MEMORY AND METACOGNITION IN CLASSROOM LEARNING: THE ROLE OF ITEM ORDER IN LEARNING WITH PARTICULAR REFERENCE TO THE INTERLEAVING EFFECT

BY

JONATHAN FIRTH

A thesis submitted to the Graduate School in partial fulfilment of the requirements
for the degree
Doctor of Philosophy
in
Education
University of Strathclyde, Glasgow, Scotland
September 2020

ACKNOWLEDGEMENTS & DEDICATION

Firstly, I would like to thank my supervisory team, Professor Ian Rivers and Professor James Boyle. I feel very lucky to have had such supportive guidance and the wealth of expertise that they bring from both the psychology and education fields, and I very much enjoyed our regular meetings over the last four years.

I would also like to thank the school teachers who hosted me in their Psychology classes and helped to supervise pupils while I engaged in data gathering. Their help was invaluable, and I am also grateful to the pupils themselves for their time and willingness to complete the tasks, and the schools and local authorities involved. I am grateful for the assistance of faculty librarian Sarah Kevill, who provided valuable advice on literature search processes ahead of my systematic review. More broadly I am grateful for the support and encouragement of my colleagues as well as the wider education community during the time of my PhD. A part-time PhD can be a lonely task at times, and I feel lucky to work in a role where I am surrounded by so much enthusiasm for research, and to have had the opportunity to attend many conferences and events at which I learned from others and shared my own findings and thinking as they developed.

Finally, I would like to thank my wife Fiona and my children Hannah and Duncan for their support and forbearance throughout the process of my PhD research, the bulk of which was conducted alongside my full-time work (and perhaps rather too many other projects). This meant sacrifices all round, but they never wavered in their support or encouragement – not to mention frequent practical help and proofreading. I love you all.

PREVIOUS ITERATIONS OF THE CURRENT THESIS; PUBLICATION PLAN

Chapters 1 & 2

The work in these chapters has been adapted into an article and accepted by the *Journal of Education for Teaching* under the title "Boosting learning by changing the order and timing of classroom tasks: Implications for professional practice." (Vol 47, issue 1, pp. 32–46).

Chapter 3

A version of this chapter has been published in *Review of Education* (Vol 9, issue 2, pp. 642–684); an article based on the pre-registered protocol was published by *Social Science Protocols* (July 2019, pages 1–7).

Chapter 5

Study 1 was published as a practitioner research study in *Impact*, the Journal of the Chartered College of Teaching (spring 2018). Study 2 will be submitted as a paper to *Scholarship of Teaching and Learning in Psychology* in spring 2022.

Chapter 6

The introduction and Study 3 are partially based on a paper presented at BPS, Psychology of Education Section Annual Conference 2017, Edinburgh, and later written up for *Psychology of Education Review* (Firth, 2018c; reproduced in Appendix 11). Study 4 has been published by the journal *Studia Psychologica* for a special issue entitled "Individual and social predictors of irrational behavior and beliefs" (Vol 63, issue 2, pp. 204–220).

MEMORY AND METACOGNITION IN CLASSROOM LEARNING: THE ROLE OF ITEM ORDER IN LEARNING WITH PARTICULAR REFERENCE TO THE INTERLEAVING EFFECT

Abstract

Education needs to be effective, but previous research suggests that teachers and learners alike are not always aware of which practices lead to lasting, transferable learning and which do not. In particular, research into evidence-based teaching strategies such as the spacing effect, interleaving and retrieval practice have shown a striking mismatch between what classroom choices are supported by the evidence and metacognitive beliefs on the part of learners. In part, this is because these strategies make the process of learning more challenging and error-prone; what Bjork and Bjork (1991) refer to as desirable difficulties tend to lead to poorer performance in the short-term but better learning in the long-term. As such, they are often mistakenly rejected by learners who cannot easily perceived their benefits.

This thesis focuses on desirable difficulties that relate to the timing and order in which classroom examples are presented, and in particular on interleaving – the process of mixing or alternating the order of examples during learning. Previous research has established the strength and boundary conditions of other desirable difficulties such as the spacing effect (Cepeda et al., 2006) but a clear picture of these issues was lacking when it came to interleaving. A systematic review was therefore conducted to gauge the strength of the evidence on interleaving, and its potential for application to the secondary classroom. It found that interleaving (as compared to blocking) is associated with high effect sizes which persist across experimental designs and do not appear to be biased by the work of specific labs.

However, there was also a gap in the literature when it came to classroom-based field research on the technique, and very little work had been done which related directly to higher-order skills – a key element of many exam-based courses.

The next stage of this thesis was therefore to extend investigation of interleaving to classroom situations, focusing particularly on psychology teaching at school level. In a pilot study, high school students engaged in an introductory week for a psychology course experienced spaced and interleaved learning tasks, allowing a computer-based methodology to be tested but revealing no effect of interleaving in the context of brief presentations of factual information. A follow-up which used similar methodology applied to learning the skills of application and evaluation found an advantage of interleaving over blocking. The latter study also found a trend in favour of self-explanation – another desirable difficulty – that did not reach significance.

As desirable difficulties are often counterintuitive, this thesis also aimed to investigate whether teachers would endorse these techniques, and what might discourage them from doing so. A wide-ranging survey on learning and memory suggested that teachers’ beliefs about memory are generally more accurate than prior findings among the general public, but are out of line with the scientific consensus when it comes to desirable difficulties such as spacing and retrieval practice.

A follow-up study focused on three techniques in particular – interleaving, spacing, and retrieval practice (all desirable difficulties). New student teachers and in-service teachers were shown a set of vignettes, each of which presented a classroom situation relating to one of these techniques and required a response on a 7-point scale to indicate their belief in which of two alternatives (for example, interleaving vs. blocking) would lead to better outcomes. This study found that a minority of teachers favoured the techniques overall, though spacing was more widely endorsed (49% overall) than retrieval practice (30%), and interleaving was endorsed least of all (4%). No relationship was found between years of experience and accuracy across the sample of in-service teachers, and this group were less accurate than the student teachers, supporting the idea that experience does not help when it comes to adopting teaching techniques which are based on counterintuitive features of human memory.

Finally, the thesis addresses the implications of these findings for both teaching and professional learning. It considers the role of both interleaving and other related techniques, as well as looking at ways of inculcating research evidence into the profession. It is noted that flawed beliefs about learning and memory often link to teacher identity, and that this is a barrier when it comes to teachers' choosing to engage with evidence (or not). Some synthesis from the ideas can be achieved by considering the role of desirable difficulties as professional learning tools, and a series of recommendations are set out. The methodology used in the research – systematic reviewing and computer-based field experiments – is also evaluated, and directions for future work identified.

Table of Contents

List of Tables (main text only)

List of Figures (main text only)

1

Introduction: Learning, Meaning, and Desirable Difficulty

1.1 A Problem Facing Teachers

Teachers face a number of problems in the classroom. One of the major ones is ensuring that their students take on board new concepts within the limited amount of time available, and that they then retain those new concepts for future use, potentially for months or years.

More broadly, learning in school contexts can be seen as a matter of effectiveness – with effective learning techniques being preferable – and also of efficiency, given the limited time available. The work of the teacher, alongside the day-to-day business of managing discipline and motivation, involves ensuring the development of knowledge and skills among their students (whether via direct teaching or other methods) in such a way that they will be able to use those things themselves in future. As Soderstrom and Bjork (2015, p. 176) put it: "The primary goal of instruction should be to facilitate long-term learning—that is, to create relatively permanent changes in comprehension, understanding, and skills of the types that will support long-term retention and transfer".

This thesis explores a set of techniques which promise to be effective and efficient in helping learners[1] to take in, retain and use new conceptual learning – perhaps more so than standard educational practice. The techniques focus particularly on manipulating the order or timing of classroom tasks or activities – by 'interleaving' the order of presented items, or 'spacing' out practice opportunities.

[1] In this thesis I will use the term 'students' when referring to those who are engaged in school and university courses, regardless of age or level, and the term 'learners' when discussing learning more broadly. The latter term, for example, may be used to refer to points that encompass both classroom-based learning and informal learning.

Spacing and interleaving fall within a broader category of educational techniques which have been termed *desirable difficulties* because they can make the learning process slower and more error-prone during practice tasks, but also more effective over the long term. These techniques are preferable to some other interventions in that they are free to use, and in that they can be implemented directly by a classroom practitioner (or even by a student), but yet they appear to be little-known outside of educational and cognitive psychology (Roediger & Pyc, 2012). And when the techniques are encountered, teachers and students alike often mistakenly think that they are less effective (e.g. McCabe, 2011), and therefore avoid using them.

In Chapter 2, I will analyse the techniques that involve the manipulation of item order in some theoretical detail, exploring their potential for use in education. First, though, I intend to briefly set out exactly what is meant by ‘learning’ in classroom contexts, and thereby precis the key concepts and issues which will underpin the discussion and analysis throughout this thesis.

1.2 Defining Effective Learning in Classroom Contexts

Learning, in school and other contexts, involves a process of taking in new facts, skills, information and concepts (Bransford et al., 2000). Both the individual students on courses, their teachers, and society as a whole can benefit if learning proceeds in an engaging and effective way. Learners need to take new concepts on board successfully in order to achieve academically, and because much of this material (if a curriculum is well designed) will be of use in their life more generally.

Outside of self-regulated learning situations such as project work and revision, teachers make the key decisions about what tasks learners do and in what order they do them. They also make ad-hoc decisions about remediation when learning does not appear to be going to plan. As a result, teaching can be defined as “the process of making and implementing decisions, before, during, and after instruction, to increase the probability of learning” (Hunter, 1979, p. 62). It is therefore important to examine the basis of choices which lead to effective and durable learning. As the actual material taught is often externally mandated by the curriculum, this thesis will focus on an area that teachers can control and improve: the issue of how and in what order examples and tasks should be presented to students.

For all types of learning – at least as viewed from an information-processing perspective – a logical set of processes must take place. Information must enter a memory store (that is, it must be *encoded*). Learners then need to be able to *retrieve* and use what they have learned when relevant situations arise in the future. The *durability* of retention refers to whether a given fact, concept or skill persists over time after being learned, making it accessible to the learner for longer, or (ideally) not being forgotten at all. Finally, knowledge may need to be *transferred*, i.e. flexibly applied to situations which resemble the learning experience but also differ from it across one or more dimensions (Barnett & Ceci, 2002).

For students engaged in school and university courses, the durability of learning is important in terms of their need to retain what they have learned at least until the time of a relevant exam, and preferably beyond. They may need to transfer learning in the sense of using what they have learned in contexts outside of the classroom, or when recognising how exam questions link to the concepts and skills that they have previously learned and then applying those concepts and skills.

From a psychological point of view, learning is based on memory. This should not be taken to imply that learning is synonymous with rote rehearsal, or ‘memorisation’ (Firth, 2018a; see Appendix 1). The latter implies a very specific set of classroom or revision strategies whereby information is repeated without concern for meaning, while memory – as defined in psychology

– is the cognitive process involved in retention and use of new fact or skills. By this definition, it is impossible for a student to learn new things without using their memory in some way (Kirschner et al., 2006).

An event that leads to a student improving their skills, understanding or knowledge can be linked to changes which are simultaneously neural and psychological, i.e. affecting the structure of brain cells in a way that is reflected in what the learner can think, do and understand. Of course, such learning cannot be reduced to memorising separate pieces of information; memory is based around meaningful structures (Bransford et al., 2000). As Anderson (1984, p. 5) put it, "Knowledge is not a 'basket of facts'". I will return to the issue of how knowledge clusters together into meaningful 'schemas' in section 1.5, below.

1.3 Classic Memory Theories Versus the New Theory of Disuse

Nearly all types of learning will involve briefly representing information in a store known as working memory (henceforth WM), a multi-modal system responsible for allocation of attention and for temporary storage and processing of information (Baddeley, 2012). However, for learning to have been achieved, information must be encoded to long-term memory (henceforth LTM), as this is the cognitive store (or set of stores) responsible for long-lasting or permanent learning, and is therefore required for any retention or use of new learning over educationally-relevant timescales. Identifying effective classroom techniques that help students to form new LTM memory traces is therefore of considerable interest.

The classic *modal model* of memory (Atkinson & Shiffrin, 1968; Murdock, 1967) presents the learning process as dependent mainly on the activity of WM, with LTM taking a more passive role. The modal model is very limited, however. It does not take into account either the role of existing long-term memories in successful encoding, or the timescale of learning.

Memories are seen as being either present or absent, and there is no satisfactory explanation of why some memories might later be harder to retrieve than others.

Bjork and Bjork (1992) proposed a more sophisticated theory which accounts for some of those issues – the ‘new theory of disuse’. According to this theory, the *retrieval strength* of a memory – i.e. its accessibility at a given time – can be dissociated from its *storage strength,* which better reflects whether information has been learned as defined above. From this point of view, the aim of most classroom activities is (or at least, should be) to boost the storage strength of memories. In practice, however, many may primarily increase retrieval strength (boosting the temporary accessibility of information) and have little to no effect on storage strength (Soderstrom & Bjork, 2015). For example, overlearning of maths problems beyond the point of mastery has no benefit if assessed over a longer time scale, but spacing out the same problems across multiple distributed practice sessions does benefit learning (Rohrer & Taylor, 2006).

These features raise a counterintuitive but important implication of the theory: a period of forgetting can actually make later learning more effective, because retrieving and revising less accessible items boosts their storage strength, while more immediate repetition would only boost retrieval strength (Yan, Clark & Bjork, 2016). As a result, practice or re-study would best be delayed until information is on the point of being forgotten (Bjork, 2011).

Another implication of this theory is that any new trace in LTM is not fixed. Its strength can increase through multiple later experiences, and is subject to editing and distortion. Any initial conceptual learning experience will be subject to rapid and large-scale forgetting over subsequent weeks, and cannot it in itself be considered the end of the learning process.

Some of the classroom dynamics of encoding and retrieval are better explained by this theory than by the modal model. For example, a teaching activity may make a concept more easily accessible to a student if this is assessed within an hour or so, but this would not necessarily mean that its storage strength has significantly increased, and so student and teacher alike may incorrectly assume at the end of the lesson that a concept has been fully learned. Students may even be asked to reflect on what they have ‘learned’ as a form of self-assessment, and making this judgement on the basis of the temporary accessibility of the information – something I have observed during my own professional practice as a classroom teacher – but immediate judgements of learning tend to be inaccurate (Dunlosky & Tauber, 2016). Correspondingly, something with high storage strength (such as a concept which was practiced multiple times during a previous year of study) might be inaccessible at the point that a student needs it (for example, when questioned by a new teacher) because the item has not been activated recently, but it may nevertheless be possible to retrieve it in the right context or with a suitable prompt. The new theory of disuse is better able to account for these phenomena than is the modal model.

Despite the potential benefits of delaying practice, doing so may be strongly counterintuitive to teachers and learners (Bjork, 2011). As such, a barrier to the application of the new theory of disuse may lie in the metacognitive process such as learners' perceptions of how learning works best, or related choices made by teachers. I will explore this issue in Chapter 2, but for now, it is important to be aware that anything that increases subjective difficulty may be avoided unless the benefits are well understood.

It should also be noted that despite its potential for practical application and its more sophisticated fit with the realities of classroom learning, the new theory of disuse is still essentially an information-processing model of memory. It has been supported primarily by laboratory-based experimentation with its associated artificiality of both setting and tasks, and is relatively narrow in its conceptualisation of what memory involves. I will return to these flaws (see section 1.5). First, though, it is important to set out how the concept of delaying practice fits within the broader framework of learning techniques that have come to be known as desirable difficulties.

1.4 Desirable Difficulties

As noted above, a desirable difficulty is a technique or feature of a learning situation that makes learning both harder (and hence it is a difficulty) but also more effective (and hence it is desirable). The concept of distributing practice over a longer educational time period would fit this definition well on the assumption that it is more challenging for students to practise material after a delay because it is harder for them to retrieve the key information from LTM in order to engage in practice or answer review questions, but that doing so would arrest forgetting.

Desirable difficulties link directly to the new theory of disuse in that they make *performance* more difficult – introducing increased effort and making errors more likely – but are more likely to benefit *learning* than are easier alternative tasks. There is no definitive list of desirable difficulties, but some of the most widely discussed include:

- *Spacing*: inserting delays between different study sessions, rather than studying intensively. Examples include a teacher choosing to delay the classroom practice of a skill or review session until the following day, or students revising for tests in multiple short spells (e.g. 6 x 20 minutes across different days) rather than a single long session (e.g. 1 x 2 hours).

- *Interleaving*: varying the order of examples or tasks, such that different types of material are shuffled or alternated rather than appearing together. This makes it subjectively more difficult for learners as they have to switch to using a previously learned skill or concept rather than repeating the same thing several times in a row.

- *Self-explanation*: situations where learners think about and explain concepts to themselves (orally or in a written form) rather than taking them in more passively.

- *Retrieval practice*: situations where learners effortfully retrieve information from long-term memory rather than having it repeated to them as input. Examples include a student doing a review quiz rather than re-reading the same information, or a teacher holding a discussion about a new concept rather than verbally summarising it for his/her class. Retrieval practice could be considered a specific form of active rather than passive learning.

- *Variation*: changing the context or format of learning such that it becomes less predictable. Examples might include learning to drive in several different cars rather than just one, or practising a golf swing on multiple types of surface. Doing so would be more difficult, but would lead to a more flexible skill being developed, thereby boosting later transfer.

- *Summarising*: forming an overview of material, for example by writing brief notes on an article or lecture slide rather than copying it verbatim.

It will be noted that many of these difficulties can be combined. Practice tests could be spaced out over time via a technique that is sometimes termed ‘successive relearning’ (Rawson et al., 2013), and variation of context could be linked to the writing of summaries. Sometimes a combination is unavoidable; it would be difficult to complete a self-explanation without engaging in some degree of retrieval, for example, and interleaving inevitably involves increasing the spacing between items from the same category (see Chapter 2). However, it should be noted that from this list of desirable difficulties, spacing and interleaving are the particular focus of this thesis, while self-explanation will also play a role in the experimental work (see Chapter 5), and retrieval practice will be fundamental to the metacognitive aspects of the research (Chapter 6).

Self-explanation provides an easy-to-understand model of the desirable difficulty concept. It is an active approach to the reading of notes, textbooks or other material, whereby learners attempt to think deeply about material. As Roediger and Pyc (2013, p. 246), put it: “while reading a new page of text, [students] might be asking themselves: What facts on this page do I already know? What facts are new?” Sometimes this technique involves thinking aloud, though it could also be done silently. Most learners are unlikely to engage in self-explanation spontaneously, and it is usually prompted in some way by experimenters (for example by asking students to ‘think aloud’ as they read a text and to comment on points that are new to them; Wong et al., 2002) – a technique that can straightforwardly by applied in the classroom. Like other desirable difficulties, self-explanation can slow performance (Roediger and Pyc, 2013). However, it appears to boost transfer of skills-based learning (Berry, 1983) and is overall more effective than more passive approaches to reading (Bisra et al., 2018), meaning that there is a payoff for the difficulty.

Desirable difficulties are not just about making learning harder or more rigorous, however. Indeed, there are cases from the research where difficulties have been analysed and deemed undesirable. One notable case is disfluency. Early evidence (Diemand-Yauman et al., 2011) appeared to indicate that presenting information in a hard-to-read font such as ‘comic sans’ or in greyscale or italics would lead to better retention, and this has been referred to as the ‘perceptual-interference effect’. However, a series of studies by Yue et al. (2013) were unable to replicate these benefits when presenting participants with texts that were blurred or partially obscured. Similarly, a font that was developed specifically to make processing more difficult – ‘Sans Forgetica’ – was tested by Taylor et al. (2020). The research team were unable to replicate the claims made by the font’s creator about its benefits; participants who read texts in Sans Forgetica did no better on a later retention test than those who read the same texts in Arial, and in some cases they did worse.

Clearly, then, there is more to the desirable difficulties concept than simply making learning slower and more difficult. Why, we might ask, are some difficulties more desirable than others? A resolution comes from the work of McDaniel and Einstein (2005), who explain that the benefits (or harm) of a difficulty depend on the nature of the processing that it elicits. A difficulty that is distracting or causes extraneous processing will be harmful, while one that encourages more time and effortful focus on retrieval and usage-relevant practice of the target information and skills will be helpful.

As such, it is not enough to simply exhort learners to work harder at something; the hard work must be directed in such a way that it targets the key components of learning – that is, it increases storage strength and/or the ability to transfer what has been learned. A corollary of this is that if a difficulty causes learners to focus on irrelevant aspects of the task, then that difficulty will be undesirable as far as learning is concerned.

Developing this idea, we can posit that similarity between a practice session and a later test will improve performance, in line with the transfer-appropriate processing principle (Bransford et al., 1979). If this idea is applied to making practice – for example, in academic settings – more similar to the real world (which is often more unpredictable and less orderly), then it can be see that difficulties such as delays and variations will make learning more similar to the situations which learners will later face (Bjork & Bjork, 2019). As such, these difficulties will be desirable.

A similar argument is also made by Sweller and colleagues (2011) when they differentiate between germane (useful) and extraneous (harmful) cognitive load in classroom tasks, although their emphasis is on immediate processing and its demands on working memory (see also Chapter 2). The idea also fits well with Craik and Tulving’s (1975) classic work on depth of processing – more meaningful processing is better remembered, whereas processing which focuses on surface appearance (such as the sound or font of a text) is less memorable.

