fois montré le bénéfice de l'apprentissage par les tests qu'importe le type de planning des révisions (effet de test de d=0.66 [0.49, 0.80]). En revanche, et contrairement à nos attentes, nous n'avons pas trouvé de bénéfice global des plannings espacés des révisions par rapport aux plannings massés (d= 0.15 [-0.006, 0.30]). Plusieurs explications ont été avancées pour cette absence d'effet d'espacement dans la discussion du Chapitre 5 (e.g, espacement de plusieurs jours entre lecture du cours et séances de révision dans les conditions avec planning massé).

Malgré l'absence d'interaction entre type d'apprentissage et type de planning, nous avons comparé les quatre conditions d'apprentissage entre elles pour mesurer la valeur de leur bénéfice respectif par rapport à la condition contrôle (révision par relectures massées dans le temps la veille de l'examen). **Les résultats indiquent que :**

1) Les participants en condition de tests massés ont obtenu un score de mémorisation supérieur à celui des participants en condition de relectures massées (d=0.60 [0.38, 0.81];

2) Réviser sous forme de relectures massées ou espacées mène à des performances équivalentes (d=0.13 [-0.10, 0.36]) ;

3) La comparaison des deux stratégies opposées, celle qui n'intègre pas de difficultés désirables versus celle qui combine effets de récupération en mémoire et d'espacement, montre un bénéfice important de l'apprentissage par les tests espacés dans le temps (d=0.81 [0.59, 1.03]).

Quel est l'effet d'un apprentissage combinant l'utilisation de tests et l'espacement des sessions de révisions dans le temps ?

L'un des objectifs principaux de cette expérience était de mesurer l'effet de la combinaison des deux stratégies par rapport à l'apprentissage n'intégrant que des tests de récupération en mémoire ou bien que l'espacement des révisions. Etant donnée l'interaction non significative entre le type d'apprentissage et le type de planning des révisions, les valeurs des tailles d'effets sont à prendre avec précaution. Néanmoins, elles suggèrent que la différence de performances finales entre cette combinaison de stratégies et l'apprentissage par les tests massés dans le temps est négligeable (d=0.13 [-0.08, 0.33]).

Pour résumer, les résultats de cette expérience suggèrent que l'effet de l'apprentissage par les tests serait globalement plus « puissant » que l'effet d'espacement. Les analyses descriptives tendent à montrer que les tests d'apprentissage espacés dans le temps pourrait avantage sur les tests massés en permettant d'atteindre un certain niveau de mémorisation tout en minimisant l'oubli, mais cela reste à confirmer.

Dans quelle mesure la réponse à ces questions diffère-t-elle entre la rétention à court-terme (24h après l'apprentissage) et la rétention à long-terme (une semaine après l'apprentissage) ?

Nous n'avons pas trouvé de différence significative entre les scores mesurés à l'examen 1 (un jour de rétention) et ceux à l'examen 2 (une semaine de rétention). Aussi, nous n'avons pas trouvé que les effets de test et d'espacement, ni même la combinaison des deux stratégies d'apprentissage, variaient selon l'intervalle de rétention. Autrement dit, tous les effets décrits plus haut restent les mêmes à court et à long terme.

Quelle est l'influence de ces différentes strategies d'apprentissage sur la capacité des apprenants à prédire leurs performances aux examens finaux ?

Les analyses réalisées sur les jugements d'apprentissage (*Judgements of Learning* en anglais) révèlent que globalement les apprenants avaient tendance à surestimer leurs futures performances avant chaque examen (jugements prospectifs). En revanche, cette surestimation disparaissait lorsque la prédiction se faisait après chaque examen (jugements rétrospectifs). Ces résultats montrent l'intérêt de réaliser des tests pour réajuster la perception de nos propres capacités à se rappeler des connaissances tout juste apprises. Enfin, les analyses portant sur l'effet des différentes conditions d'apprentissage sur la justesse de ces prédictions montrent que les participants ayant appris avec la relecture ont tendance à avoir un excès de confiance plus important que ceux ayant appris avec des tests. Cependant, ces résultats ne sont pas significatifs. Des analyses supplémentaires sont nécessaires pour prendre en compte plus de prédicteurs dans un même modèle de régression à savoir les stratégies d'apprentissage, l'emplacement des jugements d'apprentissage (prospectif vs rétrospectif) et l'intervalle de rétention (examen 1 vs examen 2).

6.2 Synthèse générale

- Réponse à la problématique de la thèse

Au cours des dernières décennies, le bénéfice de l'apprentissage par récupération en mémoire (ou effet de test) a sans doute été l'effet le plus étudié par la communauté de chercheurs dans le domaine de la psychologie de l'apprentissage (Benjamin & Pashler, 2015; Dunlosky & Rawson, 2019; Roediger, 2013; Roediger & McDermott, 2018). Etant donné la robustesse de son effet comme outil de révision sur la rétention à long terme par rapport à la relecture, les recommandations pour son intégration dans les pratiques pédagogiques sont justifiées (Putnam et al., 2016; Suzuki, Nakata, & Dekeyser, 2019; Yana Weinstein et al., 2018). Désormais, la question à laquelle la recherche doit répondre est comment augmenter, moduler l'efficacité de cette stratégie d'apprentissage ? Un certain nombre d'incertitudes subsistent encore sur l'optimisation de l'effet de test c'est la raison pour laquelle ce travail avait pour but de mesurer les paramètres qui pouvaient amplifier son bénéfice. La présente recherche a permis de répondre en partie à cette interrogation ; la question générale qui a guidé ce travail était la suivante : **comment optimiser le bénéfice des tests d'apprentissage pour une meilleure rétention à long terme en manipulant leur emplacement ?** Par le biais de trois expériences et une synthèse de la littérature, nous nous sommes intéressés au positionnement optimal des épisodes de récupération en mémoire (Chapitre 3), au bon niveau de granularité de ces épisodes (Chapitre 4), et à la manière de combiner cette stratégie à celle de l'espacement des révisions dans le temps (Chapitres 2 et 5).

Pour mener cette recherche, une méthodologie particulière a été employée dans une perspective d'application concrète des résultats. Plus précisément nous avons utilisé : 1) un terrain d'expérimentation réaliste d'apprentissage en ligne sur une vraie plateforme de formation (mais très contrôlé comme en laboratoire pour mesurer exclusivement l'effet de certains paramètres) ; 2) des contenus d'apprentissage construits à partir de connaissances issues de programmes scolaires (Expériences 1 et 3) et de formation professionnelle (Expérience 2) ; 3) des échantillons importants pour assurer une certaine fiabilité des résultats et 4) des populations de participants aux profils divers. Enfin, le travail de synthèse de la littérature a permis de fournir des mesures quantitatives solides pour étayer certaines hypothèses sur l'optimisation de l'effet de test. En lien avec la démarche d'éducation fondée sur des preuves présentée en *Introduction générale*, l'ensemble de ce travail a permis de produire de nouvelles connaissances fondamentales et appliquées (voir le point 1.1.2. du

Chapitre 1, p3). Nous avons pour cela mis en place des expérimentations contrôlées et randomisées, et nous avons systématiquement calculé des tailles d'effets pour les comparaisons des différentes stratégies d'apprentissage. De plus, les trois expériences ont fait l'objet d'un pré-enregistrement des hypothèses et du plan expérimental afin de cadrer au préalable les questions précises auxquelles nous souhaitions répondre.

Dans le cadre de ce projet, nous avons donc produit de nouveaux résultats sur l'optimisation de l'effet de récupération en mémoire. Nous avons montré qu'il est possible d'amplifier l'effet de cette stratégie sous certaines conditions pour favoriser la rétention à long terme de nouvelles connaissances. Grâce aux résultats de l'Expérience 3 et de la méta-analyse, nous avons pu mesurer de quel ordre était le bénéfice de l'apprentissage par les tests espacés dans le temps par rapport à celui généré par d'autres stratégies d'apprentissage. Il y aurait un bénéfice de l'apprentissage par les tests espacés dans le temps par rapport à des relectures massées ou espacées dans le temps. Le bénéfice des tests espacés par rapport aux tests massées dans le temps durant la révision du cours est quant à lui plus ambigu et ne semble pas systématiquement supérieur. Il semblerait que l'emplacement des révisions par rapport à la lecture du cours dans la condition de planning massé est déterminant pour obtenir un bénéfice de l'espacement (voir la section *Discussion* du Chapitre 5). Enfin, d'après la méta-analyse, le planning expansif d'espacement des tests ne semble pas être supérieur au planning uniforme.

Nous avons également déterminé l'emplacement le plus approprié des tests d'apprentissage pour une meilleure mémorisation ; c'est-à-dire placés après la lecture des contenus pédagogiques plutôt qu'avant (Expérience 1). Cependant, segmenter, granulariser le cours en différents modules pour qu'ils soient lus en alternance avec de petits tests d'apprentissage n'est pas forcément le meilleur agencement pour tirer profit de l'apprentissage par les tests (Expérience 2). A l'inverse, ce découpage semble être plus approprié dans le cas de l'apprentissage avec relectures. Pour résumer, l'ensemble de ce travail apporte une meilleure compréhension des paramètres influençant l'effet de test dans une approche fondée sur des preuves.

- Limites

Tout d'abord, plusieurs facteurs peuvent remettre en doute la validité écologique des différentes expériences. En effet, nos résultats ont été obtenus dans un contexte d'apprentissage particulier. A travers l'utilisation de Didask, nous avons essayé de simuler des situations d'apprentissage réalistes tout en bénéficiant de l'avantage du recrutement et des passations des

participants en ligne. Cependant, en terme de processus de régulation des apprentissages, apprendre de nouvelles connaissances sur une plateforme en ligne comme Didask n'est pas équivalent à un apprentissage via des manuels scolaires ou en présentiel avec un enseignant/formateur (Shea & Bidjerano, 2010; Tsai, Shen, & Fan, 2013). Alors que la régulation de l'apprentissage est à la charge de l'apprenant en distanciel, elle se fait via l'enseignant/formateur dans le cadre du présentiel. Dans notre cas, les apprenants étaient en totale autonomie mais devaient suivre les instructions données dans le cadre des expériences. Même si les participants pouvaient à tous moments interagir avec l'expérimentateur, leurs questions ne pouvaient concerner que le déroulement de l'expérience et non pas le contenu à apprendre. Or, dans la réalité il est possible d'interagir avec d'autres personnes pendant un apprentissage (que ce soit avec les pairs ou bien les formateurs), ce qui a une influence non négligeable sur la réussite de l'apprentissage. Dans notre cas, la plateforme Didask était l'unique source d'apprentissage et les participants avaient finalement assez peu d'informations sur leurs performances en temps réel, ce qui n'est pas réaliste par rapport à une situation d'apprentissage en présentiel.

Il est également important de rappeler que nous avons utilisé des contenus d'apprentissage bien spécifiques et que nous ne pouvons être certains de la généralisation des effets mesurés à des contenus différents. En dehors des considérations « pratiques » pour choisir ces contenus plutôt que d'autres (principalement la facilité de préparation car les contenus étaient déjà disponibles via Didask), il paraissait pertinent de présenter des connaissances scientifiques en lien avec des acquis fondamentaux du baccalauréat car ce sont sur ces connaissances-là que les élèves ont le plus de difficultés. Ce type de contenu se prête aussi bien à des tests sur des connaissances factuelles (ex : « *Que veux dire ADN ?* » ou « *Qu'est-ce qu'une protéine* ?») qu'à des tests nécessitant de faire des inférences et des liens entre les différents mécanismes biologiques et moléculaires qui étaient décrits dans les modules du cours. Pour le contenu de formation professionnelle que nous avons adapté pour l'Expérience 2, des connaissances spécifiques étaient abordées en lien avec des problématiques rencontrées dans le monde du travail, et le but était de pouvoir faire de la mise en pratique des connaissances abordées dans le cours via des situations fictives mais réalistes. Néanmoins, il faudrait s'assurer que nos effets se généralisent à d'autres contenus et à d'autres populations. Nous avons en effet privilégié les adultes mais il serait très intéressant de reproduire ce type d'expérimentation chez les enfants et adolescents, qui sont rarement les populations d'intérêt dans l'étude des effets de test et d'espacement (Goossens et al., 2016; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2012, 2014).

Une autre différence notable entre nos situations expérimentales d'apprentissage et la réalité est l'origine de la motivation pour apprendre. En classe, les étudiants ont une source de motivation extrinsèque forte via l'obtention de notes (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Phelps, 2012) et ont un certain niveau de motivation intrinsèque qui dépend de l'intérêt pour une discipline et qui est lié à la croyance qu'apprendre est important pour réussir (Delaney et al., 2010). Dans le cadre d'expérimentations sur l'apprentissage en laboratoire, la source de motivation pour apprendre est principalement extrinsèque puisque les participants perçoivent dans la majorité des cas une compensation financière ou des crédits bonus pour un cours dans le cas de participants étudiants. Cette compensation ne varie pas en fonction de la performance finale, et ne prend pas en compte l'effort et l'engagement dans la tâche, ce qui peut rend la motivation à apprendre plus « artificielle ». Par ailleurs, il est probable que cette compensation financière induise un biais dans le recrutement des participants.

Enfin, un dernier paramètre lié à l'apprentissage en ligne pouvant remettre en cause la validité de nos résultats est l'absence de contrôle précis sur la réalisation de la tâche par les participants. Contrairement à une situation d'apprentissage beaucoup plus contrôlée en laboratoire où l'expérimentateur peut surveiller le bon déroulement de l'apprentissage et le respect des consignes, ici nous n'avions aucun moyen de savoir si les participants étaient vraiment en train de réaliser notre tâche sans faire autre chose en parallèle. En intégrant des questions d'attention dans les quiz et en récupérant des informations temporelles sur les temps passés sur les contenus, nous avions malgré tout quelques variables pour appliquer des critères d'exclusion et s'assurer de la qualité des données comportementales enregistrées durant les différentes étapes des expérimentations.

Ces limites liées aux conditions d'expérimentation montrent que trouver un équilibre entre réalisme et rigueur expérimentale n'est pas chose facile. Malgré tout, les participants de nos expériences étaient sur un environnement d'apprentissage réel (Didask existe vraiment !) et nous avons essayé de préserver au maximum les fonctionnalités de la plateforme tout en réduisant le risque d'avoir des variables non contrôlables. Par exemple, nombre d'expériences sur l'effet de récupération en mémoire contraignent temporellement les participants pour s'assurer que le temps passé à apprendre est équivalent d'une condition à l'autre. Nous avons fait le choix de n'imposer aucune limite de temps aux participants, toujours dans un souci de réalisme. Cette variable pouvant avoir une influence non négligeable sur les effets mesurés, nous l'avons systématiquement intégrée dans nos modèles statistiques comme covariable au

moment des analyses (et nous avons fait de même avec les données sociodémographiques). D'autres variables non contrôlables par l'expérimentateur et dépendantes de chaque participant auraient également pu être intégrées dans les analyses, telles que le moment de la journée durant lequel les sessions d'apprentissage et d'examen ont été réalisées par les participants (on peut imaginer que les données comportementales mesurées peuvent être influencées par le niveau de vigilance).