However, much of the experimental work on desirable difficulties has neglected the role of meaningful context. In studies of retrieval practice, a typical task involves presenting a pair of nouns (e.g. pond–FROG) and then later testing participants using one of the pair (e.g. pond–??) as a retrieval prompt (Kliegl et al., 2019). Some studies of spacing also present information that is stripped of context – individual words or terms, for example, or concepts on flashcards (e.g. Kornell, 2015), and tasks are often designed to be minimal in terms of their real-world context – Cepeda et al. (2008), for example, tested the spacing effect with lists of facts that have no connection with a participant's prior learning or interests.

This could be a problem when trying to apply desirable difficulties to the classroom, not least because a range of research studies have not only replicated the depth of processing effect but have extended it in various ways. Information is better remembered if it is relevant to the self (Rogers et al., 1977; Symons & Johnson, 1997), if it promotes creative problem-solving (Wilson, 2016) and if it builds on prior knowledge (e.g. Cook, 2006; Hardiman et al., 1989; Piaget, 1952). It is even better remembered if it relates to factors that present personal risks, leading to even deeper and richer encoding (Nairne et al., 2007; Röer et al., 2013).

More broadly, memory is intimately bound up with meaning and context, and forgetting is likely to proceed more slowly when newly-encountered information is highly meaningful to the individual. It is therefore worth taking some time now to consider the meaningful basis of long-term memories, and how this might affect authentic classroom-based applications of research into desirable difficulties.

1.5 Long-term Memory, Schemas, Meaning, and Preparedness to Learn

As noted at the start of this chapter, an important challenge for teachers is to assist students in taking on board concepts – concepts which will be meaningfully distinct, and retained in long-term memory in a way that allows them to be later distinguished from competing memories and which allows future stimuli to be correctly categories. Meaningful understanding is therefore a critical aspect of new learning, underpinning the concepts throughout this thesis.

The idea that meaning plays a fundamental role in human memory was established in the very early days of cognitive psychology. Meaningfully-similar items are easily confused in LTM in a way that does not occur in working memory (Baddeley, 1966), and the meaning-based subsystem of LTM – 'semantic memory' – is the one in which most of a learner's knowledge of the world is assumed to be stored (Squire, 2004; Tulving, 1972).

What makes a memory meaningful? Meaning is not an inherent property but depends on the interrelations of different things that a person experiences, and is rooted in culture (Putnam, 1973). A word or sentence therefore has meaning based on how well it can be linked to existing knowledge. According to developmental psychologist Jean Piaget this begins in infancy, with a developmental process of establishing 'schemas'[2] based around a child's movements and the function of objects (such as a schema for a toy or for a dog) which then become gradually more sophisticated as they learn to make ever more subtle distinctions between concepts (Piaget, 1926). A schema represents which categories and sub-categories an item belongs to (Rosch & Mervis, 1975), as well as relevant actions and sensations (Barsalou & Wiemer-Hastings, 2005), and information about how knowledge is used (Rumelhart, 1991). Schemas are therefore complex structures combining multiple modalities.

In terms of learning, schemas can be seen as underpinning a learner's ability to both recognise a stimulus and access stored knowledge about how best to respond to it; linking new information to existing (meaningful) schema knowledge has clear consequences for retention and transfer. Prior knowledge affects how students interpret material encountered in class (Rittle-Johnson et al., 2009), their ability to make sense of verbal explanations (e.g., Schwartz

[2] Or *schemata*; in this thesis I will use the simplified form.

& Bransford, 1998), as well as their facility with problem solving strategies (Koedinger & Roll, 2012) and transferring learning to new examples (e.g., Carpenter et al., 1998).

Just asking learners to think about items in a meaningful way can boost retention. For example, presenting learners with a list of nouns (e.g. ‘dancing/pasta/eggs/music) and asking them ‘do you like this?’ for each word leads to better retention than presenting the same words with the question, “does the word contain a letter ‘E’” (Hyde & Jenkins, 1969). Meaningful story-based explanations can be used as a mnemonic – the ‘narrative technique’. In one early study, linking words into a meaningful narrative led to recall of nouns that was 6–7 times as high as a control group (Bower & Clark, 1969), although meaning-based expectations can also distort what is retrieved, and even lead to false memories (Loftus, 2018).

The important role of meaning has not escaped educators; Bloom’s hierarchy (Bloom et al., 1956) sets out the idea that factual knowledge needs to link together meaningfully (the ‘comprehension’ or ‘understanding’ level of the hierarchy) before it can form the foundation of other skills (see also Chapter 5). The flipside of this is that learning, especially autonomous learning, can be highly challenging for learners who lack sufficient background knowledge to make sense of new information (Kirschner et al., 2006). Some educators have specifically applied story-based techniques to the classroom to draw on the powerful mnemonic benefit of meaningful narrative (Arya & Maul, 2012; Egan, 1985; Kokkotas et al., 2010).

While schema knowledge can affect performance, the reverse is also true – new learning experiences both add to and modifying a learner’s schemas. Schemas are therefore less like fixed category pigeonholes for incoming information and more like a self-organising computer file system that is constantly changing in order to better hold its contents. On occasion schemas are modified and subdivided to make room for information that does not fit with the existing structure in a process known as *accommodation* (Piaget, 1950). For example, learners who know about apes, monkeys and humans may not have considered the evolutionary similarity of those species or groups of species. In studying evolution, they may come to reconsider this, and begin to see monkeys and apes as more clearly different. The previous and updated structure of their knowledge can be represented as follows (see Figure 1.1):

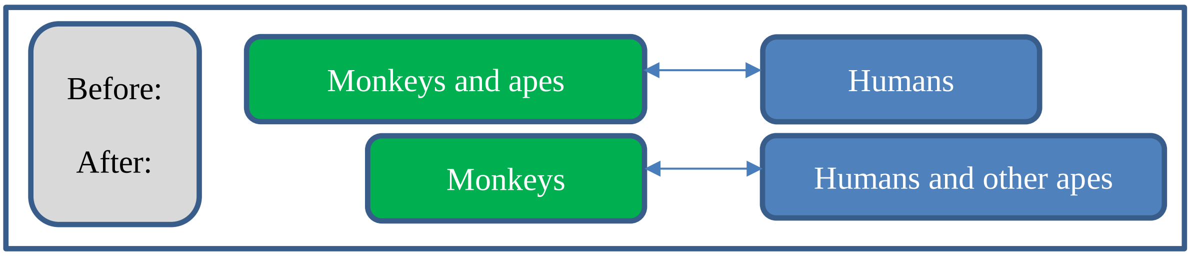


Figure 1.1: Conceptual change in learners.[3]

Conceptual change also facilitates transfer of learning (Goldstone & Son, 2005), and applications of such techniques are therefore not restricted to boosting memory for new information. As information is reorganised, new categories form in the learner's subjective understanding. This process is specific to the learner; a student who already has expertise in the topic will be able to assimilate new information rapidly without the need for conceptual change. They can also problem-solve better, seeing problems on a deeper and more conceptual level (Chi et al., 1982).

Overall, then, it is important that an analysis of how desirable difficulties can apply to the classroom takes account of the meaningful context of learning and the cumulative, social nature of schema-building. There is a subtle but important difference between taking in information and learning new concepts. And while laboratory studies of these phenomena are very promising, field studies are scarce. A key goal for this thesis will be to gather data that uses or simulates a richly-meaningful classroom context.

[3] I am very grateful to my daughter Hannah Firth for her assistance with several of the diagrams in this thesis.

As noted earlier, the new theory of disuse is flexible and provides a better fit with the dynamics of classroom learning than some classic models of memory do, yet it is a cognitive theory rooted in an information processing perspective, and does not fully account for matters such as the meaningful context as discussed above, not to mention the emotional associations or motivation to learn that might affect classroom learning. It will be important, going forward, that an analysis of the theory and of desirable difficulties remains cognisant of the fact that humans differ from information processers in that they actively attempt to construct meaning from their environment (Bruner, 1990).

This last point brings us back to the role of teachers in guiding, structuring and planning learning experiences for their students. Among the desirable difficulties that are particularly appealing, and which fall within what Roediger and Pyc (2012, p. 242) refer to as “low-hanging fruit” are those which rely mainly or entirely on changing the timing with which material is presented or practised – spacing and interleaving. Such techniques require minimal time and effort, in the sense that teachers do not need to devise entirely new materials, instead manipulating item order.

Due to this practical benefit, this thesis focuses on these two techniques (and on interleaving in particular). What exactly are the potential of interventions such as spacing and interleaving that focus on the timing of classroom examples? How should these be further researched, and how could they be applied in the classroom? And what are the major barriers in terms of professional knowledge and professional learning? Answering these questions is the main intention of the current research.

The sequence of the thesis is as follows: I carry out an exploration of educational techniques related to the desirable difficulties spacing and interleaving. This involves an analysis of how item order can be manipulated in educational contexts, what the cognitive underpinnings of the techniques in question are, and what metacognitive barriers and professional learning issues among teachers might currently inhibit their use. I also review the available research evidence and consider its limitations (Chapters 2–3).

I then consider the methodological challenges are presented by desirable difficulties and by interleaving and spacing in particular. Having reviewed and evaluated the key methodological options available, and I form a research plan for gathering supporting data in authentic contexts (Chapter 4).

2

Classroom Techniques based on Item Order – A Critical Review and Analysis of Related Literature

2.1 Introduction

As discussed in Chapter 1, there is a potential benefit to students and their teachers if a research-based understanding of long-term memory can be harnessed and put to use in the form of more effective strategies for the classroom or for independent study. Strategies based on desirable difficulties have the potential to be applied to teaching practice. Among these, spacing and interleaving are notable in that they relate mainly to modifying the timing of learning experiences.

Notwithstanding the limitations of some of the evidence as discussed in Chapter 1, these techniques may be very appealing to a teacher because implementing them is mainly a matter of planning. A classroom teacher can apply spacing or interleaving to their existing materials; it may involve some re-ordering of these materials such as changing the order of a set of slides or examples on a worksheet, but this may well be preferable to pedagogical changes which require entirely new materials or equipment. Accordingly, such interventions are low cost. They can also be applied on an ad hoc basis, for example when selecting examples to give to students verbally as part of a remediation process. Optimal timing of examples and practice may, in time, become an everyday part of a teacher's professional skills repertoire.

The spacing effect is based around the idea that even when information is successfully encoded and stored, it is at risk of being forgotten prior to the point that the learner needs to use it. During the intervening time, the likelihood of successful retrieval declines rapidly at first but at a slower rate over time (Ebbinghaus, 1885/1964). It is therefore important for educators to look not just at how people encode and make sense of new information, but also at how practice can be timed to interrupt or eliminate forgetting.

Meanwhile, within a learning task, an interleaved arrangement can boost the comparisons that learners draw between different items. Instead of the standard practice of putting like with like (for example, illustrating information about succulent plants with several examples that each show a succulent plant), like is put with unlike (so the information would be illustrated with one image of a succulent plants plus examples of other contrasting types of plants). Again, doing so is largely a matter of timing and planning for the practitioner; existing materials can be used, but presented in a different order.

Given the potential benefits of these two techniques and their relative ease of application, I will explore their theoretical underpinnings as well as scrutinising the evidence that supports them.

It should be noted that I am approaching these issues as a teacher as well as a researcher; the current research was motivated by my own professional desire to identify more effective learning techniques. The analysis in this chapter could therefore be viewed as the perspective of a 'teacher-researcher'. As such, I will attempt to explore how the techniques might be applied educationally, what key cognitive processes underpin (or limit) them in classroom practice, and what metacognitive factors might constrain their use in schools. Doing so will also provide the necessary grounding for research that appears later in this thesis: identifying boundary conditions of the techniques (Chapter 3), and obtaining a better understanding the current state of knowledge in teachers who will be the primary implementors of the strategies (Chapter 6). A larger body of applied experiments is also needed in order to inform practice, and the practical elements of this thesis will attempt to go some way towards addressing this need (Chapter 5).

2.2 Techniques for Manipulating Item Order in the Classroom

2.2.1 Spacing.

As highlighted in Chapter 1, teachers' efforts to promote effective learning may be undermined by rapid and large-scale forgetting over subsequent weeks, but the process of forgetting may be slowed or arrested by spacing out a learner's opportunities for practice. The spacing effect is a reliable and well-evidenced phenomenon which has been demonstrated in a wide range of educational contexts (Cepeda et al., 2006; Son & Simon, 2012). The essential finding is that information is more effectively encoded to LTM if it is presented in a spaced (or 'distributed') fashion than if it is presented close together in time (which is usually referred to as a 'massed' presentation).

Although often tested with simple words or isolated facts, spaced practice has also been shown to be helpful for a range of school-relevant tasks such as reading maps (Carpenter & Pashler, 2007), foreign language learning (Bahrick et al. 1993; Bird, 2010), typing (Baddeley and Longman 1978), and maths practice (Rohrer and Taylor 2006).

It appears to be important that items are well learned in the first study session. Rawson and Dunlosky (2011) found that during initial learning, students benefit from at least three successful attempts to retrieve concepts, after which re-study can occur at widely spaced intervals. It is likely that if learners only briefly look at a set of information and do not fully absorb or understand it, a long delay before a review will be less helpful, if indeed it is helpful at all.

In a typical experiment into the effect there is a gap between two study sessions which is either shorter or longer across the different experimental conditions. There is also a controlled retention interval between the end of the second study session and a final test, meaning that the period of time during which forgetting could occur is kept constant across conditions. This basic design is shown in Figure 2.1 (see also Chapter 4).

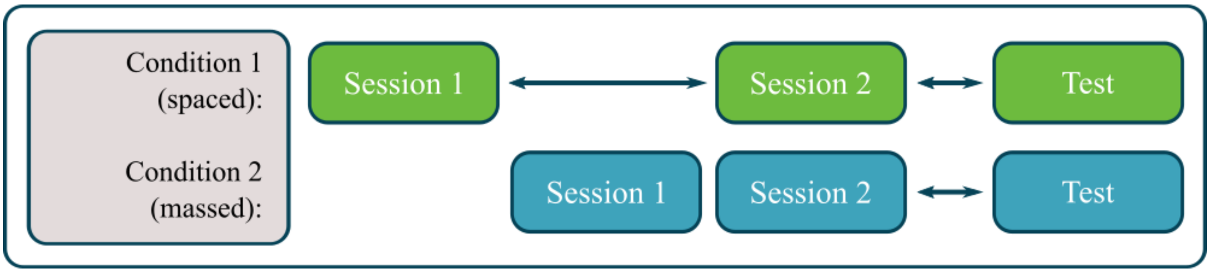


Figure 2.1: Typical design of an experiment into spacing. Gaps between coloured boxes reflect time periods which could be brief (e.g. 1 minute) or longer (e.g. 1 week).

Many of the early laboratory studies into spacing (e.g. Landauer & Bjork, 1978) involved short delays between practice sessions, but some more recent studies have used delays relevant to the timescale of classroom learning. For example, Kapler et al. (2015) followed a science lecture with a review which was delayed either by one day (condition 1) or eight days (condition 2). At a test five weeks after the review, they found superior performance among participants from the eight-day delay condition. Some early evidence suggested that the spacing effect disappears or even reverses with short delays, but a meta-analysis by Cepeda et al. (2006) found that this only occurs with delays of a few seconds, and is therefore unlikely to be relevant to classroom scheduling.

Why does spacing help? One possibility is that over longer timescales, repetition may serve to remind learners of their earlier learning experiences, thereby activating relevant episodic memories and making them easier to access in future (Benjamin & Tullis, 2010).

It is difficult to put an exact time on the optimal duration for a delay, but in a study of trivia facts, Cepeda et al. (2008) found that the benefits of spacing could extend to up to a year, and concluded that most educators are likely to be spacing too little, as the benefits of a larger delay outweighs the risk of increased forgetting. They also concluded that the optimal delay depends on the retention interval; the longer learners need to remember the information for, the more widely the initial study sessions should be spaced. There are cases, then, when it would make sense to avoid spacing (such as when a final test occurs very soon after study) but for lifelong learning, wide spacing should be recommended (Cepeda et al., 2008; Son & Simon, 2012).

Another factor that may affect optimal spacing is the level of understanding that a student has gained. As discussed in Chapter 1, prior knowledge affects most aspects of later learning, and items which are more deeply understood – i.e. those with better-developed links to existing schemas – are better remembered at a later point (Brainerd et al., 1990). This implies that forgetting has proceeded more slowly; from the point of view of the new theory of disuse, meaningful items remain more accessible for longer. As optimal spacing depends on the speed of forgetting, it is therefore meaning-dependent, too. Less inherently meaningful information (such as arbitrary vocabulary and terminology) may require more rapid follow-up.

However, given that meaningful understanding is not a property of the material to be studied but is an interaction between the learners and the material (as the extent to which something is meaningful to a learner depends in part on their existing concept knowledge), optimal timing is a highly context-specific judgement.

2.2.2 Interleaving.

The term interleaving refers to variability within a set of tasks or example items such that each item is immediately followed and preceded by an example of a different type, rather than appearing in blocks of the same type of item repeatedly. The latter is usually termed a ‘blocked’ sequence. For example, presentations of item types ‘1’, ‘2’ and ‘3’ could be shown blocked (111, 222, 333) or interleaved (123, 123, 123).

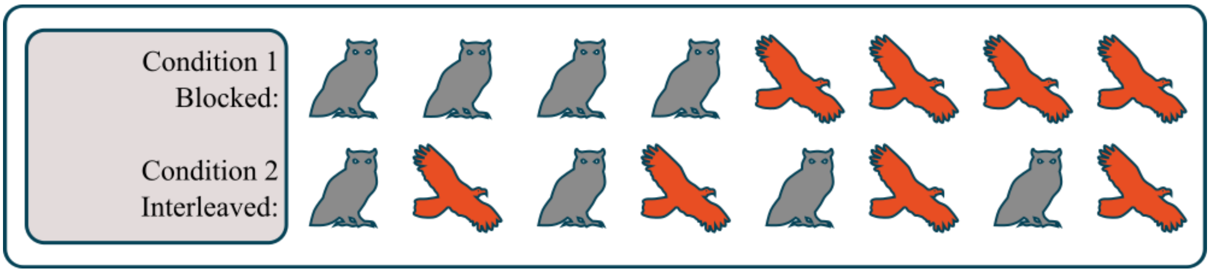


Figure 2.2: Contrast between a blocked and an interleaved sequence. In condition 1, examples of one species are shown as a block (this would consist of different examples of the same species rather than the same image repeated), followed by a block of examples of the next species. In the interleaved condition, examples of contrasting species are alternated.

If learners were being taught about species of bird, blocking would involve showing multiple examples of the same species of bird consecutively, while interleaving would involve showing examples of one species of bird, then an example of a different species, then yet another species, and so on (see Figure 2.2).

The benefit of interleaving (sometimes termed the 'interleaving effect'), has been investigated in numerous contexts, many of which are of direct relevance to education. It has been found to boost mathematics learning in seventh-grade school pupils (e.g. Rohrer et al., 2015), the inductive learning of science concepts and examples among university students (Eglington & Kang, 2017; Rawson et al., 2015), and the inductive learning of images of animal species and modern art paintings among adult experimental participants (Birnbaum, Kornell, Bjork & Bjork, 2013; Kornell & Bjork, 2008).

Interleaving appears to make learning more difficult at first but more durable over the longer term (Yan, Clark & Bjork, 2016). It links closely to the concept of 'contextual interference' in motor learning, whereby physical tasks such as throwing or catching benefit from being mixed with different movements rather than repeated directly (see Magill & Hall, 1990).

There is a connection between interleaving and the spacing effect, and the two are conflated by some authors. This is because interleaving items inevitably increases the gaps between one example item and the next (as can be seen in Figure 2.2, where the examples of item type 1 are more widely spaced in the interleaved presentation than in the blocked presentation). Indeed, when Kornell and Bjork (2008) rekindled interest in the concept with their study of interleaved learning of artists' styles, they initially attributed their findings to spacing. However, Kang and Pashler (2012) carried out a replication where spacing was held constant, using filler images to increase the temporal space between one blocked item and the next, and found that the interleaved condition was nevertheless superior to the spaced or blocked conditions. Taylor and Rohrer (2010) have found that for mathematics practice, both spacing and interleaving have separate beneficial effects, and Birnbaum et al. (2013) also found support for a separate effect.

Recent research into concept learning has tended to agree that the relationship between interleaved items is of critical importance, rather than just the interference between items, the difficulty of the task, or the spacing. An order where different types of items are presented together or consecutively makes it easier for learners to compare and contrast items and thereby notice subtle conceptual differences between them, an idea known as the *discrimination-contrast hypothesis* (Birnbaum et al., 2013). This fits with a body of research into categorisation which tends to show that highlighting differences (i.e. making discriminative contrast more salient) has a more beneficial effect on category learning than highlighting similarities (Higgins & Ross, 2011). Subtle differences are very hard for learners to notice if viewed during separate study sessions.

Supporting this idea, Hausman and Kornell (2014) mixed the study of Indonesian vocabulary with the learning of biology terms using a sample of adult participants recruited online, and did not find a benefit of interleaving, presumably because the two sets of material were too conceptually distant to be productively be compared or organised. Interestingly, though, Eglington and Kang (2017) did not find that explicitly highlighting differences during the learning phase impacted on the benefits of interleaving compared to blocking, perhaps suggesting that conscious attention to differences is not required.

There is an argument that blocking learning may at times be useful; just as interleaving boosts facilitates comparing and contrasting different items, blocking may help learners to notice similarity between items, especially when the items that make up a category are relatively diverse (e.g. members of the category 'amphibians'), and the commonalities between them are subtle. Carvalho and Goldstone (2014) have provided experimental evidence that this is the case; according to their analysis, a key theoretical factor is the increased attention that learners pay to similarities and/or differences, modulating later recall. However, learners may also overestimate the importance of surface similarities among items, after which repeated practice may simply reinforce their error. As explained by Birnbaum (2014): "*Noticing within-category correlations can thus become a liability and may steer the learner away from learning features critical for telling categories apart*" (pp. 11–12).