Concernant le travail de méta-analyse (Chapitre 2), nous avons dû faire face à des limites de puissance statistique. En raison du faible nombre de tailles d'effet dans le deuxième jeu de données (*subset 2*), les analyses ne permettent pas de conclure sur le bénéfice réel et significatif sur la rétention en mémoire des tests d'apprentissage espacés par rapport aux relectures espacées. Les mots clés choisis pour extraire des références sur les différentes bases de données n'ont pas forcément permis de cibler toutes les études nécessaires pour mesurer la taille d'effet globale de cette comparaison. De plus, comme nous n'avions pas réalisé de pré-enregistrement de notre méthodologie et notre plan d'analyse précis, nous avons omis une comparaison intéressante pour avoir un aperçu complet du bénéfice des tests d'apprentissage espacés dans le temps. En effet, pour savoir s'il existe une additivité des effets de récupération en mémoire et d'espacement, il aurait fallu rajouter les études comparant l'apprentissage par les tests espacés à l'apprentissage par relecture massée.

6.3 Implications pratiques et perspectives

- Quelles recommandations pratiques pour Didask et pour la communauté enseignante ?

Un des objectifs de ce projet de recherche était l'apport de recommandations pratiques sur l'implémentation de l'apprentissage par les tests sur la plateforme Didask. Au début de la collaboration avec l'entreprise, l'équipe se posait de nombreuses questions quant à la manière de présenter les tests d'apprentissage par rapport à l'accès aux ressources pédagogiques. De plus, l'approche de Didask était de découper les contenus très finement en unités d'apprentissage élémentaires, dans lesquelles était abordée une connaissance clé qui était testée et pratiquée par un court test de 5 à 10 questions. Ainsi, les trois expériences que nous avons réalisées sur la plateforme étaient non seulement en lien avec des problématiques encore peu explorées par la littérature, mais aussi en lien avec des questions de développement de l'interface posées par Didask. Nous avons apporté des pistes d'optimisation des tests d'apprentissage dans les résumés des résultats principaux de chaque expérience. Depuis

l'Expérience 1, l'apprentissage sur Didask a évolué, et les apprenants sont désormais invités à consulter la ressource pédagogique dans un premier temps puis sont orientés vers le test pour s'exercer dans un second temps. Les résultats de l'Expérience 2 invitent à repenser la manière de granulariser les contenus en fonction de leur densité et de leur complexité. A ce jour, Didask incite fortement les apprenants à laisser passer 24h entre deux tentatives pour réaliser un même test d'apprentissage, mais il n'y a pas d'algorithme d'espacement intégré dans les fonctionnalités de la plateforme. L'Expérience 3 montre qu'il est finalement aussi efficace d'insérer un espacement d'un jour entre deux sessions de pratique par les tests que de ne pas en insérer, même si les résultats sont à interpréter avec précaution.

En termes d'implications pratiques pour l'enseignement, les résultats présentés dans cette thèse sont également pertinents. Pour autant, **nous recommandons très fortement l'utilisation de tests comme véritables outils pour apprendre et consolider les nouvelles connaissances** (Agarwal, Bain, & Chamberlain, 2012). Pour intégrer dans une routine l'utilisation des tests d'entraînement dans les milieux éducatifs, le recours aux outils numériques paraît une solution prometteuse. Par exemple, des dispositifs simples tels que des *clickers* (boitiers permettant de répondre instantanément à des questions) permettent de tester les connaissances sur tout un groupe d'élèves (Bjork, Soderstrom, & Little, 2015; Hunsu, Adesope, & Bayly, 2015). Ce type d'outil à faible coût a l'avantage de donner un aperçu en temps réel du bon niveau d'acquisition des connaissances pour fournir aux élèves des retours correctifs immédiats et ciblés en fonction des difficultés rencontrées. Il semblerait également que l'utilisation de ce type d'outils dans le cadre de l'évaluation formative favorise l'engagement et la motivation des étudiants (e.g., Healy, Jones, Lalchandani, & Tack, 2017). De nombreuses plateformes permettent aux enseignants de programmer et de partager des quiz d'entrainement (à titre d'exemples : Quizlet, Polleverywhere, Socrative, Google forms, Kahoot!). Ces tests peuvent être présentés avant la présentation d'un nouveau chapitre du cours pour évaluer le niveau de connaissances préalables et préparer les élèves à l'introduction de nouveaux concepts clés. Dans un second temps, ces mêmes tests peuvent être présentés au moment de la conclusion du chapitre pour 1) consolider les connaissances abordées depuis le début, et 2) identifier les notions qui posent encore des difficultés et revenir sur le point du cours qui les aborde.

Combiné à l'apprentissage par les tests, nous recommandons également l'espacement des révisions dans le temps étant donné le bénéfice de cette stratégie pour ralentir l'oubli jusqu'au moment d'un examen. A titre d'exemples, les programmes

anglophones Anki[27] et Mnemosyne[28] permettent de créer des cartes du type « question-réponse » en intégrant un planning de répétition des connaissances espacées dans le temps.
Les performances obtenues aux tests d'apprentissage ne doivent pas nécessairement constituer une note, mais la régularité avec laquelle les élèves s'en servent et leur taux de progression pourraient servir de points bonus au moment de l'évaluation sommative. Cela permettrait de récompenser les élèves n'ayant pas les meilleures notes aux examens mais qui travaillent régulièrement via ces tests d'entrainement. Indirectement, utiliser l'apprentissage par les tests de manière optimale en classe et en dehors de la classe permettrait aux élèves de mieux réguler et planifier leurs révisions (Littrell- Baez, Friend, Caccamise, & Okochi, 2015). En leur donnant des retours riches et centrés sur la tâche lors de la réalisation des tests d'apprentissage, les élèves seraient plus à même d'identifier les connaissances qui ne sont pas totalement acquises et celles qui le sont, ce qui favoriserait une meilleure régulation des apprentissages tout en évitant le risque d'illusion de maitrise (R. A. Bjork et al., 2013; Fernandez & Jamet, 2017; Littrell, 2011). Une recommandation d'ordre plus général pour promouvoir l'utilisation de tests d'entrainement et de l'espacement serait donc d'informer explicitement les étudiants sur les bénéfices de ces méthodes sur la rétention à long terme et sur la métacognition (Einstein, Mullet, & Harrison, 2012; Jonsson, Hedner, & Olsson, 2012; Özsoy & Ataman, 2017).

- Comment aller plus loin ?

Ces travaux de recherche ouvrent sur un certain nombre de perspectives, toujours dans l'objectif de pouvoir produire des connaissances fondamentales et appliquées.

Tout d'abord, en lien avec les limites qui découlent de nos résultats, il faudrait s'assurer de la leur reproductibilité. Répliquer les effets mesurés dans nos expériences permettrait de renforcer leur robustesse. Dans le même ordre d'idée, évaluer leur généralisation à d'autres contenus, d'autres populations, et dans d'autres conditions d'apprentissage apporterait une meilleure compréhension des mécanismes qui influencent ces effets. Aussi, pour assurer une meilleure validité écologique, nous pourrions reproduire nos expériences dans un cadre plus réaliste avec des apprenants qui doivent réellement suivre une formation. Cela permettrait d'impliquer des enseignants ou des formateurs dans la création des contenus et dans la mise en place des sessions d'apprentissage. L'analyse de la réplicabilité de résultats du laboratoire vers la classe est finalement récente et les expérimentations qui se déroulent en classe sur le bénéfice

[27] https://apps.ankiweb.net/

[28] https://mnemosyne-proj.org/

de l'apprentissage par les tests restent encore à la marge (Karpicke & Aue, 2015). Dans leur méta-analyse, Adesope et collaborateurs (2017) indiquent que seulement 11% des études sur le l'effet de test qui sont incluses ont lieu en classe. Malgré les résultats très prometteurs produits en laboratoire sur le bénéfice de la récupération en mémoire pour apprendre, il manque encore des preuves sur les conditions d'application de l'effet de test en classe et sur son efficacité à différents niveaux scolaires (Chew et al., 2018; Greving & Richter, 2019). Pourtant, il est possible de retrouver le bénéfice de cet effet dans un environnement moins contrôlé. Même si le calcul de la taille d'effet de test en classe concerne un petit effectif de 30 études dans la synthèse d'Adesope et collaborateurs, leurs résultats sont encourageants puisqu'ils trouvent un bénéfice similaire pour les études menaient en classe par rapport à celles réalisaient en laboratoire ($g = 0.67$ versus $g = 0.62$). De façon similaire, Schwieren, Barenberg, & Dutke, (2017) trouve un bénéfice de $d = 0.56$ dans leur méta-analyse s'intéressant à l'apprentissage par les tests chez les étudiants en psychologie (sur 19 études). Enfin, il serait intéressant dans le contexte d'expérimentation en classe de comparer les bénéfices de l'apprentissage par récupération en mémoire à d'autres stratégies d'apprentissage dite « actives » telle que l'élaboration ou l'utilisation de carte mentale (Moreira, Pinto, Starling, & Jaeger, 2019).

Deuxièmement, la collaboration avec Didask a permis de monter qu'il est très intéressant d'utiliser des outils numériques comme terrain d'investigation des processus d'apprentissage et de leur optimisation. Ce type de plateforme permet de suivre les parcours d'apprentissage d'un nombre considérable d'apprenants. Actuellement, Didask suit environ 25000 apprenants engagés dans différents parcours de formations dont les thèmes sont très variés. Cela offre des perspectives de projet de recherche et développement très intéressantes. De plus, ce type de plateforme permet de générer un nombre considérable de variables comportementales, tant sur les comportements d'apprentissage (réponses aux tests et temps de réponses) mais aussi sur les interactions et la navigation sur la plateforme. Pour répondre aux questions des différentes expériences nous nous sommes concentrés sur l'analyse d'un nombre restreint de variables, mais nous aurions pu explorer d'avantage les autres mesures enregistrées pour affiner nos interprétations (notamment des données temporelles : temps passé sur les retours correctifs et sur les questions des tests d'apprentissages, nombre de tentatives de réponses aux questions avant de valider…). Pour que les chercheurs puissent profiter d'avantage de ce grand choix de contenus d'apprentissage et de cette grande diversité de profils apprenants il serait important de réfléchir bien en amont à la manière de mieux contrôler la

réalisation d'expérience sur ce type d'outil d'apprentissage tout en préservant une certaine validité écologique.

Dans la continuité de ces recherches et en lien avec la recherche sur l'apprentissage personnalisé, il serait pertinent de prendre en compte certaines caractéristiques individuelles des apprenants pour comprendre leur importance dans la modulation des effets de récupération en mémoire et d'espacement. Une littérature de plus en plus riche s'intéresse aux différences individuelles qui pourraient expliquer des différences de bénéfice de l'apprentissage par les tests d'un apprenant à l'autre (Agarwal, Finley, Rose, & Roediger, 2017; Brewer & Unsworth, 2012; Callender & McDaniel, 2007). Dans notre cas, nous avons utilisé très peu de données individuelles en dehors de l'âge, du niveau d'étude, et du degré de connaissances préalables sur les thématiques des cours. De nombreuses caractéristiques individuelles peuvent être prises en compte ainsi que des variables liées à la réussite scolaire en général (Credé & Kuncel, 2008; Hattie, 2015; Komarraju & Nadler, 2013; Robbins et al., 2004). Les habilités cognitives telles que la capacité de compréhension, de mémoire de travail, le niveau intelligence fluide peuvent expliquer une part de la variance dans les performances finales à l'issue d'un apprentissage par les tests (Argawal, 2017). Enfin, les compétences métacognitives constituent un paramètre intéressant pour prédire la réussite (Kizilcec, Pérez-Sanagustín, & Maldonado, 2017); elles peuvent être mesurées via des auto-questionnaire (ex., Schraw & Dennison, 1994) mais aussi à travers une évaluation des habitudes de travail (ex.,: stratégies mises en place pendant les révisions, allocation du temps pour apprendre, Robey, 2017).

- Pour finir

Dans une démarche de production de données empiriques, l'ensemble de ce projet répondait à des questions sur l'optimisation de l'effet de récupération en mémoire, un des effets le plus étudié dans le domaine de la psychologie de l'apprentissage. A travers une série d'expérimentations et un travail de revue systématique de la littérature, nous avons mis en lumière des paramètres pouvant moduler le bénéfice de l'apprentissage par les tests sur la rétention de nouvelles connaissances. Au-delà de la problématique scientifique à laquelle nous avons cherché à répondre, des projets comme celui-ci montrent l'intérêt de travailler à l'interface entre recherche fondamentale et appliquée ; et en collaborant avec des acteurs économiques de l'éduction. Des projets de type « Recherche & Développement » avec des entreprises s'intéressant à l'innovation pédagogique et à la création d'outils fondés sur des preuves doivent continuer d'être valorisés.

Aussi, durant ce projet j'ai eu l'occasion de rencontrer de nombreux enseignants, mais aussi des inspecteurs, des formateurs de futurs enseignants, ainsi que des associations qui promeuvent la démarche expérimentale auprès des enseignants. Ces rencontres m'ont fait prendre conscience qu'en parallèle de la démarche expérimentale pour comprendre les paramètres de l'optimisation de l'apprentissage, il est également nécessaire d'instaurer un dialogue solide avec ces différents acteurs de l'éducation. D'une part, cela permettrait à la recherche d'impliquer les enseignants dans des projets de ce type pour répondre à des problématiques plus proches du terrain et en prenant en compte les contraintes de leur métier. D'autre part, cela permettrait aux chercheurs d'assurer un véritable travail de transmission et de dissémination des résultats de la recherche en psychologie.

Enfin, ce type de projet ouvre la voie vers des collaborations interdisciplinaires prometteuses. L'implication d'autres laboratoires qui travaillent sur des thématiques annexes dans le domaine de l'optimisation des apprentissages permet d'intégrer des méthodologies complémentaires. Des disciplines comme le *machine learning* appliqué à l'éducation qui s'intéresse au développement d'algorithmes de planification adaptative et personnalisée des séances d'apprentissage enrichissent notre compréhension des mécanismes de la mémorisation (e.g., Choffin, Popineau, Bourda, & Vie, 2019). De plus en plus d'études réunissent des chercheurs en sciences cognitives et en informatique pour tirer profit des apports de chaque discipline. Ce type de collaboration permettra à terme de développer des outils numériques intégrant des stratégies d'apprentissage efficaces tout en prenant en compte des caractéristiques individuelles des apprenants et des variables précises sur les contenus à apprendre (Mozer & Lindsey, 2016; Mozer, Wiseheart, & Novikoff, 2019).

7 Annexes

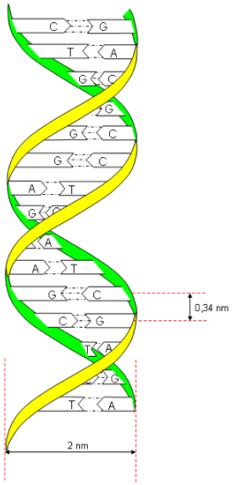
Le matériel expérimental utilisé pour chacune des trois expérimentations est présenté dans cette section. Les trois annexes incluent soit l'intégralité soit une partie des contenus pédagogiques que les participants devaient lire ; ainsi qu'un exemple de quiz ou de questions issues des quiz d'apprentissage. Dans le cas des Annexes 1 et 3, des exemples de visuels de la plateforme Didask sont fournis.

7.1 Annexe 1

Textes et illustrations des sept modules du cours sur l'ADN créé dans le cadre de l'expérimentation 1 présentée dans le Chapitre 3 (p67-74). Le premier quiz associé au module 1 est également présenté.

Module 1 : Généralités sur l'ADN

L'ADN est le support de l'information génétique. Il est contenu dans le noyau de nos cellules (unité biologique fondamentale de tous les êtres vivants connus). La molécule d'ADN est constituée d'un enchainement de nucléotides, et toutes les cellules d'un même organisme ont exactement la même information génétique. Cette molécule universelle dans le monde vivant, se transmet de générations en générations lors de la reproduction. L'ADN propre à notre espèce comprend entre 25 000 et 30 000 gènes qui contiennent l'information pour la production des protéines, composants principaux qui assure l'activité de nos cellules.



Quiz associé à ce module contenant six questions (les bonnes réponses apparaissent en vert et sont accompagnées d'une explication directement issue du texte du module) :

1. **Qu'est-ce que l'ADN ? (Une réponse possible)**

- La cellule d'un organisme contenant l'information génétique.
- La molécule qui crée des protéines à partir de l'information génétique.
- La molécule portant l'information génétique d'un organisme.
- La molécule protégeant l'organisme de mutations génétiques.

Explication : L'ADN est la molécule qui « porte en elle » l'information génétique qui est unique pour chaque individu. Sa structure contient une sorte de code qui peut être « lu » par certaines molécules.

2. **L'ADN d'un individu est transmis à … (Une réponse possible)**
- son entourage par contact physique
- ses descendants lors de la reproduction
- son conjoint lors de la reproduction
- ses parents lors de sa naissance

Explication : L'ADN d'un individu est constitué de celui de sa mère et de son père biologique. Un individu transmet son ADN à ses descendants, au moment de la reproduction sexuelle.

3. **L'ADN détermine la synthèse de … (Une réponse possible)**
- gènes
- cellules
- nucléotides
- protéines

Explication : L'ADN détermine la synthèse des protéines, qui sont les composants principaux assurant l'activité de nos cellules et plus globalement de notre organisme.

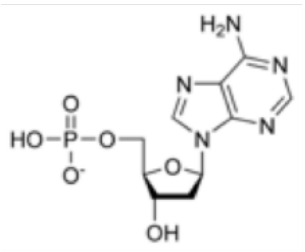
4. Dans quelles cellules d'un organisme trouve-t-on de son ADN **(Une réponse possible)**
- Dans toutes les cellules.
- Dans les neurones.
- Dans les cellules en division.
- Dans les cellules embryonnaires.

Explication : L'ADN se trouve dans toutes les cellules de l'organisme. Toutes les cellules d'un même organisme contiennent exactement la même information génétique.

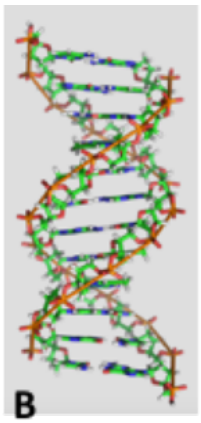
5. Dans une cellule humaine, on trouve principalement l'ADN dans... **(Une réponse possible)**
- sa membrane plasmique
- son noyau
- son cytoplasme
- les mitochondries

Explication : Notre ADN se trouve essentiellement dans le noyau de nos cellules, il s'agit donc d'une structure cellulaire importante.

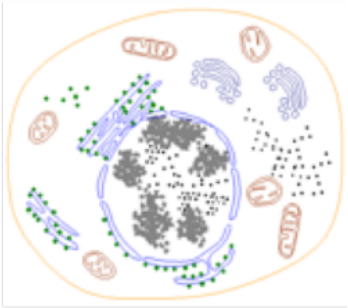
6. Lequel de ces trois schémas représente une molécule d'ADN ?



A



B



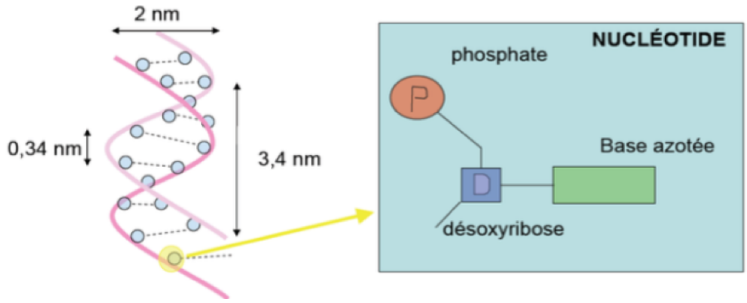
C

- A
- B
- C

Explication : La figure B est une molécule d'ADN, que l'on reconnaît à sa forme typique en double hélice. La figure A correspond à la formule chimique d'un nucléotide, unité élémentaire de l'ADN ; et la figure C est un schéma d'une cellule humaine.

Module 2 : Composition de l'ADN

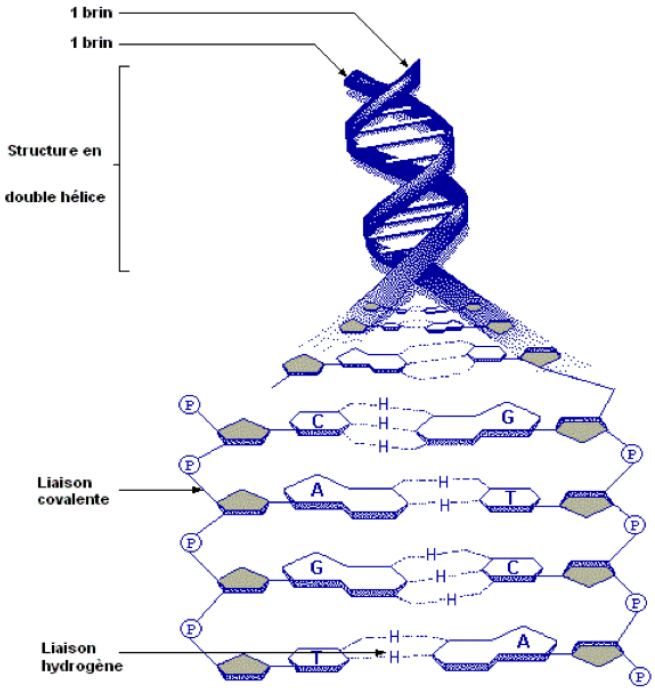
Le premier Acide DesoxyriboNucléique a été identifié et isolé en 1869 par le Suisse Friedrich Miescher. La molécule d'ADN est composée de deux chaînes, appelés aussi brins. Ces brins sont formés d'une succession de nucléotides, et ils s'enroulent l'un autour de l'autre en formant une double hélice. Chaque nucléotide est composé d'un sucre (le désoxyribose), d'un groupement phosphate et d'une base azotée parmi les 4 possibles : A (Adénine), T (Thymine), G (Guanine) et C (Cytosine). L'ADN est donc une succession de ces 4 nucléotides.



Module 3 : Structure de l'ADN

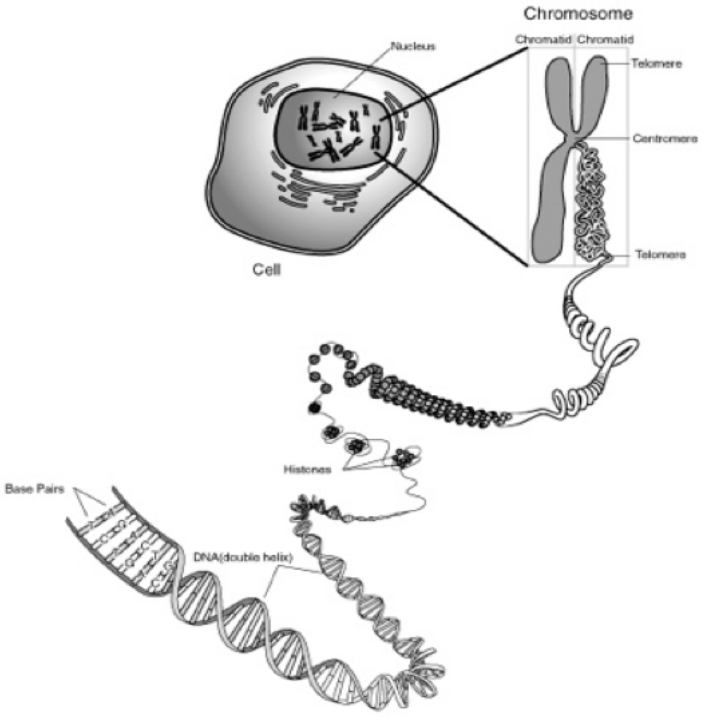
La structure en double hélice a été mise en évidence en 1953 par les chercheurs Watson et Crick, on reconnait également un rôle non négligeable de la biologiste Rosalind Franklin dans cette découverte. Ces scientifiques ont cherché à comprendre comment s'assemblent les différents éléments de l'ADN dans les trois dimensions. Ils montrèrent notamment que les

phosphates et les désoxyriboses forment les montants de la structure en "échelle" de l'ADN. Les "barreaux de l'échelle" sont formés par les paires de bases azotées de façon complémentaire : le nucléotide A s'apparie avec le nucléotide T, et C s'apparie avec G. Pour maintenir les deux brins entre eux, les nucléotides complémentaires sont reliés par des liaisons hydrogènes dites « faibles », c'est à dire qu'elles ne se défont que dans certaines circonstances (par opposition aux liaisons covalentes dites « fortes »). C'est notamment durant la transcription que les deux brins sont séparés momentanément.



Module 4 : L'ADN dans la cellule

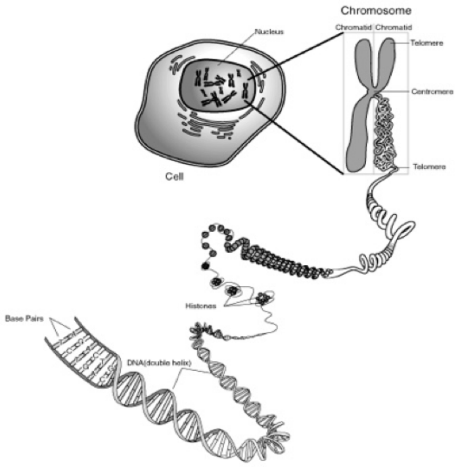
Après les travaux menés par Avery et collaborateurs dans les années 40 démontrant que l'ADN était le support de l'hérédité ; les chercheurs comprennent dans les années 60 que chaque chromosome est en fait la forme compactée d'une très longue molécule d'ADN. Selon le stade du cycle cellulaire, l'ADN peut avoir plusieurs formes possibles dans le noyau cellulaire. On parle du phénomène de condensation : sous forme décondensée, l'ADN est un long filament qui s'appelle la chromatine (très peu visible sous microscope); et sous forme condensée, l'ADN compacte s'appelle un chromosome (bien visible sous microscope). Qu'importe le degré de condensation, l'ADN est une molécule linéaire et possède deux brins complémentaires. En dehors du noyau, on retrouve une petite quantité d'ADN dans un organite appelé mitochondrie. Cet ADN est spécifique de la mitochondrie, et il est circulaire contrairement à l'ADN du noyau.



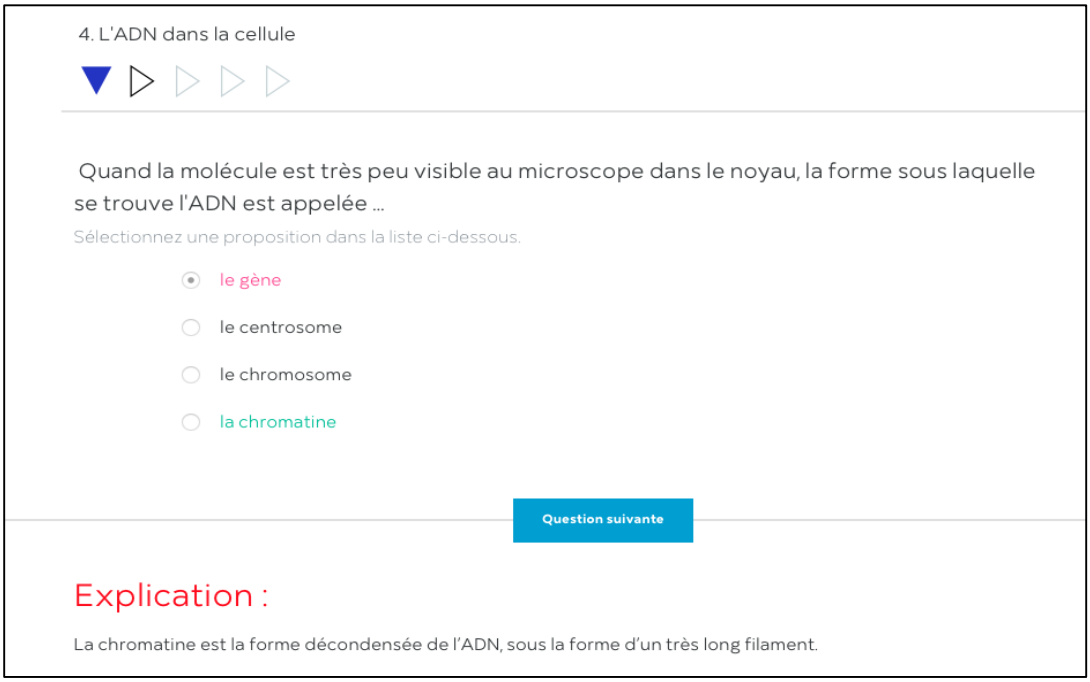
Visuel de ce module sur Didask :

Ressource pédagogique

Après les travaux menés par Avery et collaborateurs dans les années 40 démontrant que l'ADN était le support de l'hérédité; les chercheurs comprennent dans les années 60 que chaque chromosome est en fait la forme compactée d'une très longue molécule d'ADN. Selon le stade du cycle cellulaire, l'ADN peut avoir plusieurs formes possibles dans le noyau cellulaire. On parle du phénomène de condensation : sous forme décondensée, l'ADN est un long filament qui s'appelle la chromatine (très peu visible sous microscope); et sous forme condensée, l'ADN compacte s'appelle un chromosome (bien visible sous microscope). Qu'importe le degré de condensation, l'ADN est une molécule linéaire et possède deux brins complémentaires. En dehors du noyau, on retrouve une petite quantité d'ADN dans un organite appelé mitochondrie. Cet ADN est spécifique de la mitochondrie, et il est circulaire contrairement à l'ADN du noyau.