Overall, then, there is some promising evidence that interleaving could be an important technique for developing new conceptual knowledge if applied to formal learning settings. Unlike spacing (Cepeda et al., 2006; Son & Simon, 2012), however, there had not (at the time of writing) been a systematic review into the effect to comprehensively explore the effect and its boundary conditions.

The *systematic review* and *meta-analysis* are widely used in education as a means of identifying objectively which techniques are promising as interventions, and under what circumstances they are best used. A systematic review sets out objective criteria for inclusion, allowing the methodology to be replicated, and minimising researcher bias when answering a focused question (Siddaway et al., 2019). Often, papers are weighted in terms of the strength of evidence including the quality of included papers' methodology, such as whether participants were randomly allocated to experimental conditions or not (Higgins & Thomas, 2019; Robinson & Lowe, 2015). They contrast with narrative reviews, broader summaries which tend to be non-statistical. The latter do not always feature a precise research question with inclusion and exclusion criteria (Robinson & Lowe, 2015) and can therefore be more prone to bias in terms of the selection of studies reviewed.

Given the variability in the format and material of interleaving studies, a systematic literature review would be a worthwhile approach to summarising the evidence and gauging the sizes of any effects. This would allow researchers and educators a more nuanced picture of its use than is presently available. For example, is the interleaving of images subject to the same considerations as interleaving of verbal examples, and are there conditions of study that reverse the effects? Such questions need to be answered before interleaving as a whole can be confidently recommended to educators. Such a review would also identify future areas for further research, for example gaps in the literature.

I therefore aimed to carry out such a review, and this can be found in Chapter 4. To briefly pre-empt the main findings, this review found that found that interleaving had a large advantage over blocking for both memory of examples and transfer to novel examples.

2.3 Theoretical Underpinnings

In order to fully understand how the two techniques discussed above can be applied, it is necessary to understand the more basic processes at work when a learner experiences them. In particular, the cognitive processes at work are likely to be critical in understanding how and when these two strategies are likely to be beneficial, and for whom. In addition, metacognitive factors such as awareness of the learning benefits of each technique will potentially constrain their use. These considerations will help to guide the application of the techniques to classroom situations, and provide insight into possible new applications as well as limitations or concerns, and will be explored next.

2.3.2 Working memory.

If any information is taken in, held for a short period of time, and compared to other information or to retrieved ideas from LTM, then this processing must make use of working memory (WM).

There have been multiple models of working memory, but all consider it to be a system responsible for processing (Engle, 2002), or storage (Cowan, 1988), or both (Baddeley, 2012; Baddeley & Hitch, 1974; Daneman & Carpenter, 1980). Most models show multiple components within working memory rather than a single generic store, and attention plays a major role according to most theories (Cowan, 2017).

For example, in the well-known Baddeley model, a component called the *central executive* controls the other parts of the system and allocates the attention needed for complex tasks. The central executive is not limited to a particular modality, and oversees the correct functioning of verbal and visual processing (which are each carried out by separate ‘slave systems’). This model also suggests that different subcomponents of working memory can work largely independently. For example, simple visual and verbal tasks can typically be done simultaneously with relatively little decline in performance (Logie, 2016).

Although the central executive was the least well-explored area of the model as initially outlined (Logie, 2016), it is assumed to contribute to a much-studied set of abilities called executive functions. Broadly defined, executive functions are control processes, and they include the ability to stop and switch task or to monitor one's own progress towards a goal (Diamond, 2013), or to plan and maintain the next step of a complex task – processes which are self-evidently relevant to a large range of classroom activities. These abilities develop slowly over the childhood and teenage years relative to other aspects of cognition (Blakemore & Mills, 2014). As might be expected, they depend upon the development of the brain – in particular the prefrontal cortex (Diamond, 2002).

In spacing, studied items still need to be held in working memory at about the same rate, but this is done on different occasions. While that might appear to have the potential for reducing overload to this limited system, it will be in effect neutral given that the same set of material or a similar task is completed in each spaced/distributed practice session. However, the demand on the central executive may increase for the learner, as there is an increased need for a mental search for previous semantic or episodic memories from LTM.

When it comes to interleaving, comparison and categorisation of similar items is more demanding, because it requires a learner to maintain previously presented items (within a schedule such as ABCABC). This results in a form of n-back task (see Jaeggi et al., 2010) for the learner's working memory as they try to recall whether previous items were the same or different. However, as discussed in Section 2.3, the benefit of interleaving may derive from the way it juxtaposes different items. The reverse to the previous scenario applies, then, if the learner tries to focus on (or passively notices) differences rather than similarities – the demand on working memory is reduced when contrasting items are together rather than separate. As working memory can normally only hold around four discrete chunks of information (Cowan, 1988), it has sufficient capacity to maintain before/after contrasts with relative ease, while contrasts with earlier items in a series would be considerably less salient.

The demands on working memory will also relate to the exact procedure through which information is presented, which in turn depends in the nature of the material. With visual examples, it is possible to present several items simultaneously, allowing similarities and differences to be observed by the learner. With verbal presentation, this would only be possible if the items are very short, for example single words, digits, letters, or letter digit pairs. If a concept is expressed as a whole sentence, then it will be beyond the capacity of verbal working memory to maintain more than one such sentence at once. The sentences would have to be meaningfully interpreted and then compared on a conceptual level; this would be demanding to the central executive and would draw on LTM (for even items that were just seen a matter of minutes ago are registered in LTM, albeit not consolidated enough for permanent storage; see Chapter 1).

Individual differences in WM structures could also be very significant in terms of educational performance. For example, a greater capacity in verbal working could affect a learner's ability to process complex verbal tasks or examples, and to take in new vocabulary. In the interleaving literature, however, Guzman-Munoz (2017), in a visual task, found no significant interaction between the benefits of interleaving and scores on a test of WM. The interleaving advantage can persist even when learning is incidental (Birnbaum, 2014), and is steady across different numbers of categories and examples (see Chapter 3). However, for a statistics task, Sana et al. (2017) found that the benefit of interleaving disappeared for learners (undergraduates) with the highest WM scores, suggesting that for the most demanding learning situations, processing demands may play a role (Sana et al. describe their task as "rule based, with a relatively heavy memory component"; p. 89).

Overall, benefiting from the compare-and-contrast opportunity afforded by an interleaved presentation appears to be well within the capabilities of most learners given that the nature of the task makes contrast easier by putting these items together. And indeed, even pre-school learners – whose working memory capacity is still developing – appear to profit from interleaved presentations (Vlach et al., 2008). A greater factor in individual differences is likely to derive from an individual student's prior knowledge. For example, the difficulty of learning a new concept is in part dependent on how much it conflicts with learners' prior assumptions (Chi et al., 2012), with implications for processing time (see also Chapter 1 for a discussion of prior knowledge and schemas).

Sweller et al. (2011) have argued that because working memory has limited capacity, any non-essential tasks will harm learning, as they will interfere with the processing of information in WM which according to the modal model of memory (Atkinson & Shiffrin, 1968; Murdock, 1967; see Chapter 1) is essential for encoding to LTM. Essentially, the concept of cognitive load implies that the level of detail and complexity of a task should be 'just right', neither under-using or overloading the mind's limited capacity. However, this appears to conflict with certain predictions of the new theory of disuse. Difficulties are often more demanding of limited working memory resources, but have been shown in numerous ways to boost longer-term learning (see Chapter 1). Accordingly, it could be argued that cognitive load theory is too focused on facilitating short-term processing (linking to performance) and insufficiently focused on long-term learning. It also does not take account of the argument that difficulties are desirable if and when the processing that they elicit are consonant with conditions of later use (Bjork & Bjork, 2019; McDaniel & Einstein, 2005).

While reducing the demands on working memory will tend to make a task easier for learners, it is therefore worth asking whether doing so may be undesirable. In the case of spacing, a briefer delay would reduce demand on working memory, but (according to evidence discussed earlier) would also hinder long-term learning. Presenting simultaneous or near-simultaneous examples in an interleaved format is similarly demanding on working memory.

Overall, then, working memory is necessary for any classroom learning task to proceed, but it should be remembered that maintaining information in working memory during learning is necessary but not sufficient. Performance does not equate to learning. Distractions or cognitive load which lead to irrelevant processing will be unhelpful, but simplifying tasks or reducing their subjective demand may not be a good idea if doing so interferes with target desirable difficulties. Instead, educators should aim to build in desirable difficulties that boost storage strength, even when this leads to a counterintuitive sense of difficulty or harms immediate classroom performance.

2.3.3 Transfer of learning.

As discussed earlier, transfer refers to learners' ability not just to retrieve what they have learned, but also to use and apply it when relevant situations arise in the future. This is of great practical importance to educators and their students. For example, if a student has learned about the dangers of sunburn, it is not enough to retain this factual knowledge in long-term memory; they also need to be able to recognise when it will be an issue (for example, when they are outside for a long period of time on a sunny day) and take appropriate action.

Most studies of interleaving involve an element of near transfer; new concepts are taught via the inductive learning of examples, and learners are then shown novel examples and asked to categorised them. In contrast, research into spacing tends to focus on exact repetitions, such as repeated practice of vocabulary or of identical problems, and as a technique it is assumed to mainly boost memory rather than promote transfer. However, Kapler et al. (2015) demonstrated a benefit to spacing for transfer to application problems in science. It seems likely that spaced practice of foundational knowledge and skills would at least facilitate transfer, even if it does not guarantee it, because transfer would be easier if the initial learning was well consolidated (see also Foot-Seymour et al., 2020). This principle was supported in a study of retrieval practice by Butler (2010), where information about bat wings and sonar was both better remembered and more successfully transferred to questions about aircraft design by learners in a retrieval condition than their peers in a re-reading condition (see also the discussion of skills in Chapter 5).

Can students be taught to engage in more transfer? Schooler (1989) argues that cognitive training does not generalise well from one subject area to another, a conclusion that was supported by Simons et al. (2016) in their study of 'brain training' games. Gick and Holyoak (1980) looked at the ability of people to problem solve when given an example that was similar by analogy – a general having to move troops via several roads, and a surgeon who needed to operate on a tumour but couldn't use a single laser. These results suggest that despite the practical importance of transfer, it does not happen spontaneously, even when the connections between problems may seem very obvious to the educator. However, participants who try Gick and Holyoak's task were more likely to make the connection when prompted, and a larger number of examples tended to be beneficial (Bauernschmidt, 2017). This has relevance for interleaving, where multiple examples tend to be made, and learners are (often) expected to induce generalisations across these.

If repeated use of transfer-type strategies become easier with time, there would be a case for a form of '*meta-transfer'* whereby the ability to transfer learning is itself transferred. Some initial evidence on this is provided by Birnbaum (2013), who taught learners a 'compare and contrast' rule on one interleaved context, and measured the extent to which this was used in a different one. However, this study found limited benefit of the strategy across learning situations, mirroring the failure to transfer spontaneously found by Gick and Holyoak (1980).

Transfer also depends on having well-developed schemas. Brown and Kane (1988) found some evidence that children could learn to transfer a biological concept – animals using mimicry as a defence from predators – but this was more successful when understood on a conceptual level rather than via surface examples, suggesting that a degree of expertise is needed (and fitting well with the benefits of consolidating foundational knowledge via spacing and retrieval practice, discussed above).

While further research is needed, these findings together suggest that the techniques discussed in this chapter are not generic 'thinking skills' that can be taught to students and then straightforwardly applied to many situations. However, they may be used to support transfer in any given topic, both by boosting levels of the requisite prior knowledge, and (in the case of interleaving) by directly juxtaposing different examples.

2.3.4 Metacognition, reflection & learner beliefs.

As the research into transfer shows, learners do not always appreciate what they have learned or spontaneously adopt successful study strategies (Koriat, 2000; Kornell & Bjork, 2009; Nelson & Dunlosky, 1991; Pintrich, 2002). It is worth considering, then, what learners may be thinking about when appraising their own learning process, and whether the use of the techniques under discussion and other desirable difficulties are characterised by misconceptions.

When studying material with a view to learning or developing new concepts, learners not only think about the material, but also think about their own thinking and learning, and make predictions (accurate or otherwise) about their future ability to use what has been learned. The processes involved are broadly termed *metacognition*, and can usefully be divided into two main areas (Nelson & Narens, 1994):

- *Metacognitive knowledge*: learned information about strategies that can be brought to bear on future learning.

- *Metacognitive monitoring*: online focus on a particular learning task as it proceeds.

In terms of classroom activities, metacognitive monitoring includes learners noticing errors, keeping track of the steps taken when solving a problem, and judging the progression of one's own learning and understanding, while metacognitive knowledge can include awareness of the benefits of techniques such as spacing and self-testing for revision purposes, or of the typical impact of forgetting.

Metacognitive monitoring depends on executive function, and is therefore difficult for those who are tired or distracted. It is also difficult for younger students because it depends on executive function and thus on the development of the prefrontal cortex as discussed above (see Section 2.3.1). Metacognitive monitoring can also be inaccurate because introspection is a limited way of theorising about thinking. Indeed, leaners are generally poor at judging what they do and do not understand; tasks requiring 'judgements of learning' – a person's subjective anticipation of being able to recall the information at a later point in time – tends to be biased by the ease of retrieval, without making allowances for the context of this retrieval (Benjamin & Bjork, 1996; Kornell & Bjork, 2009).

Metacognitive knowledge includes beliefs about learning and about one's own attainment level. These are also characterised by inaccuracies, perhaps because memory is often counterintuitive in its functioning. Its workings are not easy to figure out even with practice (Bjork, 2011), do not fit with 'common sense' assumptions (Simons & Chabris, 2011), or with the way memory is typically portrayed in the media (Karpicke, 2016).

Multiple laboratory studies of spacing (e.g. Zechmeister & Shaughnessy, 1980) and interleaving (e.g. Kornell & Bjork, 2008; Yan, Bjork & Bjork, 2016) have found that learners mistakenly think that these strategies are less effective. Students also appear to favour sub-optimal learning strategies outside of the laboratory. For example, Hartwig and Dunlosky (2012) found that over 60% out of a sample of 324 students favoured ineffective revision strategies such as re-reading, cramming and underlining/highlighting, and that the choice of such strategies was statistically related to a poorer grade point average. Even the most basic 'common sense' assumptions about learning – such as forgetting occurring over time or practice being associated with improvements – are not matched by actual self-regulated study behaviours; Kornell and Bjork (2009) found that learners tend to exhibit a *stability bias* in terms of memory, assuming that their current ability to remember an item is a good guide to their likelihood of doing so in the future (see Chapter 6 for more on learners' flawed beliefs about effective learning).

Metacognitive monitoring and beliefs can interact, and at times this is also a source of errors. An example comes from a study by Glenberg and Epstein (1987) where students were shown two texts, one about music and one about physics. Learners were experienced in one subject but not in the other. Their judgements of their own understanding of the text proved to be more accurate about the subject that they had *less* knowledge in. In other words, it is easy for a learner to overestimate their understanding based on beliefs about their own attainment, which could include underestimating the difficulty of the task itself.

These common flaws are important because accurate metacognition undoubtedly helps with evaluating one's own work and learning from mistakes, and as such it is widely seen as a potentially powerful education intervention (Education Endowment Foundation [EEF], 2018). One of the most obvious inputs to metacognitive monitoring is corrective feedback, and this appears to be especially helpful if learners attempt to retrieve information first, and are then given feedback (Rawson & Dunlosky, 2007).

Timing also affects the accuracy of student metacognition. A study by Nelson and Dunlosky (1991) compared judgements of learning for word pairs either immediately or after ten minutes. The ten-minute delay greatly improved the accuracy of learner judgements, which was little better than chance level when done immediately. This suggests that as well as its role in learning, spacing may improve metacognition. Indeed, Bahrick and Hall (2005) have suggested that metacognitive appreciation of forgetting via multiple failed attempts to retrieve is one of the primary causes of the spacing effect, although this conflicts with evidence that retrieval practice is more effective when retrieval attempts are successful (Agarwal & Bain, 2019; Racsmány et al., 2020).

2.4 Analysis

Having reviewed the underpinnings of spacing and interleaving on a cognitive and metacognitive level, I now return to the classroom implications of the two techniques, and seek to synthesise and critique each technique in terms of the broader evidence, as well as identifying commonalities, ways that they combine, and priorities for further investigation.

2.4.1 Potential of spacing.

Spacing has considerable potential as an evidence-based teaching strategy. In particular, the technique has a very strong evidence base when it comes to factual information and vocabulary. There is evidence that it can benefit more complex material such as studying texts and maps, though, and given that forgetting is a universal, the benefits of spacing may also be universal. The bulk of the research is laboratory based, but it has begun to be applied to classroom-relevant tasks and timescales, with strong evidence emerging that permanent learning necessitates longer delays. These findings are supported by the biology of long-term memory; LTP appears to require consolidation hours later, and to benefit from the consolidation effect of sleep.

The use of spacing is likely to interact with the meaningful nature of the information, though, as well as with learner experience, and there is as yet little research on how it is modified by individual differences, limiting the extent to which teachers can confidently judge how the technique will affect different learners in their class. Going forward, more information is needed about how exactly it is best used in the classroom, particularly in the context of authentic materials and entire lessons. On a metacognitive level, students are likely to avoid spacing believing it to be harder and therefore less effective, and it is important to understand whether teachers will share this concern.

2.4.3 Potential of interleaving.

Interleaving has been explored in a number of studies, and theories of the effect have also developed rapidly. It has potential to be applied to any science or social science, and probably many more areas, too. There are also many studies that have interleaved maths problems, visual images of artworks, or images of animal species, suggesting applications to a broad range of subjects, especially where categorisation and later transfer is an educational target.

As with spacing, though, there are concerns from the metacognition evidence that students will tend to avoid interleaving, and again it will be important to establish whether teachers feel the same way. Important questions remain about the extent to which interleaving interacts with delays, given that a classroom interleaved presentation (e.g. of a few examples) could in itself be quite brief. This is a question to return to, and as noted earlier, a gap in terms of the lack of a broader review of the technique was identified as a research target for this thesis, and will be addressed in the following chapter (Chapter 3).

2.4.3 Combinations of the techniques.

As the techniques under discussion both relate to the manipulation of item order in the classroom but are not direct alternatives, they can at times be combined. Early evidence conflated interleaving with spacing, but it has now become clear that its benefits depend on comparisons, and that spacing these out can be detrimental. A gap between one example and the next reduces a learner's ability to compare and contrast – a key benefit of interleaving. However, Birnbaum (2013) found that as long as discriminative contrast is not undermined, spacing and interleaving can be combined. This can be achieved by using longer spaces between sets of un-spaced and interleaved examples, as shown in Figure 2.4:

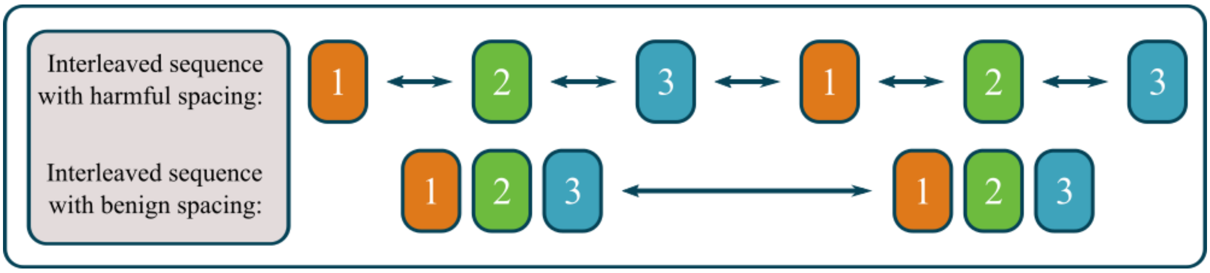

Interleaved sequence with harmful spacing:
1
2
3
1
2
3
Interleaved sequence with benign spacing:
1
2
3
1
2
3

Figure 2.3: Interleaved sequence with spaces. Delays are only problematic when they interfere with discriminative contrast.

In a classroom, a teacher could implement a version of the ‘benign spacing’ condition shown in Figure 2.3 by completing an entire interleaved activity, for example during a particular lesson, and then repeating this or carrying out a similar activity the following day.

Both of the techniques discussed depend on the cognitive apparatus of working memory. While an analysis and review of different contemporary models of working memory is beyond the scope of this chapter (though see Cowan, 2017, for an excellent review), suffice it to say that this thesis will assume that working memory is a multi-component system which includes (at least) a verbal, visual and attentional components. As such, theories which see WM/short-term memory as a simple generic passive store are seen as too simplistic to account for the phenomena under investigation.

The metacognitive evidence explored above suggests that many learners will lack insight into their own learning process, and this could apply to the benefits of any of the three techniques under discussion. This could happen either because key learning process occur without learners’ conscious awareness (for example, neural consolidation during sleep), or because they harm short-term performance on a task, and learners mistakenly view this as something to be avoided in favour of less effective techniques. In metacognitive terms, then, the evidence suggests that many will show flaws in monitoring – failing to notice effective learning when it is happening – and in metacognitive knowledge of what a good learning strategy looks like.

What’s more, most learners won’t realise their own errors, or spontaneously work to correct them. There is potentially a very important role for the teacher to guide their students towards more effective learning behaviours, therefore, and an important question to ask is whether teachers themselves have sufficient expertise in memory to provide this guidance and to implement spacing and interleaving where appropriate. However, can we be sure that teachers know any better? I will return to this question in the final section of this chapter (2.6), which explores the implications of the three techniques for professional learning.

First, though, it is important to be cognisant of the limitations of the evidence discussed so far, and it is to these limitations that I now turn.

2.5 Criticisms of the Techniques Discussed

The analysis above has identified that interleaving and spacing are techniques with considerable potential and low cost to use. They are desirable difficulties and stand to benefit student learning. An obvious interim conclusion, then, is that teachers should engage more with these techniques, adding them to their professional repertoire and boosting their students' attainment via their frequent application to classroom work, alongside other evidence-based desirable difficulties.

However, it could reasonably be argued that what has been presented so far takes an overly sympathetic view of desirable difficulties and their supporting research, and I will therefore now consider the counter-evidence and contrary arguments.

2.5.1 Arguments against evidence-based teaching practice.

A broad and philosophical complaint applying to most desirable difficulties is that teaching should not focus on adopting evidence-based practice, an approach to education which borrows from evidence-based medicine. Learning, it can be argued, is not really like curing an illness – it is cumulative, has no clearly defined end point, and there are important subtleties such as how well it can be transferred to new situations. What's more, just as medical treatments can have unwanted side effects, so can educational ones (Zhao, 2017). Simply measuring the learning *benefits* – for example, better recall following spacing or retrieval practice – might neglect consideration of certain *costs*, for example to motivation.

An evidence-based approach to education (sometimes referred to as a 'what works' approach) is well exemplified in EEF's (2018) 'Teaching and Learning Toolkit', in which interventions such as feedback and phonics are ranked by cost and impact. A second issue is that this can be seen as over-simplistic; particularly as what 'works' for one group might not work for all. To take one example, Kalyuga (2007) has described the 'expertise reversal effect' whereby tasks that are effective with beginners become ineffective or at least inefficient when used with more advanced learners. Another example is that homework appears to be more effective for secondary students than for primary (Cooper et al., 2006). While such issues don't rule out the application of evidence-based practice, they do suggest caution over one-size-fits-all solutions, and this consideration must be borne in mind when determining when and how techniques such as spacing and interleaving should be used.

Thirdly, even if we accept that a cautious use of evidence is worthwhile, there can be concerns about the appropriateness of some of the methodologies used to gather this evidence. The randomised controlled trial (RCT) appears at or near the top of many hierarchies of evidence (e.g. Evans, 2003), and in medicine it has been described as the *gold standard* in that it is large scale and makes use of a control group and random allocation of participants to conditions (Meldrum, 2000). However, RCTs can lack nuance and fail to account for the many variables that affect education (Sullivan, 2011). The concerns raised above about individual variation suggests that it could be problematic if findings of such studies are simplistically interpreted as showing the right or wrong way to teach, and then applied more widely. As such, there is a case for preferring more interpretive methodologies, less influenced by a positivist epistemology (see Boyle, 2012; see also Chapter 4). There are also concerns about the feasibility of conducting RCTs in schools due to the training and resources required (Robson, 2002). Therefore, despite their high scientific standards in helping to establish cause–and–effect at a population level, applying RCT findings to the classroom is therefore neither a straightforward nor uncontroversial endeavour.

So far, however, the key evidence base supporting desirable difficulties comes not from RCTs but from laboratory studies of learning and memory. Such studies are often carried out with university students as participants, and should be generalised to school pupils with caution – and preferably only after replicating the key findings in the school classroom – given the major differences in the structure of learning experiences, and in students' developmental stage and academic attainment level between secondary and tertiary education. Chapter 4 of this thesis will return to methodological issues, and try to establish more satisfactory options.[4]

It could also be argued that research evidence is too granular and clear cut to usefully inform the complex, multi-faceted nature of teaching practice. Wiliam (2019), for example, argues that classrooms are too complicated for research to be able to guide teachers as to what they should do at any given time.

However – while I do not wish to dismiss the idea that teaching is complex and nuanced, or even to critique this argument in general terms – when it comes to the three techniques discussed in this chapter, there tends to be a relatively straightforward choice facing the teacher. For example, examples appear to be better learned if presented in an interleaved rather than a blocked fashion. This finding cannot inform all aspects of a teacher's planning or their classroom performance, but evidence that interleaving is superior to blocking for a particular type of material and a particular type of learner is certainly relevant and applicable. A similar argument can be made for spacing.

[4] I previously made these arguments in a blog post, "what is evidence-based education?" (see Appendix 4).

Finally, educators may be concerned that the promotion of such evidence-based practice undermines teacher autonomy, and forms part of a broader movement of de-professionalisation and creeping centralised control (see Sachs, 2016). If we viewing teachers as the primary decision-makers in the classroom, then imposing a more evidence-based approach upon them could indeed be seen controlling, and may risk undermining professional agency in favour of centrally-approved approaches to pedagogy. As Biesta and colleagues have put it:

"*Some see teacher agency as a weakness within the operation of schools and seek to replace it with evidence-based and data-driven approaches, whereas others argue that because of the complexities of situated educational practices, teacher agency is an indispensable element of good and meaningful education*." (Biesta et al., 2010, p. 624).

However, as we have seen, the precise application of evidence-based techniques is highly context specific. The state of a student's existing knowledge has implications for how and when techniques should be used, and this existing knowledge is specific to the learner and the material. The use of evidence-based practices therefore depends on judgements made in a particular classroom which will vary across different students and from day to day as learning progresses, even if the underlying principles do hold more generally. Professional understanding of learning and memory processes could be seen, then, as a part of the teacher's professional toolkit, rather than as something that can be imposed from above (Firth, 2017).

Indeed, as Biesta et al. (2010) conclude, agency is best not seen as a purely individual process, but depends upon developing a collective professional understanding among staff, sited within school and national policies. Despite the fact that responsibility for learning processes and outcomes is a key part of a teacher's professional role, fully realising this aspect of teacher agency can be held back by “the absence of a robust professional discourse about teaching and education more generally” (Salomon, 1992, p. 638).

2.5.2 Evaluation of evidence on metacognition.

Moving on to the evidence on metacognition, in section 2.3 I highlighted some potential conflicts between desirable difficulties and accurate metacognitive monitoring and/or knowledge. However, a limitation with this analysis is that it relies on a mainstream research-based view of metacognition, which tends to view the processes in highly individualistic terms. It is important to bear in mind that adolescent decision-making is strongly influenced by peers (Steinberg & Monahan, 2007). Indeed, the brain areas underlying metacognition undergo major reorganisation in early adolescence, resulting in important new opportunities for learning (Blakemore & Mills, 2014), but also leading to a tendency for students to identify with peer groups which have strongly anti-education social norms in an attempt to manage their social uncertainty (Cruwys et al., 2017). This could impact on their choice of learning strategy, or, indeed, on whether they choose to study at all. The social context, then, should be taken into account when analysing students’ choices to engage in interleaving or spacing. Viewing it purely in terms of their awareness of the strategies or avoidance of desirable difficulties is limited.

More specifically, engagement with evidence-based learning strategies may depend on how these are positioned with respect to learners’ identity. Identity and social norms can have a major impact on educational behaviour, such as discipline and values (e.g. Reynolds et al., 2015), how smoothly students transition between academic levels (Cruwys et al., 2017), and how highly they rate a course (Sonnenberg, 2017). Student engagement with evidence-based learning techniques could depend on how they socially categorise themselves.

A similar argument could be made when it comes to teacher metacognition and identity. Information about effective learning techniques is not enough; if such evidence-based strategies are to be adopted, teachers need to want to use them, and will preferably see application of these techniques as the norm in their profession. The position of research as part of their professional identity, as described by self-categorisation theory, is likely to play a role in their behaviour (Mavor et al., 2017; Tajfel & Turner, 1982). To facilitate this, a willingness to engage with and critique the continually developing evidence base relating to successful learning would need to become a part of their professional identity.

Increasingly, this kind of research engagement has been prescribed at a policy level, with research engagement seen as one of the standards of professionalism teaching (e.g. GTCS, 2012). A major question going forward, then, is whether teachers are likely to engage in the specific techniques under discussion, and – if not – what might motivate them to do so. This is addressed next.

2.6 Implications for Professional Learning

Having introduced, analysed and critiqued the three evidence-based strategies relating to item order in the classroom from the perspective of a teacher-researcher, what implications are their implications for teachers' knowledge of memory? And what barriers are there to teachers' professional engagement with and use of the techniques?

The techniques outlined above have been presented as evidence-based, and have the potential to improve teachers' classroom effectiveness, impacting on student attainment. However, viewing professional learning as a process of improving teacher *knowledge* ignores the complexity of learning to apply this knowledge. Simply giving teachers information about evidence-based techniques is likely to have limited impact. As Wiliam (2010) has stated, professional development based on factual input alone has resulted in “*teachers who are more knowledgeable, but no more effective in practice*” (p. 4).

Instead, teachers need opportunities to put their new learning to work in the classroom, perhaps gradually and with some scaffolding of the process. They may also be able to acquire naturally occurring classroom evidence, for example exam scores, to gain a further insight into the learning process (Firth, 2019a; Wall & Hall, 2019) and augment this professional learning.

At times, though, techniques like interleaving and spacing will conflict with teachers’ assumptions. It seems at least possible that teachers, like their students, may harbour beliefs about the learning processes in their classroom (or about psychological processes in general) which fail to match the evidence. That is to say, their metacognitive knowledge may be incomplete or flawed. After all, misconceptions about the mind and learning are widespread among adults. Furnham (2018) questioned participants aged 19 to 66 about popular myths from developmental and neuropsychology, and found that a large number were rated as definitely or partly true, findings which held across all demographic groups. Myths under investigation included “Most babies can learn to read with the right learning program” and “People suffering from amnesia typically cannot recall their own name or identity”. Focusing on memory, Guilmette and Paglia (2004, cited by Kornell, 2015) found that 41% of respondents agreed with the statement, “sometimes a second blow to the head can help a person remember things that were forgotten [as a result of a first blow to the head].” Simons and Chabris (2011; 2012) also found large differences between beliefs about memory among the general public and those of memory researchers (see also Chapter 4).

Clearly, professionals differ from members of the public given that they have received specific training to carry out their role, and it might therefore be assumed that teachers will not be subject to these flawed beliefs. However, there is considerable variability in what is taught to new teachers (Carter, 2015), meaning that there is no guarantee that what new teachers have been told about learning, for example by tutors, mentors and colleagues, will concord with the scientific consensus. In addition, misconceptions about memory have been found even among those for whom it is clearly relevant to their professional role, for example among judges and among psychologists who act as expert witnesses to legal trials (Magnussen & Melinder, 2012).

In the teaching profession specifically, the popularity of 'neuromyths' demonstrates that popular beliefs among practitioners can conflict with the scientific consensus. As discussed by Howard-Jones (2014), over 97% or teachers surveyed in Turkey subscribed to the idea that people learn better when taught according to their preferred 'learning style', while 91% of UK teachers agreed that differences between learners can be explained by their being 'left brained' or 'right brained'. Morehead et al. (2016) found that while teachers held slightly more accurate views of learning than did their students, the difference was small and the overall pattern was broadly similar; they also found no correlation between teachers' endorsement of myths such as learning styles and their number of years of teaching experience.

Indeed, despite experience being near synonymous with skill in some contexts, there is not a strong statistical relationship between a teacher's years of classroom experience and student outcomes in their classes (Hanushek & Rivkin, 2010). This could be in part because improvements tend to plateau after a small number of years (Hood, 2016; Rivkin et al., 2005), with further experience beyond this point having a negligible impact on classroom performance. This fits with what is known from other domains; as Ericsson et al. (2007) note in relation to sports expertise, practice alone can lead to the consolidation and automation of mediocre performance. Additionally, the sources of gains in effectiveness made in the early years is unclear; they could potentially link to factors such as an improved repertoire of classroom tasks, rather than a better understanding of learning and memory.

On the basis of this analysis, it is reasonable to question the extent to which teachers will make use of the techniques discussed in this chapter, either spontaneously or even when informed about them. However, beyond the study of neuromyths discussed above, relatively little is known about what teachers believe about learning and memory. Studies of metacognitive knowledge have tended to focus on the students themselves, neglecting the issue that it is often teachers who are the primary decision makers about what is studied and when. They also face challenges in accounting for desirable difficulties; teachers may mistakenly think that performance equates to learning, though again, the studies of decision making and choices over what to study have tend to focus on students and on the process of independent study and revision (perhaps because such work has tended to use university student samples, e.g. Kornell & Bjork, 2007; Geller et al., 2018; Hartwig & Dunlosky, 2012).

It is also important to ask whether teachers have access to information about desirable difficulties in the first place. One Dutch study found that spacing and retrieval practice are largely absent from textbooks aimed at new teachers (Surma et al., 2018). Governmental bodies could in principle provide good-quality generic guidance but experience suggests that they often don't; instead, their use of evidence displays considerable distortions based on ideology and agenda (Alexander, 2014). More positively, however, there is a developing grassroots movement among teachers seeking to engage with research-based practice; the teacher-led *researchED* conferences are important means of sharing knowledge, as are social media and practitioner blogs. UK-wide publications such as the *Times Educational Supplement* and *Schools Week* are increasingly featuring research evidence based around memory, too (e.g. Enser, 2019; Scutt, 2020).

A major priority going forward, therefore, will be to survey the beliefs about memory about the teaching profession. What do they think about how memory works, and do they endorse the desirable difficulties that their students apparently avoid? How would they choose to schedule learning if presented with a set of alternatives, such as interleaving vs. blocking? And do such beliefs change in line with experience as professionals come to have a better understanding of how their students learn? These questions are explored in more detail in Chapter 6.

2.7 Concluding Comments

The evidence presented here explores the benefits of spacing and interleaving – both easy, quick and cheap techniques to apply. They are evidence-based and have been demonstrated with a wide range of subject matter over many decades. This doesn't mean that they should be used in every situation, but there is a strong argument that each should form part of a teacher's professional repertoire.

In particular, spacing and interleaving can both be seen as evidence-based techniques with broad applicability to students of all ages, and even to professional learning. Notwithstanding some limitations in the evidence, they can be applied in a range of situations. Spacing in particular stands to boost memory for facts, while interleaving can help with concept knowledge and transfer. Some field research of spacing has been done, but overall the research literature is highly lab-based, lacks work on curriculum-relevant materials, and does not fully explore how the techniques interact in the context of real classrooms. An aim for the current research will be to extend the literature in these areas.

Metacognitive research strongly suggests that we cannot rely on either learners or their teachers to have accurate beliefs about the optimal timing or order of learning tasks. Those involved in teacher preparation should be aware of the many myths and misconceptions about learning that are present among the general population, and that experience alone appears to be insufficient to counter this problem. In addition, teachers may not always have easy access to accurate, up-to-date evidence about the psychology of learning and memory.

2.8 Interim Summary and Outline of Thesis

In this chapter so far I have introduced the theoretical considerations which can guide the application of spacing and interleaving to teaching practice from the perspective of a teacher-researcher. I have briefly reviewed and analysed the current evidence on timing-related teaching interventions themselves, considered the metacognitive barriers that may limit their application by the teaching profession, and addressed certain professional learning implications that stem from these points.

The next sections will proceed as follows:

- Chapter 3 presents a systematic review focusing on interleaving. This piece of work was pre-registered with PROSPERO (the protocol is shown in Appendix 2, and a short article based around the pre-registration is reproduced in Appendix 3). The review consists of both a narrative section and a meta-analysis, with the former addressing boundary conditions and variables that impact the interleaving effect, and the latter synthesising a set of 28 studies for which the methodology was consistent enough to be comparable. This review demonstrated large standardised effect sizes of up to +0.66 but also confirmed that the existing research largely consists of laboratory studies, generally using novel stimuli that are very brief and which do not pertain to the participants' specific course(s) of study.

- Chapter 4 reviews the methodology options for the thesis as a whole, and identifies a research plan. It will consider the unique difficulties that pertain to an area of study that is essentially counterintuitive, as well as the factors that may bias self-reports when it comes to teachers' assessment of their own teaching practice. Factors that should be accounted for are detailed, as are the difficulties in gaining good classroom-based samples of both students and teachers, and the ethical issues that apply. Some of the approaches that have been tried in similar studies are reviewed. Drawing on all of these issues, this chapter sets out a plan for the empirical aspects of the thesis. A key conclusion is that the available research is largely lab-based, and more needs to be done to develop procedures for field experiments.

- Given that the systematic review (Chapter 3) found that the body of work in this area is heavily lab-based and the methodology evaluation recommended field experimentation with authentic materials, **Chapter 5** presents two exploratory studies which investigated the application of interleaving to classroom concepts. Study 1 was a small-scale pilot which tested both spacing and interleaving in the context of a school-relevant task on types of phobias, using a cohort of teenaged pupils. Study 2 was a larger-scale school-based study, testing the learning of key exam skills (explanation, analysis and evaluation) with current school pupils ahead of one of their final exams.

- A major issue that had arisen in the research process was the concern that even in as far as interleaving and other desirable difficulties are evidence based and justified there is the potential for professionals to mistakenly think that alternative options are superior. **Chapter 6** presents two further studies, both of which investigated teachers' beliefs about learning. Study 3 was a broad-ranging pilot study used to identify practitioners' beliefs about memory and learning in general. This identified that some of the areas which are most out of line with the scientific consensus are beliefs about spacing, interleaving and retrieval practice. Study 4 therefore focused on these in particular, and used a vignette-based methodology to ask pre-service teachers and practicing teachers to judge a set of nine scenarios in terms of which would lead to more successful learning.

Finally, **Chapter 7** reviews the findings of the research, explores some of the implications that the findings have for practice, and presents practical recommendations based on the work of the thesis as a whole.

3

Systematic Review of Interleaving as a Learning Strategy

3.1 Introduction

3.1.1 Background to the ‘interleaving effect’.

In this chapter I present the findings of a systematic review into interleaving as a learning strategy. As discussed in Chapter 2, interleaving means varying the order of a set of tasks or examples, whereby each item is immediately followed and preceded by an example of a different category/concept rather than appearing in blocks of the same type of item repeatedly (which is termed a ‘blocked’ arrangement). It could arise due to a randomisation or ‘shuffling’ of the order of items or a more deliberate alternation of items. For example, if presenting example paintings by each of three artists (paintings 1 and 2 by Smith, Jones and Rigg respectively), learners could be provided with examples from each artist in sequence or these could be interleaved, as shown in Figure 3.1.

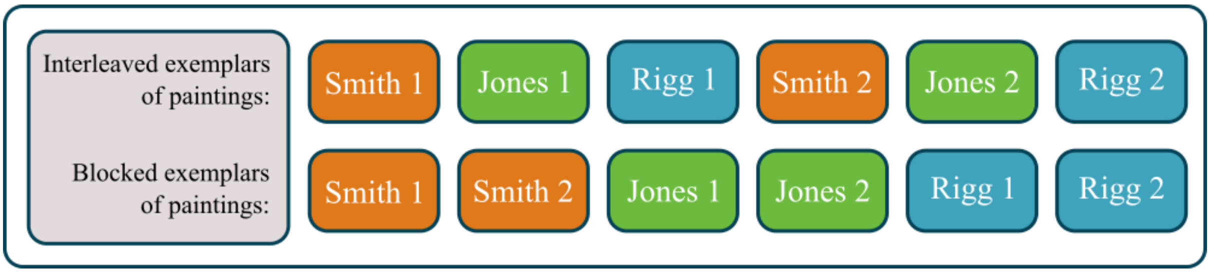


Figure 3.1: Example of interleaved versus blocked sequences

Presenting examples to learners facilitates inductive learning — a form of learning where the learner develops concepts gradually through exposure and experience. Such learning plays a role in numerous aspects of education and everyday life. For example, when people learn to distinguish different species of tree, they may do so by seeing multiple examples in their surroundings, and being told or otherwise finding out what each one is. With time, they form a schema for each tree species, allowing them to independently categorize previously-unseen examples as either belonging to the category or not, even though each specimen that they see is slightly different from those seen before. The learner thereby develops the capacity to transfer their learning to new examples.

Early research in cognitive psychology assumed that viewing examples together in blocks would be beneficial to the process of forming new categories, while spacing them out over time — with unrelated examples in-between — would be harmful (Elio & Anderson, 1981). However, when Kornell and Bjork (2008) put this to the test, the findings ran counter to this assumption. In their seminal study, the researchers presented learners with artwork in blocked or interleaved formats, in a similar way to that suggested in Example 1, above. In their study they presented the work of twelve obscure modern artists (six paintings by each), and then tested participants' ability to identify the artist who had painted novel example paintings. The study had two key findings; firstly, that interleaving was superior to blocking in this test. Secondly, participants' metacognitive awareness of their learning was faulty, as they tended to believe incorrectly that they had learned better via blocking. Both findings have since been replicated multiple times, and extended to other domains such as the learning of science concepts (e.g. Rawson et al., 2015; Eglington & Kang, 2017).

Outside of inductive learning, interleaving has a longer history. Its benefits were discovered by William Battig in the 1960s via his research into the learning of word pairs, and he believed that the interference caused by interleaving can make learning more resilient. As such, Battig (1979) tended to describe the effect in terms of 'contextual interference'. This term is still widely used in the domain of motor learning; numerous studies of contextual interference have shown an advantage of mixed/interleaved physical practice over blocking (e.g. Shea & Morgan, 1979).

3.1.2 Explanations of the interleaving effect.

At first, Kornell and Bjork and other researchers assumed that these findings were due to the spacing effect. Spacing is a memory phenomenon whereby delaying a repetition or practice session leads to items being better remembered, compared with practicing sooner (see Cepeda et al., 2006, for a review of the effect).