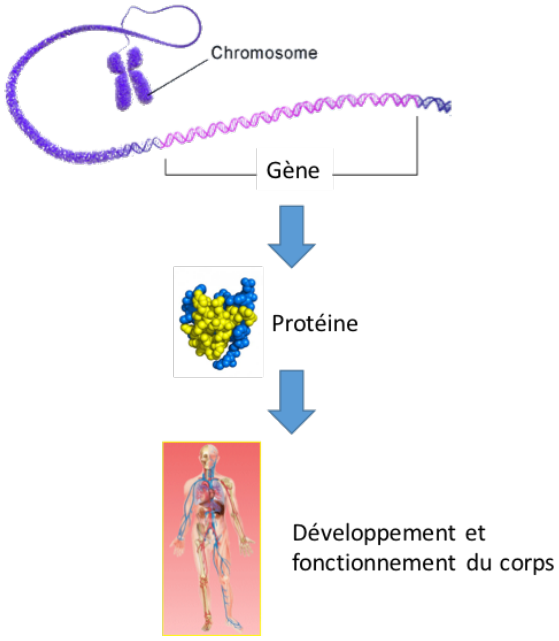


Visuel de la présentation d'une question du quiz associé à ce module sur Didask :



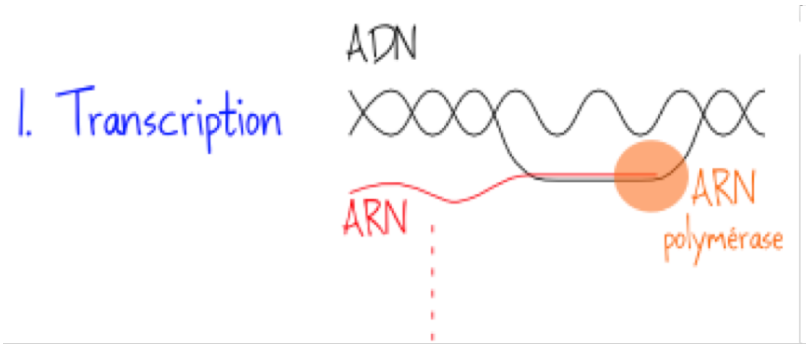
Module 5 : Les gènes

Chaque organisme peut être caractérisé par son phénotype, c'est à dire un ensemble de caractéristiques issues de l'expression de ses gènes. Ces caractéristiques peuvent être biochimiques (présence d'une enzyme clef du métabolisme), physiques (couleur des yeux) ou comportementales (comportements liés à la survie de notre espèce par exemple). Le gène est une portion de la molécule d'ADN nécessaire pour l'apparition d'une caractéristique, il est situé à n'importe quel endroit d'un chromosome. On peut simplifier cette définition en établissant la correspondance : un gène est la portion d'ADN codant pour une protéine. Le génome d'un individu correspond à l'ensemble de ses gènes. Chaque gène est présent en double dans notre génome. Ces deux copies d'un même gène appelées "allèles" sont pour l'une d'origine paternelle et pour l'autre d'origine maternelle. Les gènes occupent un emplacement déterminé sur les chromosomes : on parle de "locus". La quantité de gènes que nous possédons ne reflète pas forcément le niveau de complexité de l'espèce, par exemple le crapaud possède plus de gènes que nous ! Et nous ne possédons que 1% de différence avec le chimpanzé dans notre génome.



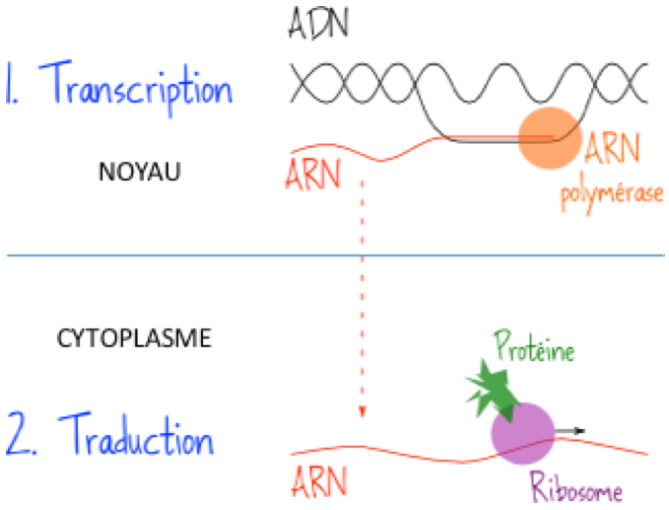
Module 6 : Du gène à l'ARN

La production d'une protéine nécessite 2 étapes : la transcription puis la traduction. L'ADN est localisé dans le noyau alors que les protéines sont produites dans le cytoplasme, donc à l'extérieur du noyau. L'ADN ne pouvant sortir du noyau, une molécule intermédiaire est donc fabriquée : l'Acide RiboNucléique messager (ARNm). L'ARNm peut passer du noyau vers le cytoplasme et peut donc jouer son rôle de messager. La transcription correspond à l'étape de copie d'une séquence d'ADN (= d'un gène) en un ARNm. Lors de cette étape, les deux brins d'ADN où se trouve le gène sont séparés momentanément. C'est l'enzyme l'ARN polymérase réalise la synthèse de l'ARNm à partir de la séquence d'un gène. L'un des deux brins est transcrit, c'est à dire « copié » et l'autre brin est non-transcrit. L'ARNm est en fait complémentaire au brin transcrit, il est donc monobrin et fait de nucléotides qui sont A, C, G, et la Thymine est remplacée par l'Uracile. L'ARNm sera ensuite "lu" pour mener à la synthèse d'une protéine.



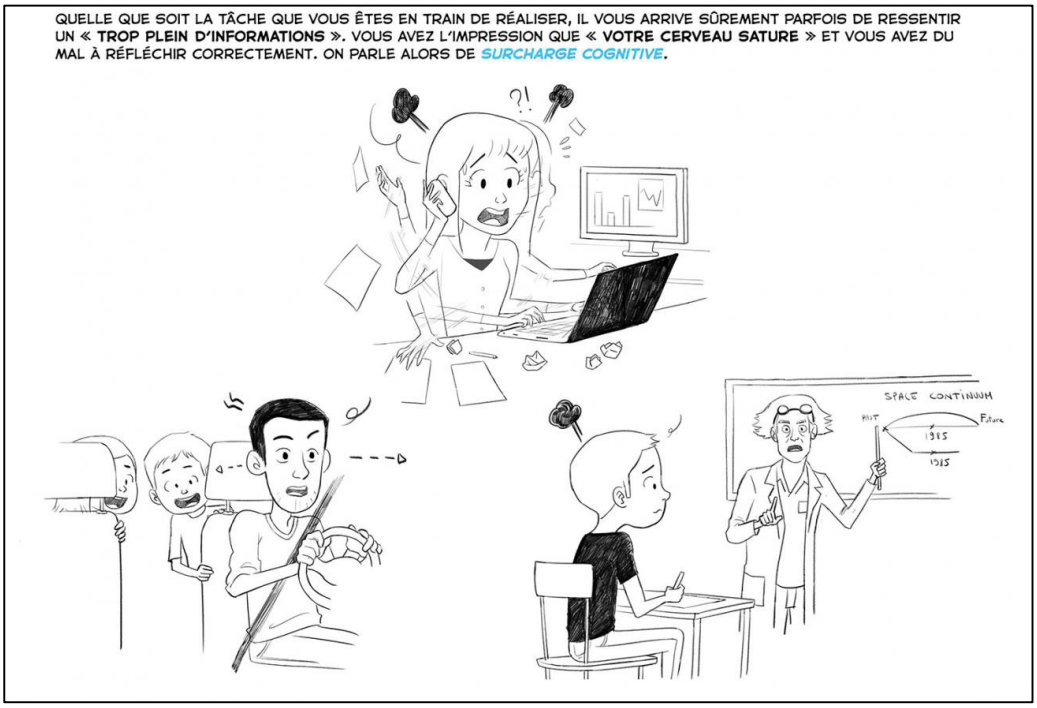
Module 7 : De l'ARN aux protéines

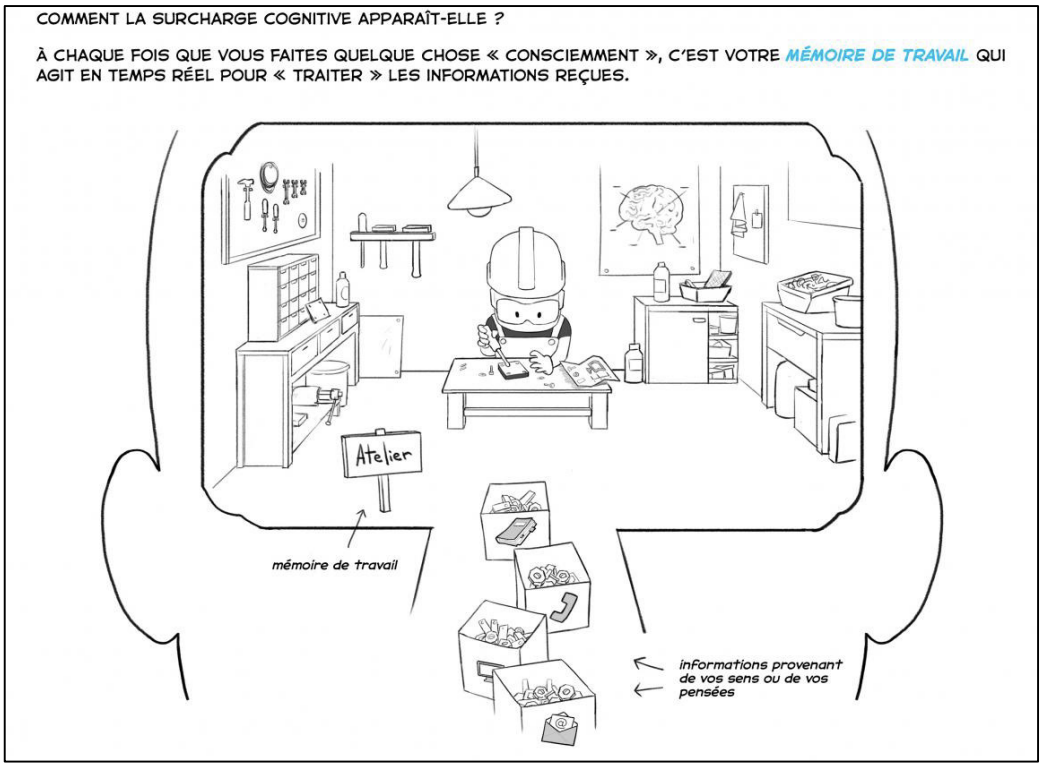
L'ARNm sort par de tous petits trous du noyau appelés les pores nucléaires. Une fois dans le cytoplasme, l'ARNm est en contact avec de nombreux ribosomes qui permettent la formation des protéines. Le ribosome est constitué de 2 parties : une grande et une petite. La petite est fixée à l'ARNm et la grande permet la production de la protéine. Ce processus s'appelle la traduction, et repose sur le code génétique. La traduction permet la production d'une protéine par assemblage des acides aminés les uns à la suite des autres. Le code génétique est le système de correspondance mis en jeu lors pour traduire l'ARNm en protéine : pour un triplé de nucléotides de l'ARNm, correspond un acide aminé. Les ribosomes traduisent donc l'ARNm en protéine par bloc de 3 nucléotides : on parle de « codons ». Le code génétique est présenté sous forme de tableau qui permet d'associant chacun des 64 codons avec l'un des 20 acides aminés. De nombreux ribosomes sont actifs sur le même fragment d'ARNm. Ceci permet la production de nombreuses protéines à partir d'un seul ARNm. A la fin de la traduction, la protéine est un enchainement linéaire d'acides aminés. Elle doit subir des changements dans sa structure pour être définitivement fonctionnelle : les acides aminés forment des liaisons entre eux pour replier la protéine et lui donner une structure tridimensionnelle.

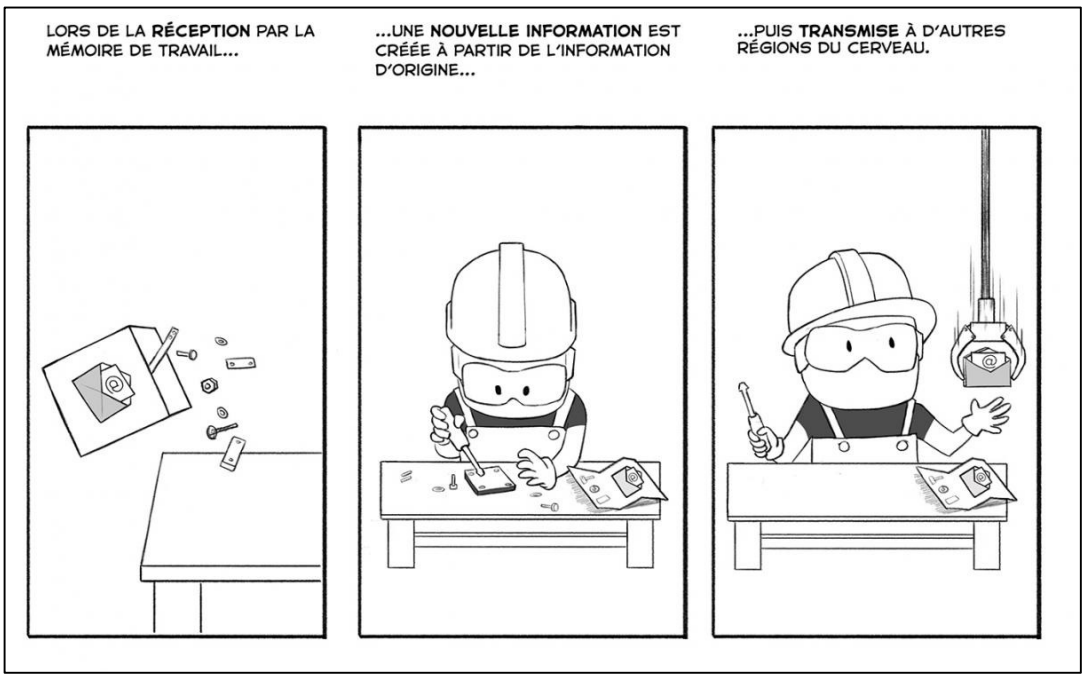


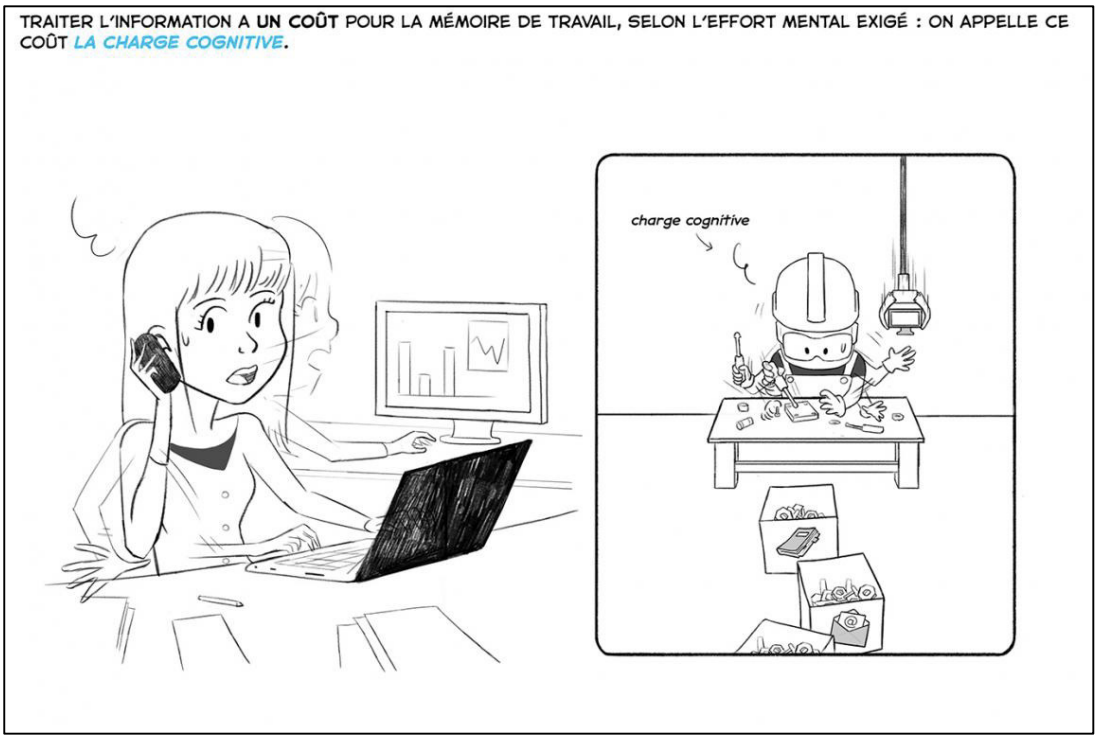
7.2 Annexe 2

Textes et illustrations du premier module « Reconnaître un problème de surcharge cognitive » de la formation créée par l'entreprise Cog'X (*Travailler autrement à l'ère du digital*) et retravaillée dans le cadre de l'expérimentation 2 présentée dans le Chapitre 4 (p75-98). Deux exemples de questions du quiz de ce module sont également présentés.





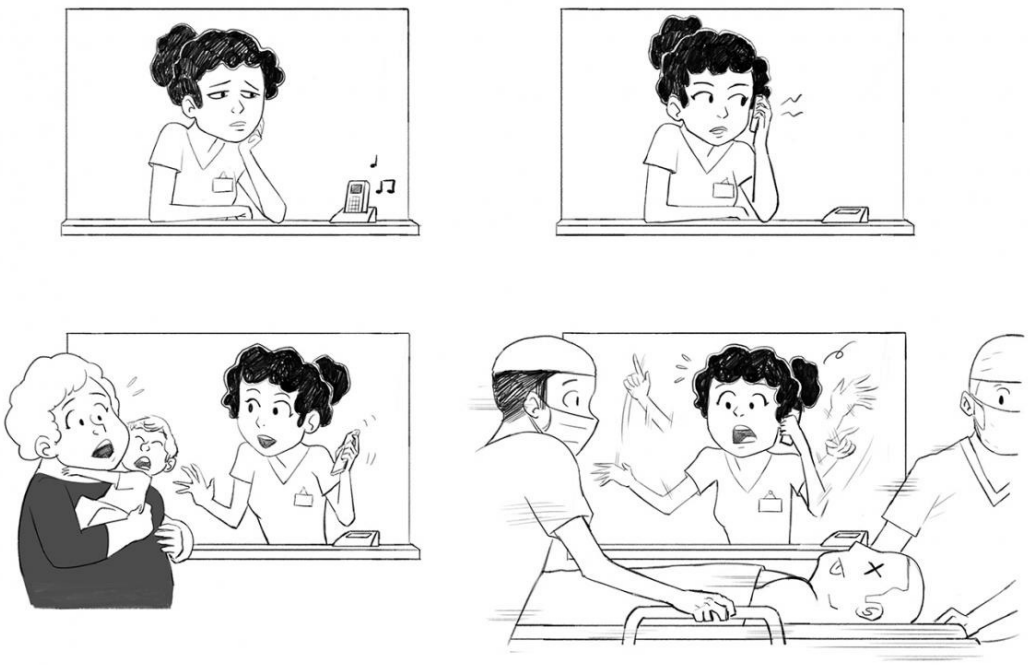

LORS DE LA **RÉCEPTION** PAR LA MÉMOIRE DE TRAVAIL...
...UNE **NOUVELLE INFORMATION** EST CRÉÉE À PARTIR DE L'INFORMATION D'ORIGINE...
...PUIS **TRANSMISE** À D'AUTRES RÉGIONS DU CERVEAU.


TRAITER L'INFORMATION A **UN COÛT** POUR LA MÉMOIRE DE TRAVAIL, SELON L'EFFORT MENTAL EXIGÉ : ON APPELLE CE COÛT *LA CHARGE COGNITIVE*.
charge cognitive

SI VOTRE MÉMOIRE DE TRAVAIL N'ARRIVE PLUS À TRAITER EFFICACEMENT LES INFORMATIONS, VOUS PASSEZ DANS UN ÉTAT DE *SURCHARGE COGNITIVE*.

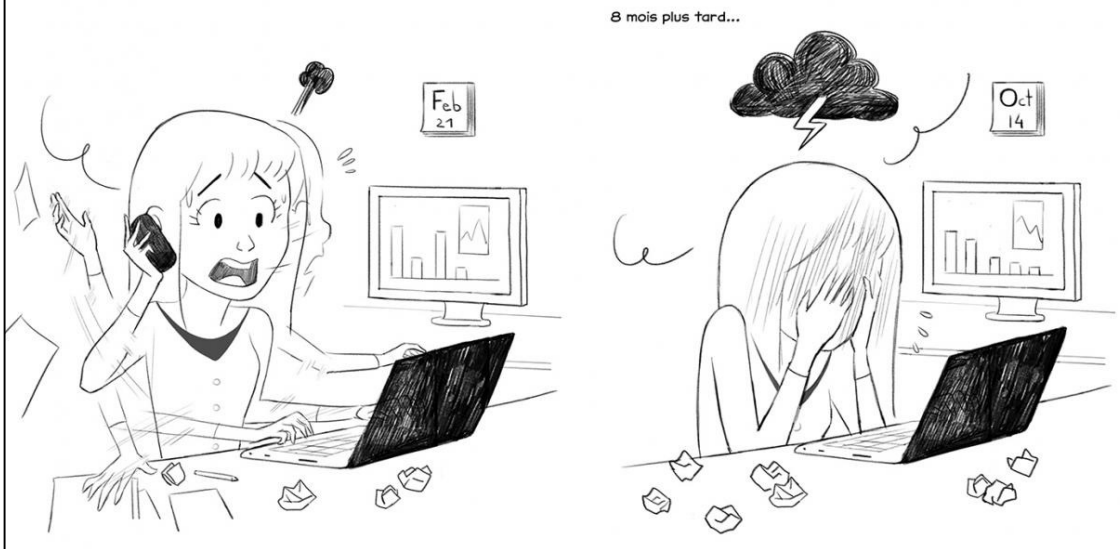
LA SURCHARGE MENTALE **ALTÈRE** LES FONCTIONS EXÉCUTIVES ASSURÉES PAR LE CORTEX FRONTAL.
VOUS **PERDEZ ALORS EN ATTENTION ET EN MOTIVATION, LA FATIGUE APPARAÎT** RAPIDEMENT ET **VOS ÉMOTIONS SONT AFFECTÉES.**

LE PASSAGE À L'ÉTAT DE SURCHARGE COGNITIVE EST **RÉVERSIBLE** ET PEUT SE PRODUIRE EN **QUELQUES SECONDES** SEULEMENT ! **OPTIMISER** LA QUALITÉ DE VOTRE TRAVAIL ET VOTRE BIEN-ÊTRE SUPPOSE DE MAINTENIR EN PERMANENCE VOTRE CHARGE COGNITIVE DANS **UN ÉTAT D'ÉQUILIBRE.**



ATTENTION À **NE PAS CONFONDRE** SURCHARGE COGNITIVE ET BURN-OUT !

LE BURN-OUT, OU SYNDROME D'ÉPUISEMENT PROFESSIONNEL, EST UN ÉTAT PATHOLOGIQUE GLOBAL QUI S'INSTALLE SUR LE LONG TERME. SON APPARITION EST FAVORISÉE PAR UNE SURCHARGE COGNITIVE FRÉQUENTE, MAIS COMBINÉE AVEC D'AUTRES FACTEURS.



Que vous soyez en train de travailler, de discuter avec des amis ou de regarder un film, votre mémoire de travail fonctionne de la même manière. La mémoire de travail est une sorte de mémoire vive, qui stocke temporairement les informations captées consciemment par l'attention, le temps que l'on puisse les analyser. Une conversation entre collègues à côté de votre espace de travail, ou une notification sur votre portable sont des informations perçues consciemment. A l'inverse, l'intensité de la lumière de votre bureau ou bien la plante verte qui est posée devant vous ne sont pas perçues consciemment.

Cette mémoire de travail a une capacité limitée en termes de :

- rétention des informations (quelques secondes)

- quantité d'information qu'elle peut stocker à un moment donné (entre 6 et 10).

La surcharge cognitive ou mentale intervient lorsque la mémoire de travail est « saturée », soit parce que les informations sont trop nombreuses, soit parce qu'elles sont trop complexes. Cette surcharge est un phénomène de court terme qui correspond donc à un effort mental très élevé. Elle peut se produire sur un temps très court et empêcher le traitement des informations durant ce court instant seulement. Un grand nombre d'information à traiter à la fois peut créer un sentiment désagréable et de la fatigue.

Attention, contrairement à ce que l'on pourrait penser :

- la surcharge mentale et la surcharge de travail ne vont pas toujours de pair ! La charge cognitive est liée à la gestion d'une tâche précise à un moment donné, indépendamment de la liste de tâches à réaliser. En revanche, on peut être en surcharge de travail mais avoir des ressources cognitives en adéquation avec les tâches à faire, évitant d'être en surcharge cognitive.

- La surcharge mentale ne correspond pas non plus au syndrome du burn-out, qui est un état installé dans la durée avec de nombreux autres symptômes qui ont des conséquences lourdes sur la santé et la vie socio-professionnelle. L'état de burn-out est un état d'épuisement professionnel lié à une perte progressive d'intérêt pour le travail, et une dévaluation de son propre travail. A l'inverse, la surcharge mentale peut être un état très temporaire durant lequel l'information à traiter excède les capacités de votre mémoire de travail.

Comment est-il possible de réduire notre charge cognitive et le risque de surcharge ? Certains facteurs peuvent aider à faciliter le traitement des informations comme l'expertise, la concentration, et la vigilance.

Question 1 du quiz associé à ce module :

« La mémoire de travail correspond à notre mémoire immédiate. En comparaison avec un ordinateur, ce serait sa mémoire vive, très différente de la mémoire à long terme. Cette mémoire de travail stocke temporairement les informations captées par notre attention pour qu'elles puissent être analysées. On dit que l'information est alors traitée : une autre information est créée et/ou une action est générée.
Selon vous, dans cet exemple ci-dessous, quelles informations sont traitées par la mémoire de travail de Mathieu, le jeune homme assis à son bureau ? » (Plusieurs réponses à sélectionner)

- La conversation de ses collègues ;
- La tâche affichée sur ses écrans ;
- L'intensité de la luminosité de l'open-space ;
- La plante posée sur son bureau ;
- Son portable qui clignote à cause d'une notification.

Explication : Les informations traitées par la mémoire de travail sont ici les informations perçues consciemment par Mathieu : il tend l'oreille vers la conversation qui a lieu juste derrière lui, et il est aussi attiré par la notification de son portable. Enfin il a pour objectif de réaliser la tâche affichée sur son écran et y porte donc une partie de son attention. Ainsi ces 3 éléments sont traités par sa mémoire de travail. En revanche, la plante verte ou l'intensité de la lumière de son bureau ne sont pas perçues consciemment. A moins d'y porter son attention volontairement, ces informations ne vont donc pas entrer dans la mémoire de travail.

Question 2 du même quiz :

« Parmi les propositions suivantes, laquelle définit le mieux la surcharge cognitive selon vous ? » (Une réponse à sélectionner).

- La surcharge cognitive est un phénomène de court terme qui provient d'un excès d'informations à traiter.
- La surcharge cognitive est un phénomène de moyen terme qui provient d'une charge de travail excessive.
- La surcharge cognitive est un phénomène de moyen terme aussi appelé syndrome d'épuisement professionnel (ou burn-out).

Explication : La surcharge cognitive correspond à un effort mental un effort mental très élevé lié à un grand nombre d'informations à traiter par la mémoire de travail. Elle peut se produire sur un temps très court et empêcher le traitement des informations durant ce court instant seulement. Durant ce quiz, il se peut qu'une phrase trop longue vous ait mis en situation de surcharge cognitive !

7.3 Annexe 3

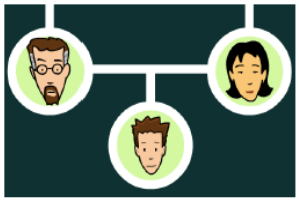
Textes et illustrations du cours sur l'ADN « Discovering DNA » créé dans le cadre de l'expérimentation 3 présentée dans le Chapitre 5 (p99-134), et adapté en anglais à partir des contenus de la première expérimentation (Annexe 1). Deux exemples de questions du quiz de ce module sont également présentés.

Overview

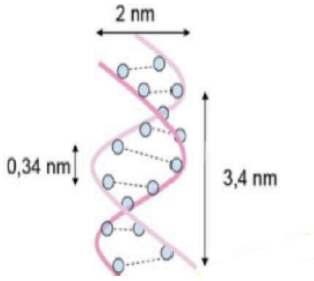
Heredity is the passing on of traits from parents to their offspring. Heritable traits are passed from one generation to the next via DNA.

DNA is a molecule embedded in the nucleus of each of our cells (the fundamental unit of all living beings). It consists of a series of units called «nucleotides».

All cells within a given organism carry the same genetic content. Human DNA carries an estimate of 20 000 to 25 000 genes. Gene expression, is the process that enables the synthesis of proteins. These proteins determine the characteristics of the organism (e.g. eye color).



Every gene is present twice in the genome. There is one «maternal» copy and one «paternal» copy. These 2 copies of the same gene are called alleles.



Schematic representation of the DNA double helix

DNA composition

DeoxyriboNucleic Acid (DNA) was first identified in 1869 by a Swiss scientist, Friedrich Miescher.

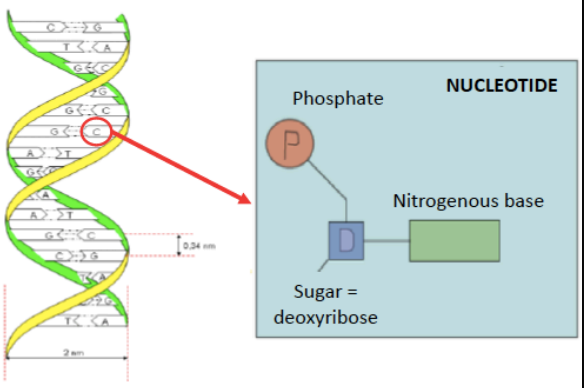
DNA is made of two long strands of nucleotides that are oriented in an antiparallel direction, because they run in opposite directions.

Each nucleotide consists of

- a sugar (deoxyribose);
- a phosphate;
- a nitrogenous base.

There are four nitrogenous bases: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). DNA bases pair up with each other - A with T and C with G - to form units called base pairs.

DNA is thus a succession of bases pairs linked to a sugar-phosphate backbone.



Zoom on the nucleotide unit: a sugar-phosphate compound forms a backbone from which the base protrudes

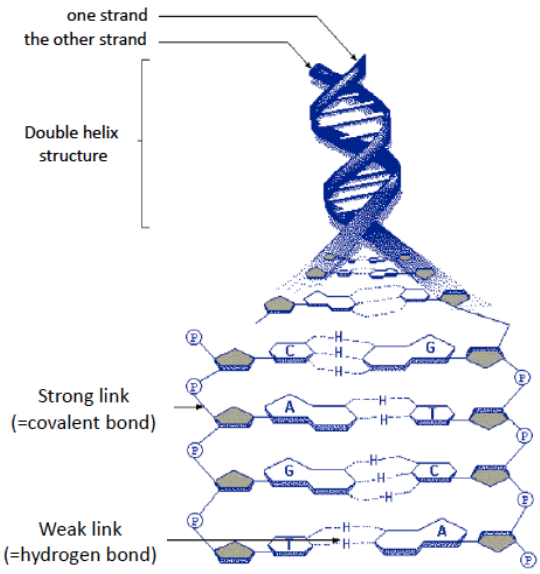
DNA structure

The DNA Double Helix was identified in 1953 by Watson and Crick (awarded a Nobel prize) and Rosalind Franklin, a major, although less honoured, contributor.