Interleaving examples of a concept with contrasting examples will inevitably lead to a degree of spacing. For instance, in Example 1 above, each painting by Smith is separated by two contrasting examples, leading to a slight time delay between the presentation of one example and the next. However, interleaving has now come to be distinguished from the spacing effect. Several studies (e.g. Birnbaum et al., 2013; Kang & Pashler, 2012; Taylor & Rohrer; 2010) have kept the level of spacing constant in interleaved vs. blocked conditions by inserting filler items, such as trivia questions or cartoons.[5] These studies concluded that spacing is not the cause of the effect; indeed, spacing appears to be unhelpful at times, with a delay making it harder to contrast items. From around this time, the effect began to be described in terms of 'interleaving' (see Zulkiply & Burt, 2013a; Birnbaum et al., 2013).

If the benefit of interleaved presentations is not due to spacing, then why does it occur? The attention-attenuation hypothesis (Wahlheim et al., 2011) suggests that blocking is inferior because attention dwindles when learners see repeated examples of the same type; a variation of an account that has also been used to explain spacing (Dempster, 1989) and primacy effects

[5] Please note that in experiments on spacing, the comparison condition is usually referred to as 'massing', implying that items are close together in time. This contrasts with 'blocking', which refers to sets of items in a consecutive, non-interleaved order.

(Tulving, 2008). Evidence supporting the idea comes from Metcalfe and Xu (2016), who presented paintings in a similar way to the Kornell-Bjork artist paradigm along with a mind-wandering probe. Mind-wandering was found to be at a higher level during the blocked condition compared to the interleaving condition. Wahlheim et al. (2011) assessed performance across six exemplars in a set, and found that although the probability of correct classification remained fairly steady, there was some evidence of reduced attention being paid to later items in the blocked conditions.

Repetition of the same type of item could also lead to a sense of reduced demand, perhaps explaining learners' false perception that blocking is superior to interleaving. Bjork (2018) warns that difficulties such as spacing and variation are desirable but often avoided for this reason, and exhorts learners to be "suspicious of a sense of ease" (p. 146).

The classroom clearly differs from an experimental procedure, but it is nevertheless likely that school pupils would experience mind-wandering if presented with multiple examples of a category in succession. Increased mind-wandering tends to be found when a task requires some effort but is undemanding (Baird et al., 2012), and it is plausible that it may be reduced due to a higher subjective sense of difficulty if category exemplars were interleaved.

The discrimination-contrast hypothesis, proposed by Kang and Pashler (2012), is another candidate explanation of the interleaving effect, and is based around the increased opportunities to compare and contrast exemplars if they are seen side by side or adjacent in a series. Interleaved presentations may make it easier to notice differences because contrasting examples are experienced together. If exemplars are blocked, such contrasts are only salient at the beginning and end of each block, and key differences (which the learner might not even realize are important on first viewing) would need to be retained in memory.

The discrimination-contrast hypothesis is supported by evidence which suggests that manipulating an interleaved list to make contrast more difficult tends to reduce or eliminate the benefit of interleaving. In particular, spacing the items out — a feature which leads to improved learning in most situations — appears to be harmful to interleaving. That is, these two phenomena are not additive in their effects; as Birnbaum and colleagues put it, “two desirable difficulties are not always more desirable than one” (Birnbaum et al., 2013, p. 401). This is more easily explained by discrimination-contrast than by explanations of interleaving that focus on attention. It also helps to explain why studies which have interleaved unrelated items have not found this to be beneficial, such as when Hausman and Kornell (2014) mixed science terms with foreign language vocabulary. The hypothesis has been endorsed by numerous authors; Zulkiply and Burt (2013a) went as far as to call interleaving “the compare and contrast effect” (p. 18).

Perhaps surprisingly, relatively few studies have shown exemplars simultaneously in a set rather than in series. One study to do so was conducted by Wahlheim et al. (2011), who found that this increased the benefit of interleaving slightly (and interestingly, their research therefore supported both explanations discussed so far). Presumably in other cases, a learner’s working memory retains previous examples for long enough for contrast to take place. Problematically for this account, recent evidence linking working memory capacity and the interleaving advantage has been mixed (e.g. Guzman-Munoz, 2017; Sana et al., 2018). This issue is discussed further below (see Section 3.3, Narrative Synthesis).

As Goldstone (1996) notes, it may be difficult to learn a new category either because of high between-category similarity (items from two different categories are very alike) or low within-category similarity (items in a particular category are very diverse). For example, an artwork may be hard to categorize as the work of Smith rather than Jones if Smith’s work is generally very similar to Jones’ (high between-category similarity), or if Smith’s body of work as a whole is very variable in style, making it unclear if a particular painting is likely to have been done by Smith or by some else (low within-category similarity). To interleave examples in the case of low within-category similarity could actually be unhelpful, making it hard for learners to notice that a set of examples can all be grouped together. In this case, an argument can be made for blocked learning (this could also be seen as a form of reverse interleaving, with diverse category members being interleaved and compared). This idea was supported by the findings of Carvalho and Goldstone (2014a) who found interleaving to be helpful if exemplars were similar and therefore easily confused, but blocking to be preferable if items were more distinct.

Carvalho and Goldstone (2014a) explain this finding in terms of attention, but rather than focusing on the amount of attention, they argue that the key issue is what is attended to by the learner:

> "*If the previous trial consisted of an object in one category and the current trial consists of another object in a different category, participants' attention will be directed toward the differences between the two objects, by comparing the current object to the previous one (or their recollection, in the case of successive presentations). Conversely, if the two objects come from the same category, learners will attend to similarities between the objects*" (p. 493).

This view is commonly referred to as the attentional-bias hypothesis. In principle, it is compatible with the other two candidate explanations described so far, but has different implications for education – the similarity of examples may be at least as important as the order in which they are presented.

3.1.3 Applications of the effect.

Interleaving is an area of educational importance, given how critical it is for learners to take in new concepts and to be able to transfer this learning to new situations, and it therefore important to assess its effectiveness for school contexts. It is increasingly recommended as part of an evidence-based approach to effective teaching practices (e.g. Benassi et al., 2014; Kang, 2016). Nevertheless, it remains unclear how consistent the supporting evidence is, whether it applies to all learners and tasks, and how strong an effect interleaving has overall. I aim to provide a review of these issues.

On the basis of what has been said so far, interleaving is likely to be beneficial when it allows learners to contrast different problems or examples, or prompts them to pay more attention (to mind-wander less), or both (perhaps by paying more attention to key differences). It is less likely to be useful for unrelated items. It is also unlikely to be useful to interleave very large chunks/sections of classroom activities, as such a format would reduce the salience of useful contrasts. I would also like to assess the role of the timing of learning in the research studies reviewed, whether the benefits of interleaving (if established) apply to short or longer items, or both, and whether they are affected by a delay. More broadly, it will be useful to review the type of items (examples, tasks or images) that have been used in all of the relevant research, and to make some evidence-based proposals for additional domains that are likely to benefit.

There are differences between spacing and interleaving in terms of how materials and follow-up tests are timed. Spacing is often scheduled over days or weeks, such as from one lesson to the next, with retention tested after many weeks. To my knowledge at the outset of this review, interleaving tends to be implemented and measured within a single session, as was done in Kornell and Bjork's experiment, with a minority of studies testing retention after a small number of days. It will be useful to review what effect a delay has on the effect, and to consider what implications this might have for applying interleaving to classroom practice or revision work. These points will be returned to in Section 3.5 (Discussion).

3.1.4 Rationale and predictions.

Given what has been said about the potential applications of interleaving to the classroom, it is vital that researchers clarify the size of any advantage that interleaving has over blocking.

Transfer in learning can be seen as a problem-solving process, requiring organisation and categorisation of novel stimuli. It is highly relevant to many learning situations. For example, most classroom tasks include items which are previously unseen by the learner, and which they must categorize on the basis of their existing knowledge. School exams, likewise, require transfer of past learning to previously-unseen exemplars, with harmful consequences if questions are wrongly categorised.

For simplicity in the remainder of this review, I will refer to tasks which involve classifying novel items as ‘transfer’ and tasks which involve classifying older items as ‘memory’, although I recognise that both processes draw on long-term memory to an extent. Use of the term ‘transfer’ for novel items is in line with Carvalho and Goldstone’s research in this review, as well as older work by J. R. Anderson and others (e.g. Elio & Anderson, 1981).

Transfer, and the inherent linking of new information to existing categories, is inherently difficult for learners (Salomon & Perkins, 1989), although the level of difficulty depends upon multiple factors such as the time delay or the similarity of the learning context (see Barnett & Ceci, 2002). According to transfer-appropriate processing principle, difficulty is reduced if there is greater overlap of the processes during initial learning and later action.

While this may seem to be advisable for educators, Schmidt and Bjork (1992) proposed that variable practice is a *desirable* difficulty, on the basis that examples in real life tend to likewise be variable rather than categorized, and variability during learning therefore contributes to transfer-appropriate processing. More broadly, they argue that many features which degrade performance during practice tasks will improve later test performance (see also Bjork, 2018). It is therefore to be expected that interleaving will benefit transfer particularly, and may have less of a benefit for memory (retrieval and categorising of previously-viewed exemplars). This review will attempt to distinguish between the effects of interleaving on memory and transfer.

Beyond the general issue of memory and transfer, there is also a need for a more complete understanding of how interleaving can be used in educational contexts, as discussed in Chapter 2. In particular, the materials used in the research are of great interest. As discussed above, some earlier studies of interleaving made use of art images, while other studies have used science-related tasks. This review provides an opportunity to compare the different types of tasks that have been used, and to determine whether they are differentially affected by interleaving. Given that interleaving seems to relate mainly to inductive pattern learning, and that “interleaving may discourage rule use (perhaps by introducing a working memory load)” according to Noh et al. (2016, p. 24), I would expect an advantage where materials include abstract patterns. In contrast, examples where a simple, deductive rule could be applied (e.g. insects have six legs, spiders have eight legs) would not require inductive learning via the contrast of multiple examples. For this reason, science materials may see less of a benefit than art materials.

As interleaving appears to boost category learning, it seems possible that increasing the number of categories would make a task more difficult (by increasing the number of distractor categories for any given item), while increasing the number of exemplars would make it easier (by providing a learner with more experience of a given group of categories). Prior research suggests that in general, a greater number of examples can lead to superior transfer when learning an abstract rule (Gick & Holyoak, 1980). On the other hand, a higher number may lead to mind-wandering or reduced attention as discussed above. It will be useful to see whether category and exemplar number interact statistically; if they do, it may be possible to formulate a recommendation for an optimal category number and size in classroom settings. The order in which exemplars are presented and learners' working memory capacity may also have an effect.

These points led to the following research questions and predictions:

(*a) What factors and boundary conditions impinge on the process of learning tasks that use interleaving?*

At this stage I can only predict that such factors will exist. Factors and boundary conditions relating to the interleaving process will be explored in a narrative synthesis.

(b) Does interleaving have a larger effect on learning than blocking?

It is expected that interleaving will be superior to blocking. A systematic review of the literature will allow us to be sure whether well-known studies such as the work of Kornell and Bjork (2008) are representative of the research field as a whole. A meta-analysis will determine the pooled effect size of any interleaving advantage.

(c) Does interleaving have a larger effect on learning in the context of tasks which require transfer than with tasks which require memory?

Given that interleaving differs from spacing and is more relevant to categorisation than to memorisation, I predict that it will have a greater impact on transfer to novel items than on a learner's ability to correctly classify previously-studied items.

(d) Does interleaving have a larger effect on learning with art materials than with science materials?

On the basis that interleaving benefits inductive learning but rule-based learning is more efficient if a rule is available to the participant, I predict that science-type tasks will show a smaller interleaving benefit.

(e) Does set size of categories and exemplars have an impact on performance?

I further predict that the number of categories and examples will affect performance. An optimal number of examples may emerge from the analysis of the data, but as yet I cannot make a specific prediction about what this will be.

3.2 Method

3.2.1 Search strategy.

The search strategy and protocol were pre-registered with PROSPERO, and this protocol was published prior to the completion of the review (Firth et al., 2019; the full protocol can be seen in Appendix 2). I carried out database searching using the PsycINFO, Web of Science, BEI, AEI and ERIC databases. Database journal categories were used to exclude items from irrelevant domains such as electronic engineering, and (where available) to specify the age of samples, according to the inclusion criteria below.

Search terms focused on the research variable interleaving, with possible synonyms, and the outcome variable learning/conceptual knowledge or induction, as shown in Table 3.1

(each row represents an ‘OR’ function). Searches were constrained to records from 2008–June 2018.

Table 3.1: Database Search Terms

interleav*	**AND**	learning
shuffl*		“conceptual knowledge”
“contextual interference”		inducti*
intermix*		

In addition, hand searching of existing narrative reviews/academic book chapters by Rohrer (2012), Carvalho and Goldstone (2015a), and Kang (2016) was conducted.

This search yielded an initial 683 studies, which were then subject to reading of abstracts or full text (see Appendix 5), in order to apply inclusion and external criteria.

3.2.2 Inclusion and exclusion criteria.

The inclusion criteria were designed to include studies that showed the effect of interleaving on memory in contexts that are relevant to education-based conceptual learning. Studies were included if:

- their participants were aged 13–65;
- they had recruited a typically (or assumed typically) developing sample;
- an experimental or quasi-experimental design was used (in order to keep the scope of the research manageable, and because these predominate in the

literature; any less common methodology would also be impossible to compare via meta-analysis);

- the study collected primary data; one or more independent research variables related to interleaved learning in an educationally-relevant context.

Studies were excluded if:

- they were neurological/fMRI-based studies;
- outcome variables did not directly relate to concept learning.

Decisions were made based on titles, abstracts, and where appropriate full text reading. Inclusions were cross-checked and discussed within the research team, and discrepancies solved via discussion. I focused on studies from 2008 onwards because during that year, Kornell and Bjork's seminal study of the interleaving of artwork was released, stimulating numerous follow up studies. This date also ensured that the studies reviewed circumscribed approximately a decade of recent research.

The main reasons for excluding studies from the original search was due to a focus on issues unrelated to concept learning, in particular studies of verbal learning (for example, modern language vocabulary), perceptual learning (for example, learning of tones), or motor learning. Other studies were excluded because they looked solely at learners' beliefs about interleaving, and did not gather data relating to the effectiveness of the strategy itself.

This search strategy yielded a total of 43 studies suitable for a full text search. The search process is summarised in Figure 3.2, below:

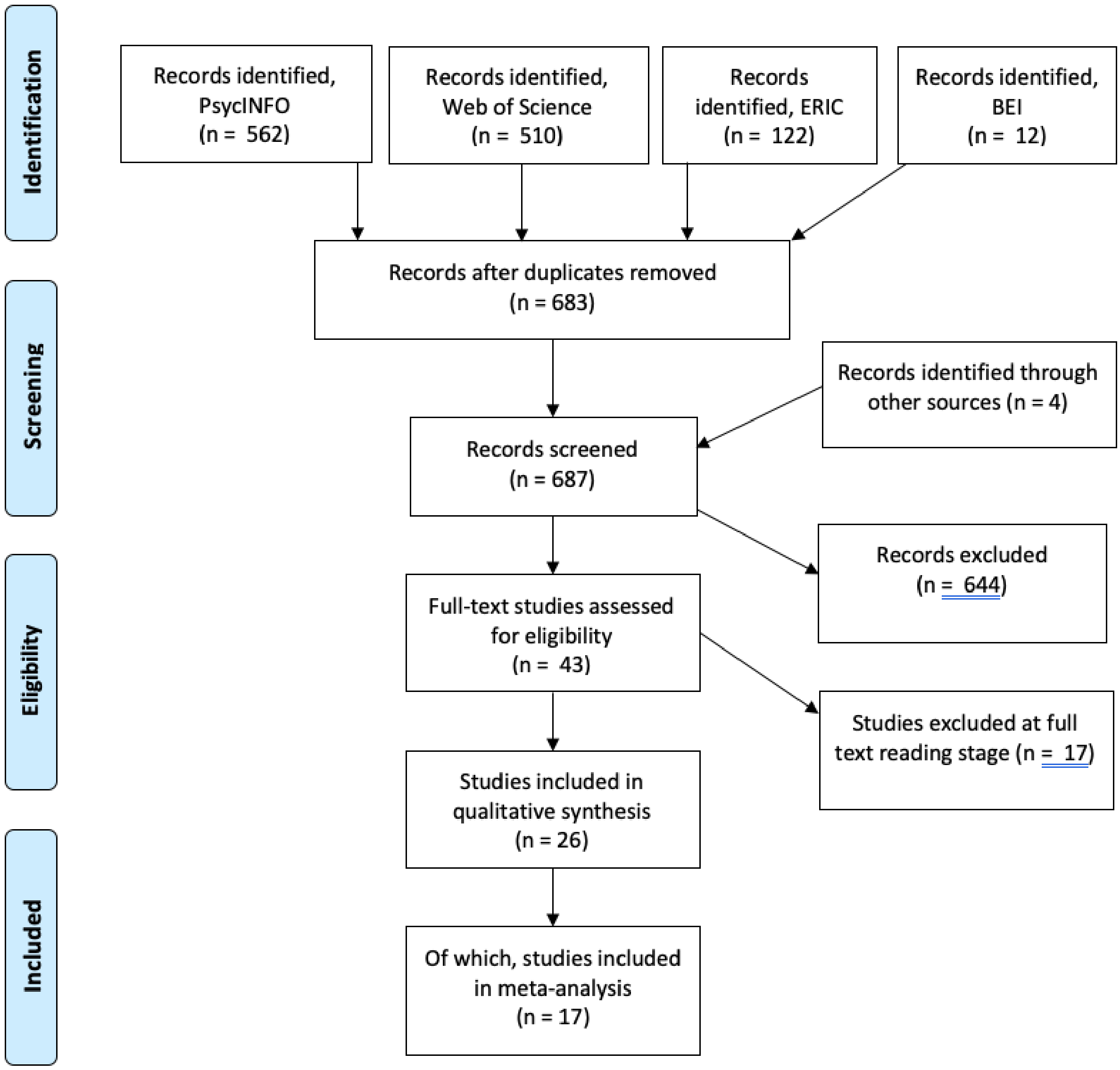

Identification
Screening
Eligibility
Included
Records identified, PsycINFO (n = 562)
Records identified, Web of Science (n = 510)
Records identified, ERIC (n = 122)
Records identified, BEI (n = 12)
Records after duplicates removed (n = 683)
Records identified through other sources (n = 4)
Records screened (n = 687)
Records excluded (n = 644)
Full-text studies assessed for eligibility (n = 43)
Studies excluded at full text reading stage (n = 17)
Studies included in qualitative synthesis (n = 26)
Of which, studies included in meta-analysis (n = 17)

Figure 3.2: PRISMA diagram summarising search and screening process

3.2.3 Full text search.

The same inclusion and exclusion criteria also applied here, together with the requirement that the full text was available. In addition, some studies from the initial search had focused on interleaved *learning* (where new examples/information were presented to learners) and others on interleaved *practice* (where items were revised, subsequent to initial learning), and I excluded the latter group because they did not compare interleaving and blocking during the initial learning phase. Overall, this resulted in 17 further exclusions, with 26 papers being retained.

Although studies of learners aged twelve and over were eligible for inclusion, none of the studies had used participants from high schools. Three used general adult samples obtained via Amazon Mechanical Turk (MTurk), and the remainder used undergraduates. This should be borne in mind when interpreting the findings reported in the following sections. In some cases, studies included several experiments, some of which met the inclusion criteria while others did not. The full list of articles included are summarised in Appendix 5; column 1 details which component experiments were included.

At times the relevance of experimental tasks to classroom learning caused some difficulty. For example, I viewed the visual blobs (presented as "alien cells") used by Carvalho and Goldstone (2014a; 2014b) and the lines and shape combinations used by Noh et al. (2016) as more science-relevant than the labelled checkerboards presented by Carvalho and Albuquerque (2012), but recognise that there is not always a clearly delineated difference between abstract and realistic visual stimuli, leading to some subjectivity.

3.2.4 Review strategy

At this point, it was important to identify which of the remaining 26 studies could feasibly be included in a meta-analysis, in order to establish an effect size (if any) of the interleaving effect, and this requires a high level of homogeneity of the inclusions. For this reason, studies which had used a very different methodology from the norm, such as Linderholm et al. (2016) – who interleaved written passages texts and tested learning via themes in essay writing – were included in the narrative synthesis only. Likewise, studies which had tested specific features and variables besides or in addition to pure interleaving (for example, in the study by Yan et al., 2017, a gradual, blocked-to-interleaved schedule was investigated) were included in the narrative synthesis, whereas the meta-analysis included only those experiments which featured a clear interleaved vs. blocked comparison as two of their conditions. This element of my review strategy brings two main benefits: improving the validity of the meta-analysis, and allowing the narrative synthesis to explore boundary conditions associated with interleaving.

Even within the studies that were appropriate for inclusion in the meta-analysis, not every constituent experiment within a particular study could be included. Studies/experiments selected for the meta-analysis are listed Appendix 6.

3.2.5 Weight of evidence.

I also reviewed the extent to which the studies and their constituent experiments provided strong evidence that could be used to answer the review questions.[6] Following Gough (2007), I considered two main aspects of quality: (i) sample size and methodology (randomised experimental methodology was considered ‘high’ quality), and (ii) relevance, which here included both the type of items used for the testing memory and transfer, and the mundane realism of the task.