These scientists used X-ray methods as well as rules on atom density and the chemical strength of bonds to establish the 3D-structure of the DNA molecule.

By contrast to strong (« covalent ») links found in the sugar-phosphate backbone, the scientists demonstrated that base pairs were linked by weak hydrogen bonds. Base pairs are complementary because G always forms hydrogen bonds with C and A with T.

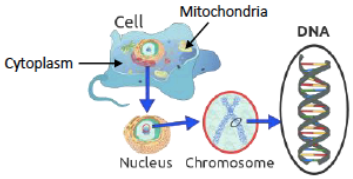
This type of bond is weak because the bond must be temporarily broken for gene transcription (reading and coping genes); and replication (another crucial process for the organism); and must be easily re-established following these processes.



The different links in DNA structure: strong for the backbone and weak in base-pairings between the two strands

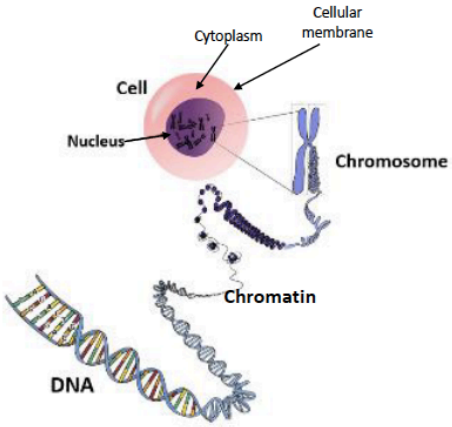
DNA in the cell

Later, scientists understood that each chromosome corresponds to a compact DNA molecule.



Indeed, DNA is condensed in a long strand called chromatin (barely visible under microscope). When chromatin is condensed it corresponds to a chromosome (very visible under the microscope when cells are dividing). There are a discrete number of chromosomes in the nucleus. Thus, several degrees of packing exist according to the cellular cycle. DNA molecule can be more or less condensed: its structure varies depending on whether access to DNA is required or not.

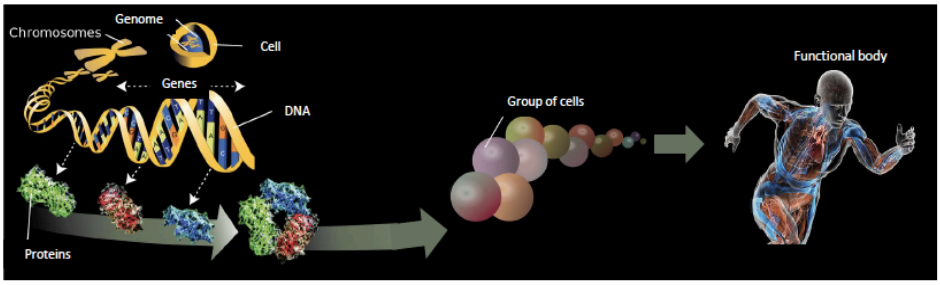
Even if the majority of DNA is in the nucleus, DNA also exists out of the nucleus, in small organelles called mitochondria. Mitochondrial DNA is circular, and codes for important molecules just like DNA in the nucleus.



The various forms of DNA in the cell: chromosome and chromatin

Genes

A gene is a sequence of DNA that codes for the structure of a protein, a molecule with a specific function. Proteins are the chief actors within our cells since they determine phenotypic characteristics. Phenotypic characteristics (=the phenotype) may be biochemical products (enzymes of metabolic pathways…), physical appearance (eye colour …) or behavioural (survival instinct).



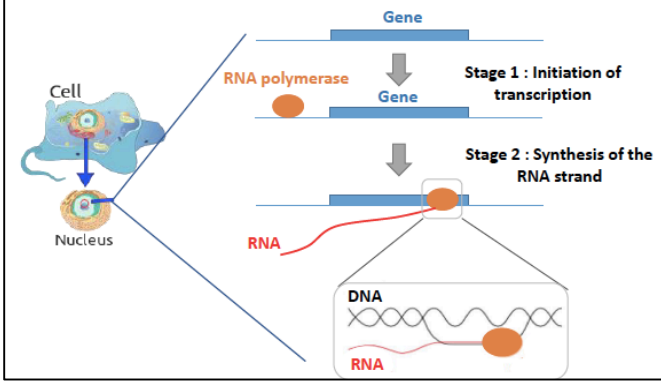
The genotype, also called genome, is defined as the complete set of genetic material of the organism, *i.e.* coding DNA + non-coding DNA. Genes are scattered all along the DNA molecule and only a tiny portion of the genes that encodes proteins.

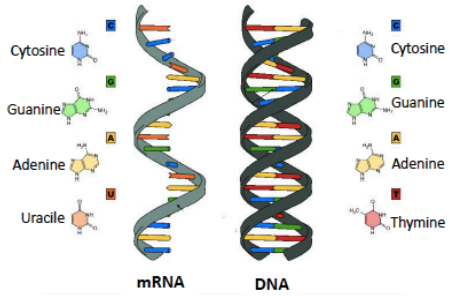
The size of the genome and the number of genes it contains do not reflect the complexity of an organism. For instance, tailed frogs have more chromosomes than human beings and only 1.5% of our genome differs from the Chimpanzee's.

From DNA to RNA

Genes expression enables proteins following two successive steps : transcription and translation.

Transcription allows a sequence of DNA with a gene to be copied into another form. The DNA strands separate where the gene of interest is located. Then RNA polymerase simultaneously unwinds the DNA double helix where the gene of interest is located and copies one strand of DNA into an intermediate molecule: a long RiboNucleic Acid (RNA). After a few processing steps, RNA is turned into mRNA in the nucleus, a perfect complementary copy of the DNA coding sequence.





mRNA is complementary to the copied DNA strand in the same way the DNA strands are, but mRNA is a single-strand molecule.

Like DNA, mRNA is made of nucleotides that are also A, C, G, but Thymine is replaced by another nucleotide called Uracile. mRNA is then transported out of the nucleus and is decoded into proteins within the cytoplasm, a gel-like material outside the cell nucleus delimited by the cell membrane.

7

From RNA to Protein

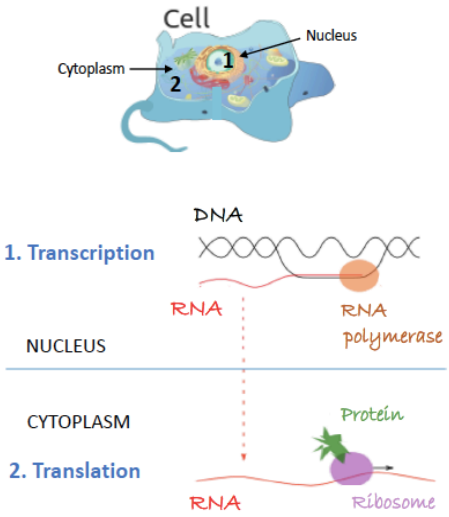
mRNA is transported into the cytoplasm of the cell *via* tiny holes in the membrane of the nucleus. Once in the cytoplasm, specialized organelles called ribosomes decode mRNA into amino acids that are chained to make a protein. This process is known as **translation**.

Specifically, translation converts the mRNA nucleotide sequence into an amino acid sequence that follows a genetic code (see table bellow). Each triplet of mRNA nucleotides corresponds to an amino acid. Thus, ribosomes translate mRNA into protein by blocks of 3 nucleotides, called "codons". There are 64 combinations of codons, but some are redundant, coding only for 20 amino-acids.

1st letter	Second Letter U	Second Letter C	Second Letter A	Second Letter G	3rd letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Amino-acids are indicated in blue (example : Phe for phenylalanine)

Ribosomes consist of two subunits, a small and a big one. Both subunits work together but undertake different actions. The small subunit decodes mRNA while the larger subunit performs protein synthesis.
Several ribosomes work simultaneously to produce large quantities of amino-acids chains from the same strand of mRNA. In the end, the linear amino-acids chain is matured and folded to acquire its 3-D structure : this is a functional protein.



Transcription in the nucleus and translation in the cytoplasm

8

Exemples de questions créés pour les quiz d'apprentissage :

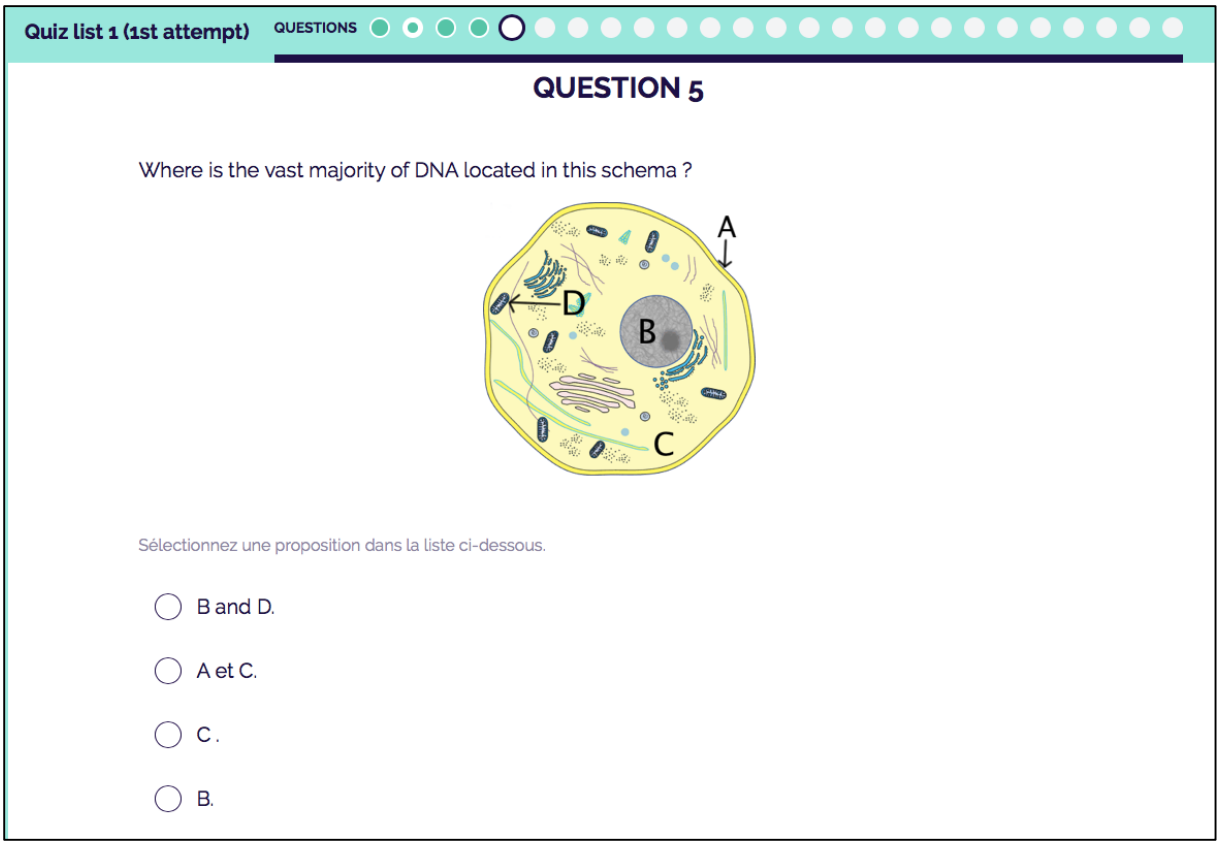
1. Which ones of these proposals are the correct ones?
- DNA and RNA nucleotides differ
- RNA is constituted by only one strand by contrast to double-strands DNA.
- Transcription may occur outside the nucleus
- DNA is in the nucleus. There, a complementary copy is transcribed as a messenger RNA (mRNA) before migrating outside of the nucleus.

Explanation: Transcription occurs in the nucleus. The process consists in copying one of the two DNA strands (coding strand) into a messenger RNA (mRNA). DNA and RNA nucleotides differ by the sugar moiety (desoxyribose and ribose, respectively) of the phosphate-sugar backbone. They contain the same three bases (A, C and G) but differ in the fourth one, U in RNA replacing T in DNA.

2. At the end of translation…
- protein is linear and functional
- protein is folded to acquire its 3D-tertiary structure becoming functional.
- bonds are created between aminoacids to fold the protein
- bonds are destroyed between aminoacids to unfold the protein

Explanation: At the end of the translation, the polypeptidic chain composed of successive aminoacids is linear (as a beads necklace). The linear chain is rapidly folded thanks to bonds between certain aminoacids giving rise to a functional protein.

Visuel de la présentation d'une question d'une question du quiz sur Didask :



Visuel de la présentation de la session de révision sur Didask (condition d'apprentissage avec cours):

8 Références

Cette bibliographie est valable pour l'Introduction Générale et la Discussion Générale de ce manuscrit.

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. https://doi.org/10.3102/0034654316689306

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

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

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory (Hove, England)*, *25*(6), 764–771. https://doi.org/10.1080/09658211.2016.1220579

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(15), 7041–7045. https://doi.org/10.1073/pnas.91.15.7041

Andler, D., & Guerry, B. (2008). *Apprendre demain. Sciences cognitives et éducation à l'ére numérique*. Retrieved from https://halshs.archives-ouvertes.fr/halshs-00791432

Arnold, K. M., & McDermott, K. B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*(3), 507–513. https://doi.org/10.3758/s13423-012-0370-3

Arthur, J., Waring, M., Coe, R., & Hedges, L. V. (2012). *Research Methods and Methodologies in Education*. SAGE.

Atkinson, R. C. (1972). Optimizing the Learning of a Second-Language Vocabulary. *Journal of Experimental Psychology*.

Auble, P. M., Franks, J. J., Soraci, S. A., Soraci, S. A., & Soraci, S. A. (1979). Effort toward comprehension: Elaboration or "aha"? *Memory & Cognition*, *7*(6), 426–434. https://doi.org/10.3758/BF03198259

Avis, J., & Fisher, R. (2018). *Teaching in lifelong learning: A guide to theory and practice 3rd Edition*.

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* (pp. 83–105). New York, NY, US: Psychology Press.

Barron, A. B., Hebets, E. A., Cleland, T. A., Fitzpatrick, C. L., Hauber, M. E., & Stevens, J. R. (2015). Embracing multiple definitions of learning. *Trends in Neurosciences*, *38*(7), 405–407. https://doi.org/10.1016/j.tins.2015.04.008

Barzagar Nazari, K., & Ebersbach, M. (2018). Distributed Practice: Rarely Realized in Self-Regulated Mathematical Learning. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02170

Bembenutty, H. (2009). Feeling-of-Knowing Judgment and Self-Regulation of Learning. *Education*, *129*(4), 589.

Benjamin, A. S., & Pashler, H. (2015). The Value of Standardized Testing: A Perspective From Cognitive Psychology. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 13–23. https://doi.org/10.1177/2372732215601116

Bianco, M. (2016). *Conférence de Consensus: Lire, comprendre, apprendre—Comment soutenir le développement des compétences en lecture? (Rapport technique)*. CNESCO.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392–402. https://doi.org/10.3758/s13421-012-0272-7

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* (New York: Worth Publishers, pp. 56–64).

Bjork, E. L., Soderstrom, N. C., & Little, J. L. (2015). Can Multiple-Choice Testing Induce Desirable Difficulties? Evidence from the Laboratory and the Classroom. *The American Journal of Psychology*, *128*(2), 229–239.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In *Practical aspects of memory: Current research and issues, Vol. 1: Memory in everyday life* (pp. 396–401). Oxford, England: John Wiley & Sons.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA, US: The MIT Press.

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

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Bjork, R., & Yan, V. (2014). Chapter 01: The Increasing Importance of Learning How to Learn. *Integrating Cognitive Science with Innovative Teaching in STEM Disciplines*. https://doi.org/10.7936/K7QN64NR

Bower, G. H. (1972). Mental imagery and associative learning. In *Cognition in learning and memory* (pp. vii, 263–vii, 263). Oxford, England: John Wiley & Sons.

Bowers, J. S. (2016). Psychology, not educational neuroscience, is the way forward for improving educational outcomes for all children: Reply to Gabrieli (2016) and Howard-Jones et al. (2016). *Psychological Review*, *123*(5), 628–635. https://doi.org/10.1037/rev0000043

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

Brewer, G. A., & Unsworth, N. (2012). Individual Differences in the Effects of Retrieval from Long-Term Memory. *Journal of Memory and Language*, *66*(3), 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Brown, P. C., Roediger (III), H. L., & McDaniel, M. A. (2014). *Make It Stick*. Harvard University Press.

Bruer, J. (2008). Building bridges in neuroeducation. *The Educated Brain: Essays in Neuroeducation*, 43–58. https://doi.org/10.1017/CBO9780511489907.005

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology*, *99*(2), 339–348. https://doi.org/10.1037/0022-0663.99.2.339

Carpenter, Shana K. (2009). Cue Strength as a Moderator of the Testing Effect: The Benefits of Elaborative Retrieval. *Journal of Experimental Psychology-Learning Memory and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, Shana 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

Carpenter, Shana K., & Toftness, A. R. (2017). The Effect of Prequestions on Learning from Video Presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

Carpenter, S.K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, *34*(2), 268–276. Retrieved from Scopus.

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

Cheryan, S., Ziegler, S. A., Plaut, V. C., & Meltzoff, A. N. (2014). Designing Classrooms to Maximize Student Achievement. *Policy Insights from the Behavioral and Brain Sciences*, *1*(1), 4–12. https://doi.org/10.1177/2372732214548677

Chew, S. L., Halonen, J. S., McCarthy, M. A., Gurung, R. A. R., Beers, M. J., McEntarffer, R., & Landrum, R. E. (2018). Practice What We Teach: Improving Teaching and Learning in Psychology. *Teaching of Psychology*, *45*(3), 239–245. https://doi.org/10.1177/0098628318779264

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology (2006)*, *70*(7), 1211–1235. https://doi.org/10.1080/17470218.2016.1175485

Choffin, B., Popineau, F., Bourda, Y., & Vie, J.-J. (2019). DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. *International Conference on Educational Data Mining (EDM 2019)*. Presented at the Montréal, Canada. Retrieved from https://hal.archives-ouvertes.fr/hal-02197659

Cnesco. (2017). *Notes remises dans le cadre de la conférence de consensus du Cnesco et de l'Ifé/Ens de Lyon «Différenciation pédagogique: Comment adapter l'enseignement pour la réussite de tous les élèves ?»*.

Coalition for Evidence-Based policy (July 2013). Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects. Coalition Evidence Reviews

Craik, F. I. M., & Tulving, E. (1975). Depth of Processing and the Retention of Words in Episodic Memory. *Journal of Experimental Psychology: General*.

Credé, M., & Kuncel, N. R. (2008). Study Habits, Skills, and Attitudes: The Third Pillar Supporting Collegiate Academic Performance. *Perspectives on Psychological Science*, *3*(6), 425–453. https://doi.org/10.1111/j.1745-6924.2008.00089.x

Davies, P. (1999). What is Evidence-based Education? *British Journal of Educational Studies*, *47*(2), 108–121. https://doi.org/10.1111/1467-8527.00106

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dekker, S., Lee, N. C., Howard-Jones, P., & Jolles, J. (2012). Neuromyths in Education: Prevalence and Predictors of Misconceptions among Teachers. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00429

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Chapter 3 - Spacing and Testing Effects: A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (pp. 63–147). Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742110530032

Demirel, M. (2009). Lifelong learning and schools in the twenty-first century. *Procedia - Social and Behavioral Sciences*, *1*(1), 1709–1716. https://doi.org/10.1016/j.sbspro.2009.01.303

Dessus, P., & Gentaz, E. (2006). *Apprentissages et enseignement*. Retrieved from https://hal.archives-ouvertes.fr/hal-00323320

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

Duchastel, P. C., & Nungester, R. J. (1984). Adjunct Question Effects with Review. *Contemporary Educational Psychology*, *9*(2), 97–103.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, *55*, 51–86. https://doi.org/10.1146/annurev.psych.55.090902.142050

Dudai, Y., Karni, A., & Born, J. (2015). The Consolidation and Transformation of Memory. *Neuron*, *88*(1), 20–32. https://doi.org/10.1016/j.neuron.2015.09.004

Dunlosky, J., & Rawson, K. A. (2019). *The Cambridge Handbook of Cognition and Education*. Cambridge University Press.

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

Ebbinghaus, H. (2013). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences*, *20*(4), 155–156. https://doi.org/10.5214/ans.0972.7531.200408

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The Testing Effect: Illustrating a Fundamental Concept and Changing Study Strategies. *Teaching of Psychology*, *39*(3), 190–193. https://doi.org/10.1177/0098628312450432

Eysenck, M. W., & Keane, M. T. (2010). *Cognitive psychology: A student's handbook*. Hove, Eng.; New York: Psychology Press.

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking Connections Between Research and Practice in Education: A Conceptual Framework. *Educational Researcher*, *47*(4), 235–245. https://doi.org/10.3102/0013189X18761042

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. *Metacognition and Learning*, *12*(2), 131–156. https://doi.org/10.1007/s11409-016-9163-9

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2010). Metacognitive Control of Learning and Remembering. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education* (pp. 109–131). https://doi.org/10.1007/978-1-4419-5716-0_6

Francis, G. (2012). The Psychology of Replication and Replication in Psychology. *Perspectives on Psychological Science*, *7*(6), 585–594. https://doi.org/10.1177/1745691612459520

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews. Neuroscience*, *6*(2), 119–130. https://doi.org/10.1038/nrn1607

Galand, B. (2009). Hétérogénéité des élèves en apprentissage: Quelle place pour les pratiques d'enseignement ? *Cahiers de recherche en éducation et formation*, *71*, 3.

Gerbier, E. (2011). *Effet du type d'agencement temporel des répétitions d'une information sur la récupération explicite*. Retrieved from http://www.theses.fr/2011LYO20029

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

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399.

Godden, D. R., & Baddeley, A. D. (1975). Context-Dependent Memory in Two Natural Environments: On Land and Underwater. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed Practice and Retrieval Practice in Primary School Vocabulary Learning: A Multi-classroom Study. *Applied Cognitive Psychology*, *30*(5), 700–712. https://doi.org/10.1002/acp.3245

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology*, *24*(8), 965–971. https://doi.org/10.1080/20445911.2012.722617

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 177–182. https://doi.org/10.1016/j.jarmac.2014.05.003

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The Trials of Evidence-based Education: The Promises, Opportunities and Problems of Trials in Education*. https://doi.org/10.4324/9781315456898

Greving, C. E., & Richter, T. (2019). Distributed Learning in the Classroom: Effects of Rereading Schedules Depend on Time of Test. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02517

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Hammersley, P. M. (2007). *Educational Research and Evidence-Based Practice*. SAGE.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

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

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

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 79–91. https://doi.org/10.1037/stl0000021

Hattie, J., Rogers, H. J., & Swaminathan, H. (2014). The Role of Meta-analysis in Educational Research. In *A Companion to Research in Education* (pp. 197–207). Springer.

Healy, A. F., Jones, M., Lalchandani, L. A., & Tack, L. A. (2017). Timing of Quizzes During Learning: Effects on Motivation and Retention. *Journal of Experimental Psychology. Applied*. https://doi.org/10.1037/xap0000123

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Hunsu, N., Adesope, O., & Bayly, D. (2015). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education*, *94*. https://doi.org/10.1016/j.compedu.2015.11.013

Isen, A. M. (2001). An Influence of Positive Affect on Decision Making in Complex Situations: Theoretical Issues With Practical Implications. *Journal of Consumer Psychology*, *11*(2), 75–85. https://doi.org/10.1207/S15327663JCP1102_01

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

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305–318. https://doi.org/10.1037/xap0000087

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621–629. https://doi.org/10.1037/a0015183

Jonsson, F. U., Hedner, M., & Olsson, M. J. (2012). The Testing Effect as a Function of Explicit Testing Instructions and Judgments of Learning. *Experimental Psychology*, *59*(5), 251–257. https://doi.org/10.1027/1618-3169/a000150

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, *38*(3), 209–215. https://doi.org/10.1007/s11251-009-9102-0

Kang, S. H. K. (2016). Spaced Repetition Promotes Efficient and Effective Learning Policy Implications for Instruction. *Policy Insights from the Behavioral and Brain Sciences*, 2372732215624708. 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

Kang, S. H. K., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2009). Metacognitive Control and Strategy Selection: Deciding to Practice Retrieval During Learning. *Journal of Experimental Psychology-General*, *138*(4), 469–486. https://doi.org/10.1037/a0017341

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., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do Learners Really Know Best? Urban Legends in Education. *Educational Psychologist*, *48*(3), 169–183. https://doi.org/10.1080/00461520.2013.804395

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, *104*, 18–33. https://doi.org/10.1016/j.compedu.2016.10.001

Kole, J. A., Healy, A. F., & Bourne, L. E. (2008). Cognitive complications moderate the speed-accuracy tradeoff in data entry: A cognitive antidote to inhibition. *Applied Cognitive Psychology*, *22*(7), 917–937. https://doi.org/10.1002/acp.1401

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences*, *25*, 67–72. https://doi.org/10.1016/j.lindif.2013.01.005

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology. General*, *133*(4), 643–656. https://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits - and costs - of dropping flashcards. *Memory (Hove, England)*, *16*(2), 125–136. https://doi.org/10.1080/09658210701763899

Korte, M. (2013). Cellular Correlates of Learning and Memory. In C. G. Galizia & P.-M. Lledo (Eds.), *Neurosciences—From Molecule to Behavior: A university textbook* (pp. 577–608). https://doi.org/10.1007/978-3-642-10769-6_26

Kozleski, E. B. (2017). The Uses of Qualitative Research: Powerful Methods to Inform Evidence-Based Practice in Education. *Research and Practice for Persons with Severe Disabilities*, *42*(1), 19–32. https://doi.org/10.1177/1540796916683710

Kramár, E. A., Babayan, A. H., Gavin, C. F., Cox, C. D., Jafari, M., Gall, C. M., … Lynch, G. (2012). Synaptic evidence for the efficacy of spaced learning. *Proceedings of the National Academy of Sciences*, *109*(13), 5121–5126. https://doi.org/10.1073/pnas.1120700109

Kvernbekk, T. (2017). Evidence-Based Educational Practice. *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.187

Landrum, T. J., & McDuffie, K. A. (2010). Learning Styles in the Age of Differentiated Instruction. *Exceptionality*, *18*(1), 6–17. https://doi.org/10.1080/09362830903462441

Latimier, A., Riegert, A., Peyre, H., Ly, S. T., Casati, R., & Ramus, F. (2019). Does pre-testing promote better retention than post-testing? *Npj Science of Learning*, *4*(1), 1–7. https://doi.org/10.1038/s41539-019-0053-1

Laurent, C., Baudry, J., Berriet-Solliec, M., Kirsch, M., Perraud, D., Tinel, B., … Ricroch, A. (2009). Pourquoi s'intéresser à la notion d' « evidence-based policy » ? *Revue Tiers Monde*, *n° 200*(4), 853–873.

Leberman, S., McDonald, L., & Doyle, S. (2006). *The Transfer of Learning: Participants' Perspectives of Adult Education and Training*. Gower Publishing, Ltd.

Lin, Y.-G., McKeachie, W. J., & Kim, Y. C. (2003). College student intrinsic and/or extrinsic motivation and learning. *Learning and Individual Differences*, *13*(3), 251–258. https://doi.org/10.1016/S1041-6080(02)00092-4

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

Litman, L., & Davachi, L. (2008). Distributed learning enhances relational memory consolidation. *Learning & Memory*, *15*(9), 711–716. https://doi.org/10.1101/lm.1132008

Littrell, M. K. (2011). *Influence of testing on memory, monitoring, and control* (Thesis, Colorado State University. Libraries). Retrieved from https://mountainscholar.org/handle/10217/51802

Littrell- Baez, M. K., Friend, A., Caccamise, D., & Okochi, C. (2015). Using Retrieval Practice and Metacognitive Skills to Improve Content Learning. *Journal of Adolescent & Adult Literacy*, *58*(8), 682–689. https://doi.org/10.1002/jaal.420

Makel, M. C., & Plucker, J. A. (2014). Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

Marzano, R. J. (2001). *A New Era of School Reform: Going Where the Research Takes Us*. Retrieved from https://eric.ed.gov/?id=ED454255

Mayer, R. E., & Anderson, R. B. (1992). The Instructive Animation: Helping Students Build Connections between Words and Pictures in Multimedia Learning. *Journal of Educational Psychology*, *84*(4), 444–452.

Mazur, J. E. (2015). *Learning and Behavior: Instructor's Review Copy*. Psychology Press.

Mazza, S., Gerbier, E., Gustin, M.-P., Kasikci, Z., Koenig, O., Toppino, T. C., & Magnin, M. (2016). Relearn Faster and Retain Longer: Along With Practice, Sleep Makes Perfect. *Psychological Science*, *27*(10), 1321–1330. https://doi.org/10.1177/0956797616659930

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). 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

McGaugh, J. L. (2000). Memory—A century of consolidation. *Science (New York, N.Y.)*, *287*(5451), 248–251. https://doi.org/10.1126/science.287.5451.248

Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *2*(1), 1–21. https://doi.org/10.1016/S0022-5371(63)80063-8

Mesnier, P., & Missotte, P. (2003). *La recherche-action.* Paris: L'Harmattan.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 978–984. https://doi.org/10.1037/xlm0000216

Miele, D. B., & Molden, D. C. (2010). Naive theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General*, *139*(3), 535–557. https://doi.org/10.1037/a0019745

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition*, *29*, 131–140. https://doi.org/10.1016/j.concog.2014.08.008

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval Practice in Classroom Settings: A Review of Applied Research. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00005

Mozer, M. C., & Lindsey, R. V. (2016). *Predicting and Improving Memory Retention: Psychological Theory Matters in the Big Data Era*. https://doi.org/10.4324/9781315413570-8

Mozer, M., Wiseheart, M., & Novikoff, T. (2019). Artificial intelligence to support human instruction. *Proceedings of the National Academy of Sciences*, *116*, 201900370. https://doi.org/10.1073/pnas.1900370116

Özsoy, G., & Ataman, A. (2017). The effect of metacognitive strategy training on mathematical problem solving achievement. *International Electronic Journal of Elementary Education*, *1*(2), 67–82.