Consistency of rating by the lead reviewer was established by training and a reliability check on a ten percent sample of the included studies by two independent trained reviewers. None of the studies were rated as ‘low’, and one was rated as ‘medium’ overall; the impact of removing this study from the meta-analysis is described later. The combined weight of evidence judgement (Gough’s WoE ‘D’) for each study can be seen in Appendix 5, column 4.

[6] I am indebted to J. Boyle and I. Rivers for their invaluable contribution to this process as raters.

3.3 Narrative Synthesis

As noted already, a list of all experiments included is provided in Appendix 5, and all of these were considered as part of the narrative synthesis. The following sections review the main factors that emerged from this stage of the review, and address research question 1.

3.3.1 Type of stimulus.

One task in particular dominates the literature on the interleaving of concept learning – categorization of the work of modern artists. The art task first devised by Kornell and Bjork was used by 21 of the constituent experiments reviewed, within 12 studies overall, and interleaving was consistently found to be superior to blocking, notwithstanding some boundary conditions that are discussed in the following subsections.

Many other research papers also use visual learning tasks, while some used verbal stimuli. A summary of all of the materials used across the different studies is shown below (see Table 2); in the case of direct repetitions, only the first study chronologically to use the material is named. Many of these tasks/materials are referred to in the subsequent sections.

Table 3.2: Stimuli Used Across All Studies

Stimuli	Research study
“Alien” cells (blob shapes)	Carvalho & Goldstone (2014a)
Abstract digital pictures	Zulkiply & Burt (2013a)
Art eras (e.g. impressionism era)	Birnbaum (2013)

Bird images	Wahlheim et al. (2011)
Butterfly images	Birnbaum et al. (2013)
Case studies of psychological disorders	Zulkiply, McLean, Burt & Bath (2012)
Chemistry molecules	Eglington & Kang (2017)
Fribble objects (Williams, 1997)	Carvalho & Goldstone (2015b)
Images of “alien” creatures	Carvalho & Goldstone (2017)
Medical/physiology passages	Dobson (2011)
Modern art painters’ styles	Kornell & Bjork (2008)
Psychology concept definitions	Rawson et al. (2015)
Statistics examples	Sana, Yan & Kim (2017)
Ziggerin objects (Wong et al., 2009).	Carvalho & Goldstone (2015b)

As can be seen from Table 3.2, many of the materials used are directly relevant to education, ranging from images to the verbal statistics descriptions used by Sana and colleagues (2017). Some items were drawn directly from educational materials; Dobson (2011) used text-based examples of physiology descriptions, and Rawson et al. (2015) took psychology definitions/examples from mainstream textbooks. The longest interleaved were textbook sections averaging 319 words, in the work of Linderholm et al. (2014). These were something of an outlier; among the examples of psychology concept definitions provided in Appendix A of Zulkiply (2013, p. 244), the longest text is 124 words long (an example of schizophrenia, labelled as “category TEM”). The longest example paragraph provided by Dobson (2011) is 55 words long, and the longest in Rawson et al. (2015, p. 485) is 34 words long. Thus, it can be seen that the texts used in studies of interleaving vary but are generally quite short. None of the included studies used lengthy tasks, or entire lessons, or entire topics.

3.3.2 Category and item similarity.

A number of studies looked at the effect of having categories which are more or less distinct from one another, or varied the similarity of example items within categories. In Birnbaum (2013, experiment 4), a comparison of learning of art eras vs. artists was undertaken. While artist styles are best learned interleaved (see above), art eras were better learned when blocked. This suggests that a diverse category (such as ‘all impressionist artists’) may benefit from blocking, due to the low level of similarity *within* each category.

Carvalho and Goldstone (2014a, 2015b) presented results which suggest that if category members are all alike then interleaving is best, while if category members are more diverse then blocking may be beneficial. They argue that blocking and interleaving prompt different attentional foci; blocking can usefully promote inter-item comparisons within a diverse category. Zulkiply and Burt (2013a, experiment 2) reached a similar conclusion.

MacKendrick (2015) took this a step further, varying both within- and between-category similarity systematically, and considering four possibilities of category similarity: high-within (HW), low-within (LW), high-between (HB) and low-between (LB). She found that interleaving benefited learning of HB-HW and especially HW-LB categories. Unlike the previous studies mentioned, there was no advantage of blocking. However, there was a floor effect in the LB-LW condition, with proportions correct at 0.08 or less.

Eglington and Kang (2017) note that it is hard to experimentally vary within-category similarity with art stimuli (although c.f. MacKendrick, 2015), but they were able to do so with their chemistry molecule stimuli. However, doing so had neither a main effect, nor an interaction with interleaving vs. blocking (experiments 2–3), suggesting that this factor may have a limited role to play with real educational materials.

On the whole, categories with high within-category variability (for example all animals, all modern artists) are often best blocked, while categories with low within-category variability (e.g. a specific species of animal, or a particular medical problem) are likely to benefit from interleaving, though more studies with realistic materials are needed.

3.3.3 Improvement during the task.

As Kornell and Bjork (2008) state, "*Recognition test trials are, inevitably, also learning events*" (p. 589). For this reason, multiple test items can lead to further learning, just as is the case in the classroom. The studies by Carvalho and Goldstone (2014a, 2014b, 2015b, 2017) usefully compare performance across blocks. In Carvalho and Goldstone (2014a), for example, participants took part in four learning phases, each time seeing the examples three times, and therefore seeing each one twelve times in total. Blocking was superior during four learning phases, but in a test phase which included transfer, interleaving was superior for high-similarity items, and neutral for low-similarity items.

3.3.4 Retention interval.

The retention interval between a spaced or interleaved presentation can also be varied; most studies in this review used a very short retention interval, measuring performance straight after the initial learning task. Seven studies in the review included tests of retention over timescales which are more in line with educational applications (see Table 3.3, overleaf).

Table 3.3: Interleaving Studies with Longer Retention Intervals

Carvalho and Goldstone (2014b)	24 hours
Carvalho and Goldstone (2017)	3 days
Dobson (2011)	up to 10 days
Eglington and Kang (2017)	2 days
Linderholm et al. (2016)	7 days
Zulkiply (2013)	7 days
Zulkiply and Burt (2013b)	7 days

Out of the studies in Table 3.3, Zulkiply (2013) and the Carvalho and Goldstone studies were the only ones to directly compare immediate versus long-term retention. While performance was (unsurprisingly) somewhat worse after a delay, all three of these studies found no significant interactions between interleaving and an immediate/delayed condition. This provides evidence that interleaving is not an artefact of the spacing effect (see Section 1), and suggests that its advantages can persist across educationally-relevant timescales.

3.3.5 Practice and hybrid schedules.

In authentic educational settings, study is not a ‘one-off’ experience with novel items, but typically builds on previous learning and is followed up by further practice (Bransford et al., 2000; Willingham, 2004). During this process, interleaving and blocking do not have to be absolute; it is possible to apply a schedule that includes some blocking and some interleaving (Kang, 2016).

Yan et al. (2017) experimented with a blocked-to-interleaved schedule; it was not advantageous over a pure interleaved schedule but was no worse either, and was superior to blocking. Birnbaum (2013, experiment 2) extended Kornell and Bjork's artist experiment by adding a pre-training phase during which participants were given exposure to the artists' names. This experiment found that not only did the interleaving advantage persist, it was greater than that found in the standard procedure (which in this experiment was the control condition).

Overall, the evidence reviewed did not strongly support or repudiate the idea of combining interleaving and blocking, and nor was it much studied.

3.3.6 Rule-based (non-inductive) learning.

Categorisation of novel objects based on a verbal rule may rely on explicit cognitive process, mediated by different brain areas than the implicit, information-integration system that was referred to in the introduction (Ashby & Gott, 1988; Birnbaum, 2013; Rawson et al., 2015). Can interleaving also be beneficial in situations where such a deductive rule is possible, and/or where such a rule is directly supplied?

Zulkiply (2015) tested learning of paintings using a verbal rule (which she described as "top down" learning) versus inductive ("bottom up") learning. As with previous studies, an interleaving benefit was found with inductive learning; this study found that there was also an advantage when learning was top down, but that it was reduced.

However, Noh et al. (2016) studied the interleaving of shapes (patterns with lines and dots) and found that blocked presentation led to better outcomes with rule-based learning, while interleaved presentation led to better outcomes where there was no clear verbal rule, with information integrated more holistically.

Rawson et al. (2015, Experiment 1b) presented examples of social science concepts with or without a definition. While interleaving was found to be of benefit when examples were displayed by themselves, the presence of a definition on screen disrupted this benefit.

Overall, interleaving appears to be more useful for subtle pattern learning where there is no clear verbal rule, perhaps because this prompts learners to pay attention to contrasts.

3.3.7 Working memory span.

Categorization draws on existing knowledge of categories/schemas and therefore on long-term memory, but the processing of a particular example and allocating it to a previously-stored category will occupy working memory and attention. WM has a limited capacity for processing, and in education, its propensity to be occupied or distracted is often described in terms of 'cognitive load', with researchers such as Sweller, Ayres and Kalyuga (2013) expressing the view that cognitive load needs to be managed effectively in the classroom. In brief, cognitive load can be measured in terms of the number of separate elements a task involves (element interactivity), and too much cognitive load is thought to overload a learner's working memory and therefore lead to inefficient learning.

However, Birnbaum (2013, experiment 3) found that an interleaving advantage can persist even when learning is incidental, suggesting that the interleaving effect cannot depend purely on effortful problem-solving within working memory. In their test of statistical problem solving, Sana et al. (2017) found an interaction such that the benefit of interleaving disappeared for learners (undergraduates) with the highest WM scores. In an apparently contradictory finding, Guzman-Munoz (2017) found a marginal correlation – significant when findings from his three experiments were pooled – which suggested that high-WM learners benefit *more* from interleaving. The results are therefore mixed, and a further limitation is the lack of studies on adolescents found in the current review.

3.4 Quantitative Analysis and Meta-analysis

From the studies accepted for inclusion in the narrative review, it was at this stage necessary to separate out the different constituent experiments reported, because often these had different features such as different samples or stimuli. In total from among the 26 studies reviewed, there were 56 eligible experiments (see Appendix 5). From among these experiments, 28 were deemed suitable for inclusion in the meta-analysis, coming from 17 out of the 26 studies overall. The following two sub-sections describe these experiments as a whole. Further subdivision of four of the experiments into memory and transfer conditions yielded 32 items overall for the meta-analysis.

In some cases, the information summarised in the coming sections relates to only part of an experiment, due to the need to separate out relevant from irrelevant conditions within the same experiment. For example, in the study by Kornell et al. (2010), both young adult participants and elderly participants were recruited, and (in line with my inclusion criteria), only data from the younger participants were analysed here. In such cases, only those part of the experiments are reported in this part of the analysis, and in the relevant tables and appendices.

3.4.1 Categories and exemplars.

There was a level of variation in terms of how many exemplars and how many categories were used by the experiments selected. The variation in number of categories used is summarised in Table 3.4; as can be seen, the modal number of categories was 12 (due in large part to the many replications of the Kornell & Bjork, 2008, methodology), while the mean number of categories was 8.38 and the median was 8. The mean number of exemplars of each category was 7.34, and the mode was 6 (again, due to replications of Kornell & Bjork).

Table 3.4: Number of Categories and Examples, All Constituent Experiments (n = 56)

Number of categories used	**Number of experiments**
2	4
3	13
4	4
5	1
6	5
7	0
8	3
9	0
10	1
11	0
12	19
13+	6

The highest number of categories used was by MacKendrick (2015) with 20, in a study designed to raise the memory load on participants. Dobson (2011) and Carvalho and Goldstone (2017) each used just two categories.

Some studies varied category number to increase difficulty. For example, Eglington and Kang (2017) raised this from four to eight between Experiments 3 & 4; however, in their same study, Experiment 2 was designed to be more challenging than Experiment 1 despite using fewer categories (four rather than five) and otherwise the same methodology. In this and several other studies, difficulty level depended on the number of distinctive features of each item, that is, their complexity (see Carvalho & Goldstone, 2017), and also on the similarity of the examples presented. An interim conclusion is that the number of categories used can be seen as a measure of difficulty only so long as other factors remain constant, with complexity and similarity the other principle factors.

The total number of examples used in the learning phase of each experiment is typically a multiple of the number of categories and the number of examples of each type. This total ranges from 17 (Dobson) to 80 (MacKendrick), with a mean of 50.9 examples overall. Full details are shown in Appendix 5. However, it is important to note that in some studies, each item from the learning phase was shown more than once. These numbers may serve as a guide to the planning of future studies or applications of interleaving.

The number listed in brackets in Appendix 5 (column 7) relates only to the learning phase of each experiment; studies which tested transfer employed additional examples of each item for the test phase, typically between 1–4 (for example, Kornell & Bjork Experiment 1a used 10 exemplars of each artist overall, 6 for study and 4 for testing).

The materials used in the studies across the whole review are discussed above. Table 3.5 summarises the materials from those items used in the meta-analysis only (see also

Appendix 6). For simplicity, I decided to categorize modern art as 'art', and the remaining types of materials as 'science'. This will be discussed further in the meta-analysis (see below).

Table 3.5: Stimuli Used in Items (n = 32) Included in the Meta-analysis

Materials used	**Number of experiments**
Modern art	14
Animal images	3
Chemistry/maths images	6
Verbal – biology/psychology	6
Verbal – statistics	3

3.4.2 Transfer versus memory.

A further issue which I had to consider when preparing the data for meta-analysis was that some experiments included both a memory and a transfer condition — testing participants on both repeated items from a learning phase and on their ability to classify new items of the same type. Others only measured transfer, and one (Dobson, 2011) measured only memory.

In order to make a valid measure of an effect size of interleaving, I felt that it was necessary to separate out the transfer and memory conditions of such experiments. Splitting the experiments in this way resulted in a total of 32 items for further statistical analysis (transfer, within-participants design: 13 items; between-participants: 3 items. Memory, within-participants design, 3 items; between-participants: 13 items). These are shown in Appendix 6. As can be seen, some of the rows in Appendix 6 relate to the same study, divided into two parts, memory and transfer, each with effect sizes calculated. All of the effect sizes were in a positive direction (interleaving superior to blocking).

However, as the participants in these parts of the experiment were the same (none of the included studies compared memory vs. transfer between groups), the findings didn't have independence, and I therefore compared them separately. Similarly, I compared within-groups and between-groups designs separately, resulting in four main analyses. This is reported in the next section.

3.4.3 Statistical findings: meta-analysis.

Research questions (b) to (e) were addressed using meta-analysis. Random effects models were used throughout (see Borenstein et al., 2009).[7] Random effect models are considered more efficient, and they help to control for unobserved heterogeneity, provided that their assumptions are met. By heterogeneity, I am referring to unexplained variation among the data such that otherwise similar data (such as findings from the same condition) differ statistically, suggesting variation in the effect of a particular variable or intervention. Random models assume that heterogeneity among the sample is uncorrelated with independent variables, while fixed effects models assume the opposite. In the present case, the random effects model appeared justified given that the heterogeneity in the findings from different experiments are likely to be due to lab- or methodology-specific factors (such as the tasks used and the local population of participants sampled) rather than being correlated to the independent variables under consideration.

[7] I am indebted to J. Boyle for his guidance with the software and process of meta-analysis.

3.4.3.1 Within-participants designs, transfer.

I first carried out an analysis on the records from the 13 within-participants experiments which had tested transfer. Using a random model in the Comprehensive Meta-Analysis software (Borenstein et al., 2014), I calculated a pooled standardised effect size (Hedges *g*) of interleaving at +0.59, *p* < .001, 95% CI [0.45, 0.72]. Hedges g is one of the main calculations of effect size along with Cohen's d, and is considered to be equally sound in most cases and superior when using small sample sizes (Field, 2018).

Egger's test (Egger et al., 1997) was used to test for publication bias for this group of records, and the results were non-significant indicating no publication bias (intercept 3.77, 95% CI [-1.47, 9.00], t = 1.58, df = 11, two-tailed p-value .142).

I noticed high levels of heterogeneity among this group of records (Q statistic (12 df, *p* = 0.00) of 45.4; I-squared = 73.6; Tau Squared = 0.04). The Q-statistic and other tests measure heterogeneity in a similar way to chi-square tests for goodness of fit of an observed distribution, and lower levels are considered preferable because any meta-analysis relies on pooling relatively similar sets of data; very high heterogeneity might indicate that the effects of the independent variable fluctuate to wildly to be comparable at all. Here, the convention of Q = 0–40 being seen as low – as described in the Cochrane handbook (Higgins & Thomas, 2019) – was adopted, with lower values within this range seen as preferable.

There were outliers in the present set of records, and with 4 of these iteratively removed — all at the high end of the effect size range — heterogeneity was fell to a point within the acceptable range (Q-value = 8.64 [8 df; *p* = 0.37]; I-squared = 7.38; Tau Squared = 0.001). The effect size from the remaining 9 records (Hedges *g*) was +0.43, *p* < .001, 95% CI [0.35, 0.52].

I also wanted to know if the learning materials used contributed to the variability found among this group of studies. I therefore conducted a further analysis on the original 13 records in this category, subdividing them by materials (art or science). In terms of the records from studies which had used art images, the pooled *g* was +0.64, *p* < .001, 95% CI [0.47, 0.81]. Surprisingly, given that all of these studies used similar materials and some were direct replications, the heterogeneity remained quite high (*Q* = 29.3).

The science-based experiments (*n* = 5) also showed one outlier. A sensitivity analysis[8] with the outlier removed resulted in the following findings: *g* = +0.38, *p* < .001, 95% CI [0.25, 0.51]. The Q-value was 1.86, that is to say, these studies were highly homogenous, despite using different types of science-based stimuli.

Together, these findings suggest that the variability in the experiments from this group was not due to the materials used, and also provide interim support for prediction (d), namely that art-based materials will have a larger interleaving effect.

3.4.3.2 Within-participants designs, memory.

I next ran an analysis on the three records from within-participants designs which had tested memory, and calculated a pooled effect size of g = +0.65, *p* < .001, 95% CI [0.50, 0.80]. Egger's test was again used to test for publication bias for these records, and the results were again non-significant (intercept -4.08, 95% CI [-58.5, 50.4], t = 0.95, df = 1, two-tailed p-value .516). This time, there was a low level of heterogeneity among the records (*Q* = 1.02), and I did not detect any outliers. However, I decided that this set of records was too small to compare sub-groups with different types of stimuli/materials on the basis that one such sub-group would have contained only one study at most; two studies is the generally accepted minimum for a meta-analysis to be carried out (Higgins & Thomas, 2019).

[8] I am indebted to J. Boyle for his guidance on this stage.

3.4.3.3 Between-participants designs, transfer.

I next analysed the records from studies of transfer, this time focusing on those 13 records with between-participants designs. Again using a random model, I calculated a pooled effect size *g* of +0.66, *p* < .001, 95% CI [0.49, 0.80]. Egger's test was again used to test for publication bias, and the results were again non-significant (intercept 3.76, 95% CI [-0.38, 7.91], t = 1.99, df = 11, two-tailed p-value 0.071).

As a measure of heterogeneity, I found Q = 18.4, which is again within the acceptably low range. There was one outlier apparent, a study that had been rated as 'medium' quality (see Appendix 5). Removing this record and re-analysing the data led to a pooled effect size of 0.61, and to a lower Q value of 10.4. Thus it can be seen that the level of heterogeneity among these experiments was low, and even after removing the record with the highest effect size, the pooled effect size remained at a comparable level.

Filtering the remaining 12 records by stimuli led to pooled effect sizes of *g* = +0.56, *p* < .001, 95% CI [0.41, 0.71] for the experiments which had used science-based materials (*n* = 8), and *g* = +0.82, *p* < .001, 95% CI [0.46, 1.17] for the experiments which had used art-based materials (*n* = 4), providing further support for prediction (d) that there would be a larger interleaving effect with art-based materials.

3.4.3.4 Between-participants designs, memory.

Finally, I analysed the 3 records relating to memory from experiments with between-participants designs. Their pooled effect size was *g* = +0.39, *p* = 0.011, 95% CI [0.09, 0.69].

Egger's test was again used to test for publication bias and the results were again non-significant (intercept 6.35, 95% CI [-165, 179], t = 0.47, df = 1, two-tailed p-value .72). Again there was a low level of heterogeneity (*Q* = 2.74). Again, this set of records was too small to make a comparison using sub-groups with different types of stimuli/materials.

3.4.3.5 Sensitivity analysis

I was aware that there were multiple records from the same researchers or labs among the data, presenting a possible source of bias. As a sensitivity analysis I therefore further analysed each set of 13 records focusing on transfer as follows.

For the records based on within-groups designs, I compared the work of Kornell and colleagues (n = 3) versus the others (n = 10), and found effect sizes of +0.60 and +0.59 respectively, a non-significant difference (Q = 0.005; p = 0.94). For the studies featuring the work of Zulkiply (n = 3), the effect size was +0.60 compared to +0.59 for the other records (Q = 0.001; p = 0.97). Finally, for Guzman-Munoz (n = 2) the effect size was +0.80 compared to +0.56 for the others (Q = 3.81; p = 0.051). Only the last comparison approached significance; it would be more parsimonious to focus overall on the lower effect size from the latter comparison – that is, an effect size of 0.56 for transfer among interleaved within-groups designs. There was a relatively high level of heterogeneity in each of these comparisons.