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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. (2007a). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. *National Center for Education Research*. Retrieved from http://eric.ed.gov/?id=ED498555

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007b). *Organizing Instruction and Study to Improve Student Learning.*

IES Practice Guide. NCER 2007-2004. Retrieved from http://eric.ed.gov/?id=ED498555

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*, *9*(3), 105–119. https://doi.org/10.1111/j.1539-6053.2009.01038.x

Pasquinelli, E. (2011). Knowledge- and Evidence-Based Education: Reasons, Trends, and Contents. *Mind, Brain, and Education*, *5*(4), 186–195. https://doi.org/10.1111/j.1751-228X.2011.01128.x

Pasquinelli, E. (2012). Neuromyths: Why Do They Exist and Persist? *Mind, Brain, and Education*, *6*(2), 89–96. https://doi.org/10.1111/j.1751-228X.2012.01141.x

Pasquinelli, E. (2013). Slippery slopes. Some considerations for favoring a good marriage between education and the science of the mind–brain–behavior, and forestalling the risks. *Trends in Neuroscience and Education*, *2*(3–4), 111–121. https://doi.org/10.1016/j.tine.2013.06.003

Phelps, R. P. (2012). The Effect of Testing on Student Achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

Postman, L., & Keppel, G. (1977). Conditions of Cumulative Proactive Inhibition. *Journal of Experimental Psychology: General*.

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L. (2016). Optimizing Learning in College: Tips From Cognitive Psychology. *Perspectives on Psychological Science*, *11*(5), 652–660. https://doi.org/10.1177/1745691616645770

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J. G., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning*, *14*(1), 21–42. https://doi.org/10.1007/s11409-019-09189-5

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261

Robey, A. M. (2017). *The benefits of testing: Individual differences based on student factors*. Retrieved from http://drum.lib.umd.edu/handle/1903/19902

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254. https://doi.org/10.1146/annurev.psych.57.102904.190139

Roediger, H. L. (2013). Applying Cognitive Psychology to Education Translational Educational Science. *Psychological Science in the Public Interest*, *14*(1), 1–3. https://doi.org/10.1177/1529100612454415

Roediger, H. L., Finn, B., & Weinstein, Y. (2012). Applications of cognitive science to education. In *Neuroscience in education: The good, the bad and the ugly* (pp. 128–151). Oxford University Press Inc.

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, H. L., & McDermott, K. B. (2018). Remembering What We Learn. *Cerebrum: The Dana Forum on Brain Science*, *2018*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6353106/

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). New York, NY, US: Psychology Press.

Roediger III, H. L., Dudai, Y., & Fitzpatrick, S. M. (2007). *Science of Memory: Concepts*. Oxford University Press.

Rohrer, D., & Pashler, H. (2010). Recent Research on Human Learning Challenges Conventional Instructional Strategies. *Educational Researcher*, *39*(5), 406–412. https://doi.org/10.3102/0013189X10374770

Rohrer, D., & Pashler, H. (2012). *Learning Styles: Where's the Evidence?* (Vol. 46). Retrieved from https://eric.ed.gov/?id=ED535732

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

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review*, *21*(6), 1516–1523. https://doi.org/10.3758/s13423-014-0625-2

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ*, *312*(7023), 71–72. https://doi.org/10.1136/bmj.312.7023.71

Schraw, G., & Dennison, R. S. (1994). Assessing Metacognitive Awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Shea, P., & Bidjerano, T. (2010). Learning presence: Towards a theory of self-efficacy, self-regulation, and the development of a communities of inquiry in online and blended learning environments. *Computers & Education*, *55*(4), 1721–1731. https://doi.org/10.1016/j.compedu.2010.07.017

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450–466. https://doi.org/10.1080/02680939.2017.1280183

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Slavin, R. E. (2004). Education Research Can and Must Address "What Works" Questions. *Educational Researcher*, *33*(1), 27–28. https://doi.org/10.3102/0013189X033001027

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

Smith, C. D., & Scarf, D. (2017). Spacing Repetitions Over Long Timescales: A Review and a Reconsolidation Explanation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00962

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

Squire, L. R. (1986). Mechanisms of Memory. *Science*, *232*(4758), 1612–1619.

Styles, B., & Torgerson, C. (2018). Randomised controlled trials (RCTs) in education research –methodological debates, questions, challenges. *Educational Research*, *60*(3), 255–264. https://doi.org/10.1080/00131881.2018.1500194

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). Optimizing Second Language Practice in the Classroom: Perspectives from Cognitive Psychology. *The Modern Language Journal*, *103*(3), 551–561. https://doi.org/10.1111/modl.12582

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: From the classroom to online learning. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00495

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of Metacognitive Monitoring Affects Learning of Texts. *Journal of Educational Psychology*, *95*(1), 66–73.

Toppino, T. C., & Gerbier, E. (2014). Chapter Four - About Practice: Repetition, Spacing, and Abstraction. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 60, pp. 113–189). https://doi.org/10.1016/B978-0-12-800090-8.00004-4

Tricot, A. (2017). *L'innovation pédagogique*. Retz.

Tsai, C.-W., Shen, P.-D., & Fan, Y.-T. (2013). Research trends in self- regulated learning research in online learning environments: A review of studies published in selected journals from 2003 to 2012. *British Journal of Educational Technology*, *5*(44), E107–E110. https://doi.org/10.1111/bjet.12017

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109–118. https://doi.org/10.1016/j.jml.2010.11.002

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior*, *13*(2), 181–193. https://doi.org/10.1016/S0022-5371(74)80043-5

Uner, O., & Roediger, H. L. (2017). The Effect of Question Placement on Learning from Textbook Chapters. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2017.09.002

Unsworth, N., & McMillan, B. D. (2013). Mind Wandering and Reading Comprehension: Examining the Roles of Working Memory Capacity, Interest, Motivation, and Topic Experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 832–842. https://doi.org/10.1037/a0029669

Unsworth, N., McMillan, B. D., Brewer, G. A., & Spillers, G. J. (2012). Everyday attention failures: An individual differences investigation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(6), 1765–1772. https://doi.org/10.1037/a0028075

Webb, S., & Chang, A. C.-S. (2015). HOW DOES PRIOR WORD KNOWLEDGE AFFECT VOCABULARY LEARNING PROGRESS IN AN EXTENSIVE READING PROGRAM? *Studies in Second Language Acquisition*, *37*(4), 651–675. https://doi.org/10.1017/S0272263114000606

Weinstein, Y., Lawrence, J. S., Tran, N., & Frye, A. A. (2013). How and how much do students really study? Tracking study habits with the diary method. *Poster Presented at the Annual Meeting of the Psychonomic Society, Toronto, Ontario, Canada*.

Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology. Applied*, *22*(1), 72–84. https://doi.org/10.1037/xap0000071

Weinstein, Yana, 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). Retrieved from https://search.proquest.com/docview/1297338409/citation/99A7F419CA124520PQ/1

Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In *The psychology of*

evaluation: Affective processes in cognition and emotion (pp. 189–217). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(2), 439–455. https://doi.org/10.1037/xlm0000047

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *Npj Science of Learning*, *3*(1), 8. https://doi.org/10.1038/s41539-018-0024-y

Yoncheva, Y. N., Blau, V. C., Maurer, U., & McCandliss, B. D. M. (2010). Attentional Focus During Learning Impacts N170 ERP Responses to an Artificial Script. *Developmental Neuropsychology*, *35*(4), 423–445. https://doi.org/10.1080/87565641.2010.480918

Zaromb, F. M., Karpicke, J. D., & Roediger, H. L. (2010). Comprehension as a basis for metacognitive judgments: Effects of effort after meaning on recall and metacognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(2), 552–557. https://doi.org/10.1037/a0018277

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995–1008. https://doi.org/10.3758/MC.38.8.995

‘Wake up, Alice dear!' said her sister;
'Why, what a long sleep you've had!'
'Oh, I've had such a curious dream!' said Alice, and she told her sister, as well as she could remember them, all these strange Adventures of hers that you have just been reading about; and when she had finished, her sister kissed her, and said,
'It was a curious dream, dear, certainly: but now run in to your tea; it's getting late.'
So Alice got up and ran off, thinking while she ran, as well she might, what a wonderful dream it had been.

— Lewis Carroll
Alice's Adventures in Wonderland (1865, 1869), Chapter XII *Alice’s evidence.*

RÉSUMÉ

L'apprentissage scolaire implique généralement une phase d'étude des cours suivie d'une phase d'évaluation pour mesurer l'efficacité de la première. Dans cette conception, la phase de test sert à quantifier la réussite de l'apprentissage mais n'est pas envisagée comme outil d'apprentissage. Pourtant, de nombreuses études ont montré l'importance des tests comme processus actif pour consolider des connaissances à long terme. Une autre pratique peu utilisée est la distribution d'un même apprentissage dans le temps. Alors que le bachotage favorise la mémorisation à court terme, l'espacement des révisions favorise la consolidation sur le long terme. A l'heure actuelle, ces méthodes sont méconnues des enseignants et des élèves alors que les bénéfices de l'entrainement par récupération en mémoire et de l'espacement ont été répliqués de manière robuste avec des populations et contenus variés.

La start-up Didask a créé la plate-forme d'enseignement numérique du même nom en incorporant les résultats de ces recherches menées sur l'apprentissage. En collaboration avec Didask, ma thèse s'est articulée autour de la problématique suivante : quel est l'agencement optimal des phases de récupération en mémoire par rapport à la présentation des contenus d'un cours ?

Le premier chapitre est une introduction générale pour contextualiser cette recherche. Le second chapitre compile les résultats de trois méta-analyses portant sur l'effet de l'apprentissage avec tests répétés et espacés dans le temps. La première méta-analyse a montré un bénéfice significatif de l'apprentissage avec tests espacés dans le temps sur la rétention en mémoire par rapport à l'apprentissage avec tests massés dans le temps (g = 0.74). La seconde méta-analyse suggère en revanche un bénéfice non significatif de l'apprentissage avec tests espacés par rapport à l'apprentissage avec relectures espacées dans le temps (g=0.46). La dernière méta-analyse n'a pas démontré de différence entre un planning expansif d'apprentissage avec tests espacés (accroissement progressif de l'intervalle de temps entre les sessions d'apprentissage) et un planning uniforme (maintien du même intervalle entre les sessions, g = 0.032). L'ensemble de ces résultats confirme le net avantage de l'apprentissage avec tests espacés dans le temps, mais le planning d'espacement optimal n'est pas nécessairement expansif.

Les chapitres 3, 4, et 5 présentent trois expérimentations menées sur Didask. L'objectif était de mesurer les bénéfices de différents emplacements et planning de tests d'apprentissage sur la rétention en mémoire une semaine à un mois après la session d'apprentissage. Les résultats de l'Expérience 1 ont démontré qu'il est préférable de faire des tests d'apprentissage après la lecture du cours pour une meilleure mémorisation plutôt qu'avant. De plus, l'avantage de faire des tests après la lecture du cours était observé sur les informations non testées lors de l'apprentissage. Les résultats de l'Expérience 2 indiquent que le degré de granularité des contenus importe lors de l'apprentissage par relectures successives : un découpage fin permet une meilleure mémorisation qu'un découpage plus grossier. Par contre, l'importance de la granularité disparait dans la condition avec des tests d'apprentissage. Enfin, l'Expérience 3 a répliqué l'effet de récupération en mémoire mais pas l'effet d'espacement. Contrairement aux hypothèses de départ, l'effet de la combinaison des deux stratégies d'apprentissage n'était pas significatif. Néanmoins, l'apprentissage avec tests espacés permettait en moyenne une meilleure mémorisation que l'apprentissage avec tests massés, avec relectures espacées, et avec relectures massées dans le temps. Les résultats de cette thèse apportent une meilleure compréhension sur la manière d'utiliser l'apprentissage par les tests. Ils suggèrent également de nouvelles pistes de recherche sur l'optimisation des apprentissages pour promouvoir la consolidation de nouvelles connaissances.

MOTS CLÉS

Apprentissage, éducation, mémoire, numérique

ABSTRACT

Learning in formal school contexts usually unfolds in a study phase consisting in studying and memorizing the contents of a lecture followed by an assessment phase that allows teachers to evaluate their students' knowledge. Within this configuration, the assessment phase is confined to a pure testing of the learning without playing any further role in the learning process. However, research in cognitive science has provided strong evidence for the use of tests during the study phase to enhance long-term memory. The effect of testing one's knowledge with retrieval practice is actually more efficient than merely repeated reading (testing effect). In addition, inserting a time interval between episodes of the study sessions promotes better consolidation than massed practice that consists in learning the same piece of information in a repetitive fashion without inter-study interval (spacing effect). As opposed to commonly used learning strategies such as reading and cramming, retrieval practice and spaced learning have been shown to promote better long-term retention among a great variety of populations and contents. However, to this day, teachers and students still rarely employ these effective strategies.

The start-up company Didask designed a teaching platform with the same name that incorporates the results of research to offer an evidence based learning tool. In my PhD project, I have used Didask to raise the following issue: how can one optimize the benefits of retrieval practice relative to the exposure to learning contents of a lecture?

The first chapter of this dissertation is a general introduction that contextualizes the present research project. The second chapter presents three meta-analyses conducted on the effect of spaced retrieval practice. Spaced retrieval practice consists of repetitions of the same retrieval event distributed through time. Results of the first meta-analysis indicated a significant benefit of spaced retrieval practice on retention relative to massed retrieval practice (g = 0.74); while the second meta-analysis did not show a significant benefit of the spaced retrieval practice relative to spaced restudying (g=0.46). The last meta-analysis indicated no difference between the expanding schedule (i.e., spacing intervals between repeated reading episodes increase with every repetition) and the uniform one (spacing intervals are kept constant throughout the study phase; g = 0.032). Together, these analyses support the advantage of spaced retrieval practice, but do not support the wide belief that intervals should be progressively increased.

Chapters 3, 4, and 5 refer to three experiments run on Didask. These experiments investigated the benefits of different placements and schedules of retrieval practice episodes on long-term retention, one week and one month after the learning session. Experiment 1 demonstrated that learning with retrieval practice was better when retrieval practice episodes were placed after rather than before reading the learning contents. Moreover, the benefit of post-testing over pre-testing transferred to untrained information during learning. Experiment 2 demonstrated that the granularity of the learning contents is especially of interest when learning with successive readings: short reading passages led to a better retention than longer readings. However, when learning with retrieval practice, granularity did not matter for long-term retention. Finally, Experiment 3 replicated the testing effect but not the spacing effect. Contrary to our predictions, the effect of the combination of both learning strategies was not significant but using spaced retrieval practice led to better retention in average than massed retrieval practice, spaced reading, and massed reading. Overall, the results of this thesis give a better understanding on how learners should use retrieval practice and suggest new research questions in optimizing learning.

KEYWORDS

Learning, education, memory, digital tool