For between-groups designs, I compared the work of Kang and colleagues (n = 6) versus the other records (n = 7), and found effect sizes of +0.63 and +0.73 respectively; in the latter case the effect size was +0.62 if the outlier mentioned earlier was removed. The difference was not significant in either case (Q = 0.002; p = 0.97 with the outlier removed). There was a low level of heterogeneity in both sets of studies when analysed without the outlier. Finally, for Sana and colleagues (n = 3) the effect size was +0.53 compared to +0.66 for the others excluding the outlier (Q = 0.705; p = 0.401), again with low heterogeneity

Overall this sensitivity analysis showed a consistent interleaving effect among the records, with comparable effect sizes across different labs.

3.5 Discussion

3.5.1 Analysis of findings.

In the previous sections I outlined the review protocol and described how 26 studies were identified. The full set of studies were taken into account in the narrative synthesis, which explored variables which interact with or limit the interleaving effect. The most homogenous constituent experiments were then investigated via a meta-analysis, broken down by experimental design and by the focus on either memory or transfer.

The findings have revealed that presenting examples in an interleaved order reliably leads to an advantage over blocked presentations. This advantage is associated with large effect sizes in the laboratory, especially for art-based materials but also for science-related category learning, and therefore shows potential for broad educational application. The sensitivity analysis for clustering by lab revealed that the findings persisted even when the work of different labs were analysed separately, with little indication that the effect sizes have been inflated by same-author dependencies. If anything, the very close effect sizes reported in the final subsection above – based on the work of different researchers using different materials – strengthen the idea that it is the interleaved order itself, rather than any other factor, that benefits memory and transfer. However, this is a counterintuitive finding, unlikely to be applied spontaneously by teachers due to the fact that it often makes the learning phase seem more effortful, and learners too tend to incorrectly assume that blocking is the better strategy (Rohrer & Pashler, 2010; Yan et al., 2017). In short, it is a desirable difficulty (Bjork & Bjork, 2011).

The limitations of the evidence base indicate that there is an urgent need to more fully investigate the use of interleaving in school contexts, and to use school-relevant materials and samples, as nearly all included studies used undergraduate participants (with three using participants from the broader adult population). The limited and mixed results on the role of working memory in the effect have not fully supported the use of interleaving for learners of other ages, but neither do they speak strongly against it. Outside of the studies reviewed, an investigation by Vlach et al. (2008) found that interleaving was advantageous for pre-school children, suggesting that the benefits found in this review can potentially generalize to learners at any educational stage. Further research into the effect and any interactions that may emerge with younger groups of learners, including those with special educational needs, is desirable.

It can be seen that the research into interleaving features sets of materials that are quite varied — ranging from modern art to statistics. From the Egger's test, publication bias was not implicated among the papers reviewed, though I note that the statistical power of this test is low in cases where there are fewer than ten studies (Higgins & Thomas, 2019), which here applies to the two samples of studies of memory. The literature does, however, primarily feature relatively artificial and short-term tasks, and most of the studies used visual stimuli. Future work should investigate school learners in real classroom contexts, ideally using materials which are drawn directly from their school courses and which build on prior learning.

This review has indicated that a number of studies have explored the degree of similarity of stimuli between and within categories, and that this factor appears to make a difference. In particular, interleaving is beneficial when stimuli are similar (and differences are therefore hard to notice without opportunities for contrast), and that the effect can even be reversed – with blocking more beneficial – if materials are very diverse. When using verbal materials, the length of examples is potentially quite variable, and there was some tentative support found for the idea that verbal examples should be short rather than extended. These aspects of the findings fit the discrimination-contrast hypothesis (Birnbaum et al., 2013; Wahlheim et al., 2011), and suggest that there is unlikely to be a major benefit if teachers were to interleaving entire activities, lessons or topics.

Although a minority of studies reviewed (e.g., Zulkiply, 2013) included an educationally-relevant retention interval (and even then, it was matter of days, not weeks or months), the early evidence indicates that the interleaving benefit persists after a delay, with or without an intervening practice session. Future field studies could investigate longer delays, and must also take account of common classroom practices; for example, the likelihood is that teachers do not present 50 or 100 examples at the start of a lesson, and instead integrate smaller number of examples into other activities over time. However, as revealed by some of the studies reviewed (Birnbaum et al., 2013; Kang & Pashler, 2012), spacing can be detrimental to interleaving, suggesting that a gradual approach to presenting examples to learners will be sub-optimal if it reduces opportunities for contrast; if contrast is maintained, then spacing and interleaving can work in concert (Birnbaum, 2013).

I predicted that interleaving would have a greater impact on transfer-based tasks than memory-based tasks – prediction (c) – but found that both showed large and comparable effect sizes. This finding supports the idea that interleaving is potentially of use when training learners in situations where they will need to encounter new example stimuli (previously-unseen new examples of cells or texts, new psychological case studies, and so forth). However, it shows that it is equally applicable to boosting recall on a practice and retrieval basis. This finding may be in part because the same transfer processes that are helpful for new items can also be applied to previously-seen examples, which are unlikely to be remembered perfectly.

Furthermore, I predicted that the set size and number of categories would affect interleaving – prediction (e) – and the findings here were inconclusive. This is mainly because very few of the studies reviewed have directly compared a larger versus a smaller category size. It would be useful if this were further explored in future, because time is limited in school settings and it could be helpful to identify an optimal number of examples.

On a theoretical level, a number of the studies provide support for the discrimination-contrast hypothesis, but as noted earlier, this is not necessarily inconsistent with the attentional-bias hypothesis, given that interleaving — if it promotes discrimination by way of contrasting differences — must also prompt learners to pay attention to those differences. A broader theory of interleaving will locate it within the psychology of category learning and of attention and memory in general, and will account for the role of multiple concrete examples (versus no examples, or only one or two) in forming new understanding.

It is interesting to consider why visually highlighting key differences between chemical molecule types did not modulate the effect. These findings support the idea that category learning processes happen on an automatic level, with a learner's limited attentional resources being occupied by the task in hand, and minimal metacognitive control of the process. This also fits with the limited evidence of an interaction with working memory.

Following the work of Kornell and Bjork (2008), interleaving came to be seen as primarily relevant to image-based inductive learning, but this review has found that interleaving is also beneficial, albeit slightly less so, for science- and maths-based tasks. It would be useful to investigate its application to science or social science topics that involve visual stimuli — geography and neuroscience are two that include many relevant examples, and among older learners, there could be obvious links with medical or dental education. The methodology of the studies reviewed was fairly consistent; nearly all of the studies reviewed were laboratory experiments, and most stimuli were presented using computer screens. The quality of the studies included tended to be high, with random assignment to groups and large sample sizes, although sampling tended to be non-random. The removal of outliers did not alter the overall conclusions from the meta-analysis.

The present review has helped to advance the literature in a novel area with major practical implications (see Chapter 2). Given that blocking is everywhere in the real world of education and that learners tend to prefer blocking due to the sense of fluency it provides, the strong findings reported here suggest a potential to improve on current teaching practice in a way that could be relatively straightforward to apply, and relates to a broad range of teaching materials and topics.

The review also benefited from following a clear and replicable methodology, as indicated by the PROSPERO-registered and published protocol. The database searches summarised in the PRISMA diagram (Figure 3.2) reflect an extensive search of the literature which uncovered a number of infrequently-cited but relevant articles. I also applied a novel and nuanced approach to exploring potential publication bias or bias according to the methods used by specific labs.

A further strength of this review was that it included both a narrative section and a meta-analysis. As discussed earlier (see Chapter 2), the application of evidence to practice is a complicated and nuanced matter. It is therefore useful to explore not just the main pooled effect sizes of interleaving for memory and transfer, but also to consider the conditions under which these might best be implemented in the classroom, as well as any variables that might counteract the benefits.

There are also limitations of the present review which are worth highlighting. Firstly, as with all review studies, the findings are an artefact of the choice of search methodology. There may have been other relevant studies of interleaving which have been overlooked due to the terminology used (for example, synonyms that I did not search for), or because they were not included in the databases searched. Similarly, the time period is a limitation; at the outset, I felt that it was important to assess the work that had been done since Kornell and Bjork's

(2008) seminal study, a period of approximately one decade. Hopefully this review can be followed up in future years with further reviews of the literature as new studies emerge.

There was undoubtedly a degree of subjectivity in applying the inclusion criteria. As noted earlier (see Section 3.2), there could be some doubt about what constitutes relevance to science or social science learning, and the division of materials into art and science categories is somewhat imprecise. Other researchers may interpret this differently, and the views of practicing teachers as judges could also be used.

Finally, the present study did not address interleaved/shuffled practice of previously-learned material, a strategy frequently used in high schools (deliberately or otherwise) as part of exam preparation. A fuller picture of the research will only emerge if it is clear how and when interleaving should be used both for new learning and for review.

3.5.2 Conclusion.

In conclusion, the findings of this systematic review show that interleaving, when applied appropriately, can be a powerful tool for learning new concepts. For the educator, this extends the use of interleaving, which has been described as a useful practice and revision technique in other work (e.g. Rohrer et al., 2015). It has powerful benefits for both memory and transfer, and applies across different subject domains. However, the benefit was restricted to the presentation of visual examples or very short texts, in contrast to a common educational recommendation to interleave entire topics or activities, and it depends on the ability to contrast similar items that would otherwise be easily confused.

Further studies in this field are likely to provide essential evidence on how practical the technique is when applied to everyday materials with elementary or high school learners, as well as guidance on how best to do so.

Having identified some of the gaps in our understanding of interleaving as well as raised certain questions about how it is underpinned by metacognitive processes and professional learning, I now turn to the range of methodological choices available for progressing scientific knowledge of these areas. How best should new research approach the question of item order in the classroom? In Chapter 4, I consider the range of methodology open to the researcher for tackling concepts such as spacing, and for the study of human memory and metacognition in general – a field where the target concepts are often counterintuitive, and where individuals may have flawed beliefs about the workings of their own psychological processes.

4

Methodology

4.1 Introduction: Investigating Memory's Role in Education

On a psychological level, there is little doubt that memory plays a fundamental role in education (Bjork, 2019; Bransford et al., 2000; Dunlosky et al., 2015). Each change that occurs when a learner processes a new example, develops a skill or learns a new concept is accompanied by a transformation in memory on both a neural and a psychological level. As one research team put it, teaching is fundamentally about altering the contents of LTM: "*The aim of all instruction is to alter long-term memory. If nothing has changed in long-term memory, nothing has been learned*." (Kirschner et al., 2006, p. 77).

Recent years have seen an uptick of interest in these processes, with multiple books for teachers focused on evidence-based instructional strategies (see https://www.retrievalpractice.org/books for a useful list), together with a recognition that the metacognitive process of 'learning how to learn' can play an important role in boosting teaching quality (e.g. Darling-Hammond et al., 2020; EEF, 2018). Concepts such as spacing and interleaving have entered the everyday vocabulary of many teachers via popular books and blogs (for example Agarwal & Bain, 2019; Weinstein et al., 2018) and through professional learning experiences provided by national bodies such as the Chartered College of Teaching as well as commercial organisations (such as *Seneca Learning*[9] and *Dragonfly Training*[10]).

It arises from this fact – as argued in Chapter 2 – that the application of education-relevant memory processes such as interleaving and spacing to classroom practice need to be

[9] senecalearning.com

[10] dragonfly-training.co.uk

investigated, as do teachers' beliefs about these processes. A fuller insight into how and when concepts such as interleaving can be applied stand to greatly benefit learners and their teachers, as does an understanding of misconceptions that teachers are likely to hold.

But how exactly do we go about investigating processes which are at times counterintuitive or obscured, and about which participants may have false beliefs? And how can researchers account for the timescales involved in investigating memory and forgetting, and overcome the difficulty in accurately assessing learning via a brief task? This chapter attempts to tackle these questions, working through some of the main methodology choices that are represented in the literature thus far, and identifying other promising options. In addition, it outlines a research plan for investigating interleaving and spacing in the context of the secondary school curriculum in Scottish classrooms.

4.2 Three Problematic Assumptions

4.2.1 Overview.

When it comes to investigating and applying education-relevant memory processes, there are certain assumptions that tend to either implicitly or explicitly guide researchers, policy makers and practitioners. In this section I will analyse three such assumptions, each of which will influence the appropriateness of the methodology choices that could be used to investigate the processes under consideration in this thesis:

- The first assumption is that education-relevant memory processes are *absolute and universal* in that they pertain to every learner. This means that memory is something that applies to all, and that any conclusions drawn from research (that technique x is better than technique y, for example) will apply to every situation and for every learner. This assumption is most obvious in the widely-discussed '*what works movement*' (henceforth WWM) in education, where various authors and policy-makers have suggested that as with evidence-based medicine, a particular set of educational practices work better than others, and these should therefore be identified and then applied to all teaching and learning situations; see, for example, the EEF's 'Teaching and Learning Toolkit' (EEF, 2018) in the UK, or the American 'What works clearing house' (Institute of Education Sciences, n.d.).

- Secondly, an assumption that memory processes are relatively *stable* and amenable to investigation and manipulation. This position portrays memory as something akin to a physical property in that it can be measured and quantified at any given point, and increased and decreased via interventions. By extension, it may also be assumed that memories will remain static if no attempt is made to interfere with them. It is an assumption that may be observed in classroom settings when referring to a person's memory competence in general (for example, a pupil stating that they 'have a good memory') as well as to specific memories (for example, a teacher stating that 'most of my class remember fact x').
- A third assumption – often implicitly held – is that memory processes are open to scrutiny by their own users, and can be guided on the basis of intuition. This again is particularly evident in educational settings. A teacher or parent might ask learners to state how well they understand something, for example, or use techniques which ask them to gauge their progress. Indeed, the entire approach of lesson observation – a staple technique in teacher education/teacher training and in the development of early-

career teachers[11] – implies an assumption that the state of students' learning can be observed by a visitor to the classroom, based on behaviour over a brief time period.

I will now analyse each of these assumptions in turn in order to identify the methodological issues that arise from each one.

4.2.2 Universality.

The first assumption that I will scrutinise is a tendency to assume that education-relevant memory processes are *absolute and universal* in that they pertain to every learner. This assumption is implicit in examples like the following, which draw on writings in education (Kirschner), psychology (Willingham), educational psychology (Sweller), and the views of prominent teachers (Didau):

- "…students learn more deeply from strongly guided learning than from discovery (Kirschner et al., 2006, p. 79).
- "That students will only remember what they have extensively practiced—and that they will only remember for the long term that which they have practiced in a sustained way over many years—are realities that can't be bypassed." (Willingham, 2004, "What Material Merits Practice?" section).
- "Most students are novices and so most of the information provided to them is novel and must be processed by a limited capacity, limited duration, working memory." Sweller et al., p. 44).

[11] In Scotland, for example, student teachers are visited and observed by a course tutor three times during their PGDE course, each time for around an hour. They are then visited approximately 10 times across the first year of practice by a line manager or member of school management. A key consideration for such visits would be whether or not there is evidence that the pupils learned what was intended, and the teacher would be expected to share "learning intentions" at the start of lessons and review these at the end.

- "It's well understood that all human being *[sic]* have a limited capacity for paying attention to information and that we can only think about four 'chunks' of information at any given time." (Didau, 2018, para. 6).

Universality is a key research concern in this thesis – do concepts like interleaving and spacing apply to all learners, for example, and to all school activities? In order to answer this question, it's worth briefly considering what exactly it would mean for a trait to be viewed as universal. As Norenzayan and Heine (2005, p. 764) put it: "The question is difficult because [it] requires one to make distinctions between the concrete, particular manifestations that can be observed in behavior and the abstract, underlying universals that have given rise to those behaviors". This is certainly the case in the classroom.

It is therefore important to ask whether proposed features of memory (such as the spacing and interleaving effects) are:

1) universally applicable for all learners in every situation, or

2) universally applicable for all learners in terms of abstract principles, but likely to be modified by the specifics of the learner and/or the situation, or

3) not universal at all.

It could be seen that preferring statement (1) to statement (2) represents a stronger version of universalism as it applies to education, while preferring statement (2) to statement (3) represents a weaker, situation-dependent interpretation of universalism.

Educationally, the implications of accepting the strong version would be that if certain techniques were held to 'work', there would be no need for educators to consider learner age, preferences, or other psychological traits (at least among those learners who can be considered developmentally typical) at all. Techniques such as spacing could be applied without any need for analysis of the learner/group, or the specifics of the situation and task.

However, as has already been analysed and discussed (see Chapters 2 & 3), there are a number of factors – including the level of meaningful contrast between examples, the length of each example, and the prior knowledge of the learner – which might contradict this strong interpretation (and the implied simplistic approach to practice). In terms of the task, for example, the level of similarity between example stimuli is likely to affect whether the interleaving of examples is an optimal strategy or not (see Chapter 3), while it is possible that the optimal timing of spacing will depend on the task difficulty (see Chapter 2).

Focusing in on the learners themselves, individual differences have at times been neglected by psychology, with many researchers studying a limited and biased sample of Western participants (Norenzayan & Heine, 2005; Sears, 1986). This is not a good reflection of contemporary Scottish classrooms, where the cohort is becoming increasingly multi-cultural. In addition, most psychology research is conducted on university undergraduates, participants who are adults, and by definition the more academically-able segment of the population. This can lead to an unrealistically idealised view of cognitive functioning, and one that does not generalise well, as (for example) over 25% of pupils have a recognised additional support need (Scottish Government, 2018).

There is also variability within individuals; a person's cognitive function is not static across all situations, and cognitive and behavioural traits are subject to prevailing conditions and to a person's emotional and motivational state (Smith, 2018; Wilson & Korn, 2007). Working memory functions can be reduced by stress, and are greater in domains where the learner has prior knowledge that enables chunking of information (Ericsson & Kintsch, 1995; Miller, 1956). Meanwhile, tasks that form the mainstay of cognitive psychology research are often short term, and participants are often paid for doing them. They can be boring and even uncomfortable to do, in a way that would not be practical in the classroom. Importantly for the arguments presented in this thesis, these points suggest to me as a practitioner that a technique which has been found to be helpful in an artificial laboratory task could, plausibly, be neutral or even harmful when applied in schools.

These matters suggest that a certain caution should be considered when generalising memory phenomena to classroom practice. It is important to remain mindful about the type of learning experience used to gather data when drawing conclusions about the efficacy of proposed classroom or revision techniques. Techniques that appear to work for Mathematics can't be assumed to transfer to other curriculum areas, for example, and even within a subject discipline, the features of the task itself and the target learning material will play a critical role in whether material should be interleaved or blocked.

More specifically, it is important that tasks that will be used in classroom-based research on pupils – especially where this is done with a view to applying the techniques more widely – must be accessible to all members of a typical school class. Each must have sufficient face validity as a learning task, in the sense that the task appears to participants to be worthwhile and authentic (Loyd et al., 2005). It must be the case that teachers and pupils could reasonably be expected to use the task spontaneously and without a research-based motivation to do so.

Where does this leave the application of memory principles to the classroom? Firstly, it should be noted that contemporary psychology tends to favour the weaker version of universality discussed above. While some features of cognition are indeed held to be universals of the human condition and may have evolutionary roots (Barrett, 2015; Cosmides & Tooby, 2013), these tend nevertheless to be seen as open to change, with trajectories that depend on developmental conditions, and features which are neither stable nor lacking in variability. While cognitive psychology researchers such as Bjork and Bjork (2011) or Diamond (2013) create models that imply universality of cognitive processes, the limitations of these models are well understood; it would not be assumed that particular features apply to all tasks and learners without very strong supporting evidence. However, such subtleties may be lost in translation when models are taken out of context and applied to education, with some authors (e.g. those quoted at the start of this section) apparently promoting techniques uncritically, and others (e.g. Willingham, 2017) actively discouraging teachers from learning their theoretical underpinnings (see Chapter 7).

It is therefore important that any conclusions deriving from the work presented in this thesis are specific when it comes to practice. Caution will need to be exercised when generalising findings to different tasks, younger learners, and across subject disciplines. Following the work of McDaniel and Einstein (2005), a key factor in whether a difficulty is desirable is whether it prompts particular processing (see Chapter 1), and this may help such generalisation to continue to align with both evidence and theory. Analysis of the specifics of classroom materials will be important, too; following the work on the type of tasks that benefit from interleaving (see Chapter 3), it will be important to consider the specific challenges that exist in authentic tasks, such as preparing for mainstream school exams that call on both factual knowledge and skills. These considerations will be important when preparing research tasks for the empirical part of this thesis, and when analysing the findings.

Can even the weaker version of universality be supported, or should (following the distinction between points 2 and 3, above) we reject the idea of universal factors in learning altogether? It is certainly possible to conceptualise the functioning of memory as having evolutionary roots that apply to all learners (e.g. Anderson and Schooler, 1991; Barrett, 2015; Nairne, 2005; Nairne et al, 2007), and spacing and interleaving would appear to fit well within that framework. An analysis by Anderson and Schooler (1991), taking into account the spacing effect and looking at retention of words from an evolutionary/functional perspective, concluded that “human memory mirrors, with a remarkable degree of fidelity, the structure that exists in the environment” (p. 404).

Notwithstanding the criticisms of evolutionary explanations outlined by Siegert and Ward (2002) – who warn against explanations based on 'just-so' stories about human origins that are devised to fit the data after the fact – both spacing and interleaving can plausibly be characterised as having evolutionary roots:

- Spacing could be seen as having been beneficial to our hunter-gatherer ancestors on the basis that it would be more advantageous for LTM to encode risks and opportunities that recur periodically than temporary conditions that repeat frequently within a short period of time and then cease to hold.
- Naturally-occurring interleaving of subtly-different stimuli such as plants, fungi or insects could aid categorisation into separate types in such a way that would improve survival chances, but only if human cognition was evolutionarily prepared to focus on and encode key differences.

The above points represent 'adaptive targets', the first steps of a process that Tooby and Cosmides (1992) recommend for carrying out an evolutionary functional analysis for proposed universal traits. Additionally, as discussed by Anderson and Schooler (1991), it is difficult to explain the various observed phenomena of human memory (in particular the power functions taken by the forgetting curve, the retention function and the spacing effect) without reference to this functional basis. Such an explanation may therefore be the most parsimonious option.

As such, there are reasons to suppose that interleaving and spacing are features of memory which – at least in principle – could hold universally due to their fit with activities that human memory evolved to carry out. Such universality would also help to explain the widespread scope of the phenomena, such as evidence of the spacing effect in studies of other species (e.g. Mauelshagen et al.,1998), and of interleaving in infants (see Chapter 3).

While further consideration of these questions is beyond the scope of this thesis, it is at least worth considering the educational importance of identifying and fully understanding the boundaries of memory factors that might hold universally rather than depending entirely on the task or the learner. Doing so needn't lead to complex philosophical debates about whether such traits are innate, or the thorny question of what 'innateness' even means (see Samuels, 2004), but can, on a practical level, offer a level of solidity to educational guidance. If phenomena such as interleaving hold across the broad landscape of human learning, then the role of educators and researchers is to identify their relevance to a particular context, and to make proposals for how to utilise them in service of educational goals. Doing so would arguably be more helpful to educators than asking them to accept that techniques 'work' without providing any further guidance about how or when.

I now move on to a second problem associated with researching memory's role in education – the issue of whether memory processes remain stable when scrutinised.

4.2.3 Whether memory processes are stable and amenable to investigation.

The second assumption under discussion is that memory processes are relatively stable and amenable to investigation and manipulation. As noted above, it may sometimes be assumed that memory processes can be measured and quantified. This is exemplified in many studies of long-term memory including some of those discussed so far in this thesis (e.g. Kornell, 2009; Roediger & Karpicke, 2006). It is also reflected in any formal examination where a judgement of what students know or can do is established via a one-off test.

Such examples illustrate a positivist position, whereby phenomena are seen as being objectively measurable and thus amenable to scientific scrutiny in a way that is separable from the observer or the specific interaction (Peca, 2000). However, the position is threatened by the tendency for memories to change as soon as they are investigated. While it may seem like common sense that a learner has a specific amount of knowledge and that this can be measured, simply making use of our memories changes their structure. One example that has already been discussed is that the process of testing a learner on a set of facts strengthens their memory for those facts (see Chapter 1).

This has major implications for research methodology; the process of studying memories typically also affects the object of the study; as Bjork (1975) put it, “an item can seldom, if ever, be retrieved from memory without modifying the representation of that item in memory in significant ways.” (p. 123). Spellman and Bjork (1992) have characterised this situation as a psychological analogue of Heisenberg’s uncertainty principle: “Any effort to take a reading of a subject’s current state of knowledge may alter that state of knowledge” (p. 316).

Indeed, for many years the benefits of testing were considered to be something of a methodological headache for memory researchers rather than a boon to educators (Karpicke, Lehman & Aue, 2014). Retrieval attempts had to be strictly monitored and minimised in experimental protocols, and covert/internal retrieval viewed as a source of bias in some experimental protocols (for example McCabe, 2008, suggests that covert retrieval and the resultant spaced practice affects performance on working memory tasks).

The fact that something in a state of flux does not in itself rule out a positivist framework for research. Unstable nuclear isotopes, for example, decay with a gradually slowing loss of energy in the form of radiated particles (an analogy for the LTM forgetting curve, although the specifics are very different). However, processes such as radioactive decay can be observed and measured without changing the speed or extent process itself. When it comes to forgetting, the problem does not lie in the changes to memory structure, but the fact that such processes change when observed. This doesn’t invalidate the positivist idea that such processes are subject to rational laws. However, it does impact on the main means of determining such laws – objective observation and verification (a key principle of science, founded on the philosophy of positivism, is that phenomena can be observed objectively, and that these observations can be used as the foundation of natural theories and laws which hold independently of the specific circumstances of the observation; Norris, 1985).

How do researchers address this problem? One way would be to minimise the extent to which testing is done during a research paradigm. As discussed, the extent to which repeated testing boosts memory has tended to be seen as useful thing and worth investigating in recent years (e.g. Agarwal, et al., 2012; Karpicke et al., 2014). But any methodological choice will have to be mindful that by asking a question that requires participants to access their memory, the thing under investigation will itself change. However, this leads to a paradox – how can the memory processes be investigated without testing the research participants?

Another option might be to embrace repeated testing, on the basis that if tests are frequent, each one will have less of an impact (a corollary of the spacing effect; see Dempster, 1989).

A third option would be to test in a way that samples the studied material rather than testing everything, so that some items are tested and others are not. However, as memories are connected to one another, there remains the possibility that testing item A will have an effect on an untested item B. This exact outcome is seen in the literature on 'retrieval-induced forgetting' (see for example Ciranni & Shimamura, 1999), where testing one item makes other items harder to retrieve. This is the case with competing items. Testing one item can make other items easier to retrieve if they are semantically connected, and the test of one thing causes another to spontaneously come to mind (an example of this can be seen in the DRM paradigm, when many participants falsely remember seeing the word *needle* when presented with a list that includes items such as *haystack, syringe, vaccination* but does not include the word *needle*).

Overall, this points to a broader problem with this assumption about memory: researchers or educators alike may view individual memories as 'items' which are stored in memory and then retrieved intact and in their entirety. However, Simons and Chabris (2011) note that this is a misconception about memory; the idea that LTM functions “like a video camera” was endorsed by 63% of participants among the general public in their study, but by 0% of memory researchers questioned at a 2010 meeting of the Psychonomic Society. In line with schema theory, memories are better seen as elements within interrelated structures, despite the problems that this brings to research.

Karpicke (2016) discuss the way this flawed idea of memory has permeated popular culture; an example that he discusses is the movie *Inside Out* (Docter, 2015) which portrays each memory as a glowing ball that is stored inside a girl’s memory bank. Each one can be lost/forgotten and can be affected by emotions – a reasonably accurate portrayal. However – inaccurately – each is portrayed as a separate item that does not connect to other memories or past knowledge schemas, and each can be replayed via a video camera inside the head of the lead character in the movie. It seems quite likely that such a view of how a memory functions will be shared by at least some educators and their students, and perhaps by some researchers as well.

A related problem that arises when investigating memory is the slippery nature of memory processes and their tendency to be affected by research tasks, leading to biases and unwanted interactions. Drawing on the general positivist position that characterises much scientific research, cognitive psychologists typically aim to isolate phenomena and study them in as pure a form as possible. In practice, this has often led to methodological approaches which involve isolating variables in laboratory contexts – outside variables are seen as a source of error, and must be kept constant or eliminated entirely in order to establish patters of cause-and-effect, and to test theories. Often memory researchers may seek to avoid bias by simplifying both the tasks studied and the context in which they are completed. A good example is the seminal work of Cepeda et al. (2008) on the spacing effect, where participants were tested on a set of 32 general knowledge facts without any broader context (for example, “*What European nation* *consumes the most spicy Mexican food? Answer: Norway*”, p. 1097), in order to test their later retention of the answers.

This problem connects to a broader problem with positivistic methodologies in education; it is often difficult to isolate variables and determine cause-and-effect in a meaningful way without simplifying a participant’s situation to the extent that it ceases to model an authentic educational experience. Isolation of variables can mean removing context without which the phenomenon under study may no longer exist in the same form. For example, in the study of homework efficacy, taking out the context (the child working autonomously, at home, surrounded by their own family members, when tired, etc) would fundamentally alter the situation, such that findings would be impossible to generalise to the real situation. That is to say, setting up an artificial model of homework (e.g. with a child working autonomously in a fake bedroom within a laboratory) would lack external validity, as environmental and social variables are intrinsic to the concept being studied.

Again, this links to positivist assumptions: the idea that memory processes can be studied in the abstract. A contrasting post-positive perspective – a framework that views the researcher as a participant in an interaction rather than an impartial observer of events – would suggest that such biases are inevitable if the research is to be meaningful. As such, rather than eliminating biases, such a perspective suggests that they should rather be embraced and accounted for in terms of the way we frame hypotheses and conduct a scientific analysis (e.g. Popper, 1963). When we are considering the applicability of memory processes to education, it seems especially important that factors such as classroom distractions, background noise, and the complexity of real-world study materials are not eliminated as ‘bias’, because such factors will inevitably affect the application of any such processes (Coe, 2020). Failure to account for the realities of the classroom could result in well-documented memory phenomena having no effect in authentic learning situations, or even being counterproductive (Küpper-

Tetzel, 2017). A further principle, then, would be to find ways of conducting a reliable scientific investigation while maximising the authenticity of the tasks, settings and materials used (Loyd et al., 2005).

4.2.4 Whether memory is open to scrutiny by its users.

The third assumption raised above is the issue of whether memory is open to scrutiny by its users. That is to say, can a person accurately reflect on what they remember and what they don't remember? And can they make accurate predictions about how well they will remember something in the future?

Again, there are multiple examples of this assumption in education. Many of the set of techniques associated with 'formative assessment' assume that learning is at least partly amenable to reflection. Such techniques aim to encourage feed-forward from the learning process, so that assessments form a springboard for future learning rather than simply a progress check (Bryce, 2019), and hence this approach is also called *assessment for learning*. However, many of the key techniques require pupils to self-assess (Black et al., 2004), for example via a traffic lights system whereby they highlight what they have or have not learned. In addition to the point raised in the previous section (that assessing what you know changes what you know), these techniques reflect an assumption that pupils are able to access their own knowledge.

Formative assessment is referred to by the General Teaching Council for Scotland in their standards for professional registration, which state that registered teachers should: "systematically develop and use an extensive range of strategies, approaches and associated materials for formative and summative assessment purposes" (GTCS, 2012, p. 17), and is widespread across other education systems, too. It can reasonably be considered to be a foundation of modern teaching practice.

However, again, there are reasonable research-based doubts about such practices. As discussed by Koriat (2000) and others, our judgements of what we know tend to reflect a feeling of knowing (FOK). They demonstrate a subjective, phenomenological sense that we know something, based at least partially on incomplete and implicit information. As such, they may be open to bias.

One such bias is that a FOK may be based at least partly on how easily information comes to mind, leading to the availability heuristic being applied. Several examples of this heuristic are discussed by Tversky and Kahneman (1974) in their article *'judgment under uncertainty: heuristics and biases'*. For example, people may assume that words beginning with the letter 'R' are more common than words with 'R' as a third letter, simply because it is easier to retrieve examples from memory. In fact, the reverse is true.

In the classroom, this may play out with pupils attempting to judge their own learning on the basis of how well they can retrieve a few facts (it would be a different matter if the self-assessment was based on a piece of work and used a marking scheme, as this is not a memory task). And in a classroom setting, if a student is asked "how well do you know this topic", it seems unlikely that will mentally retrieve most or all that they know about it before answering the question. If, for example, a student was asked "how many of the 45 American presidents could you name?", they would typically be able to give an estimate without taking the time to engage in retrieval of all of their relevant knowledge. At most, a few items would spontaneously come to mind, biasing the judgement via availability.

Additionally, any such judgements will inevitably – unless judgement is delayed – reflect the learner's current state of knowledge (*performance,* rather than *learning* – see Chapter 1). If made very soon after study, the judgement may reflect the contents of working memory rather than long-term memory (Nelson & Dunlosky, 1991). Even a judgement that is delayed by ten minutes or more but remains within a single lesson or study session will reflect LTM performance at an early stage, before forgetting has begun to make a major impact. As such, they relate to the top left of Ebbinghaus's forgetting curve; if a learner accurately concludes that they 'know' something, this judgement may well fail to take account of subsequent forgetting. And due to the stability bias discussed in Chapter 2, they tend not to take adequate account of forgetting, either[12].

As a methodological problem, this issue clearly interacts with timescale. As discussed in Chapter 1, research by Nelson and Dunlosky (1991) compared judgements of learning for word pairs either immediately or after a delay of ten minutes, and found such judgements to be much more accurate if delayed. However, there is some disagreement about why such a judgement would be superior. Most researchers focus on retrieval processes, for example noting how the difficulty of delayed retrieval is more akin to the difficulty of a later exam or practical use, or that the delay prevents the use of immediate WM recall as discussed above (see Metcalfe & Finn, 2008). However, Son and Metcalfe (2005) postulate that familiarity with the cue – in the absence of any actual retrieval – is often the basis for such judgements. For example, a pupil may see a question like "how well do you know Shakespeare's Othello" and, recognising the title of the play, judge their state of knowledge on the basis of cue familiarity and/or on their recollection of past test performance (Finn & Metcalfe, 2008).

A final point to be made is that many of these issues with flawed judgements and bias can apply to classroom practitioners, too. This may be an argument for recommending that teachers gather classroom data to inform their judgements, rather than making decisions on the basis of

[12] It is beyond the scope of the current thesis, but these flaws in a learner's judgement of their own knowledge can have an impact on independent study and revision.

intuition, for such intuition is likely to be biased by short-term performance of the class, therefore failing to take account of forgetting (Firth, 2019a).

This section has raised three further principles. First, methods for assessing student competence will best be done directly, for example by questioning them or giving them a task to do, rather than indirectly, for example by asking them about their confidence with the material. It can't be assumed that participants will have an accurate judgement of what they know or do not. Secondly, it is important to build in some form of delay when testing learning to ensure that performance does not merely reflect the content of working memory, and we should also be cognisant that even a delayed performance does not fully reflect learning, but rather LTM in its initial stage. And finally, these biases apply to teachers, too. A richer, more contextualised classroom task may make it easier for them to judge their own performance, but certain biases can be expected. Due to the obscure nature of human memory, they will neither have perfect insight into their own past actions nor will they be able to judge their learners' progress with perfect accuracy.

4.2.5 Overview, and a comment on post-positivist perspectives.

As I have suggested, there are considerable reasons to doubt the three assumptions stated at the start of this chapter; we can't state with any confidence that memory processes are universal to all learners, static and amenable to investigation, and open to reflective scrutiny by its users.

From some of the criticisms I have made of a positivist approach to studying memory, it might be assumed that I am instead endorsing a post-positive perspective. However, some of the main arguments for post-positivism do not apply here. I am not advocating a nuanced reading of classroom performance on memory-related tasks which take observer bias into account, or viewing human behaviour as culturally relative. Performance on memory-related tasks are fairly objective, and any bias that might be brought in by a culturally biased selection of tasks (for example) are a product of the education system and those who create the curriculum, not of the education researchers who choose to maintain validity by using authentic classroom materials to investigate learning.

Indeed, some of the key research methods associated with post-positive perspectives in education suffer from the same problems raised above. An interview or focus group methodology in which learners are asked about the success of their study habits, for example, would be undermined by the fact that learners are often unable to accurately appraise their own learning accurately. It will be important to remember that while such methods are useful for asking about a student or teacher's *beliefs* about their learning (or about learning in general), they cannot provide an accurate insight into actual progress. Even episodic recall of one's own past learning experience is likely to be flawed, as demonstrated by the ample literature on eyewitness memory (see Loftus, 2019, for a recent overview; see also Firth, 2020 – reproduced in Appendix 7 – for a consideration of how these issues apply to the process of lesson observation).

I do endorse one key aspect of post-positivism, however, which is to say that the biases associated with real world settings should not be eliminated, but must instead be accepted and accounted for. When studying educational processes, minimising bias often fundamentally transforms the task, for example by studying the learning of abstract shapes within a laboratory setting in lieu of real curriculum material in a classroom setting. For the aims of this thesis to be fulfilled, a correction of this approach is necessary, with the focus of investigation more on the applicability of findings than on their freedom from random error. Doing so will involve embracing the use of classroom settings for experimentation.

While the effect of experimenter biases and extraneous variables in research are well understood, a more frequently misunderstood point raised in this section is that the nature of the subject matter under investigation is fundamentally hard to access and open to being distorted by the research process itself. Accepting this problem does not mean that relativism has to be built into our methodology. However, it does necessitate a nuanced, evidence-informed approach that takes account of human metacognition, and in particular the flaws with metamemory. People often simply do not know what they know, they frequently don’t access their knowledge when asked to judge its extent, and they routinely underestimate the impact of forgetting and the benefits of practice. And these problems can apply to educators as well as to their students.

Having set out this psychological context, I will now, in the following section, outline the implications of these issues for the methodological choices that must be made when studying desirable difficulties. Firstly (section 4.3) I consider the use of experiments for direct investigation into the phenomena of interest, and highlight the considerations that should be taken account of. I then analyse the use of self-report methods both for students, and – potentially – for teachers (section 4.4). I also address issues relating to selection of participants and to research ethics (sections 4.5 and 4.6).

This leads to a specific plan for the studies which follow later in this thesis. I also briefly consider some of the less-frequently used methods, as an awareness of the broader palette of research methodology can be important when reviewing findings and setting out priorities for future investigation, as is covered especially in the General Discussion chapter (Chapter 7).

I will attempt to synthesise the problems inherent in the methods in general (for example, the idea that questionnaires are prone to misunderstandings or social desirability bias) with the specific memory-related issues discussed above. In identifying specific strengths and barriers to using each method as well as the key variations and options that must be considered, I will review the principle insights that have been gained from their use with specific reference to the interleaving effect.

4.3 Experiments

4.3.1 Experiments – introduction.

Given the limited insight which learners have into their own memory processes as discussed above, direct experimentation – either used alone or in combination with self-report – is a key research method. A comparison can be made here with social psychology (e.g. research into conformity, prejudice, and altruism), in which it is often assumed that participants' actions will differ from their intentions, predictions and beliefs about their own actions. And in a similar way to that field, self-report (metacognitive judgements of one's own learning) can be instructively contrasted with actual learning performance on an experimental task.

As mentioned above (see Section 4.2), it is important to avoid positivist over-simplifications, especially in applied work. A useful body of experimental research has studied concepts such as memory and category learning in abstract contexts and controlled circumstances, and an aim of the present thesis is to both extend and apply this. Doing so will require both control of variables to avoid confounds as with any experiment, and careful consideration of how such variables may play out in realistic circumstances, such as on classroom tasks.

A major issue with any test of interleaving or spacing is how the order and timing of learning and practice are controlled. This needs to be realistic enough to mimic a classroom learning situation, yet sufficiently controlled to eliminate confounding variables. Task phases and control are discussed here, and the authenticity of task stimuli is addressed in the following section (Section 4.3.3, below).

4.3.2 Experiments – procedure and design.

Most experiments into interleaving involve one or more of the following: a learning/input phase, a test phase, and one or more interim/delay/distraction experiences, as follows.

4.3.2.1 Learning phase.

Interleaving experiments feature a learning phase in which stimuli – typically, examples of the concept to be learned – are either interleaved or blocked. These stimuli are typically being studied by participants for the first time (and are often chosen to be novel/unusual to minimise prior exposure), but in some studies (e.g. Abel & Roediger, 2017; Rohrer & Taylor, 2007), material is learned in a more traditional format, and a follow-up practice phase is interleaved. This distinction can be referred to as *interleaved learning* vs. *interleaved practice* (see Chapter 3). I will continue to focus on the former.

The key manipulation (i.e. independent variable) studied in interleaving is the presentation of these items/examples in either an interleaved or blocked order (as described in previous chapters; see for example Chapter 3, Figure 3.1). Typically, the interleaved condition is systematically alternated, but there are alternatives. One is the use of ‘contrast’ or ‘hybrid’ sequences, which present blocks followed (in varying degrees of gradations) by an interleaved sequence, for example: AAAAAA–BBBBBB–ABABAB (e.g. Clapper, 2015; Yan et al., 2017). Carvalho and Goldstone (2015) discuss the use of blocked sequences which include relatively less contrast (alternating 25% of the time) and interleaved conditions which use relatively more (alternating 75% of the time). Overall, then, there are multiple options, but in general blocking implies contrast between consecutive examples, and interleaving implies contrast.

In some studies (e.g. Rawson et al., 2015) items are presented one at a time as a series, while in others they are presented as a simultaneous set (e.g. Wahlheim et al., 2011). Either way, the interleaved condition features an alternating order of different types of item, while in the blocked condition, the same type of item appear together or consecutively (see Figure 4.1).

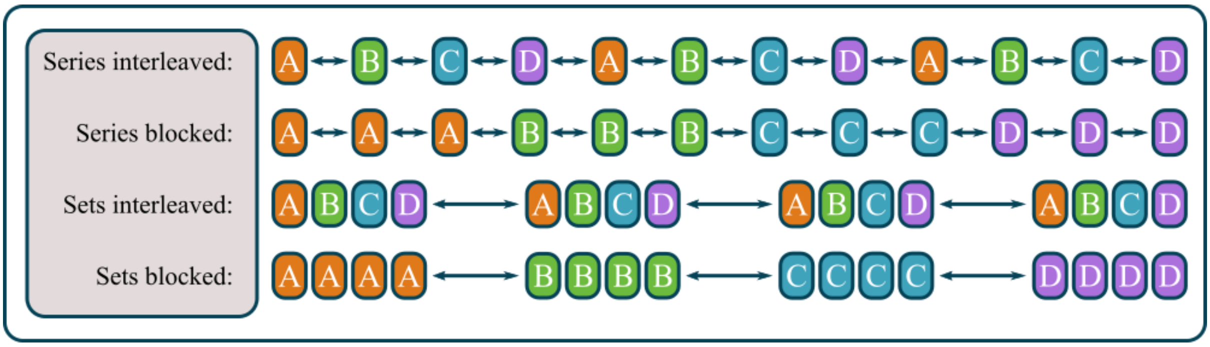
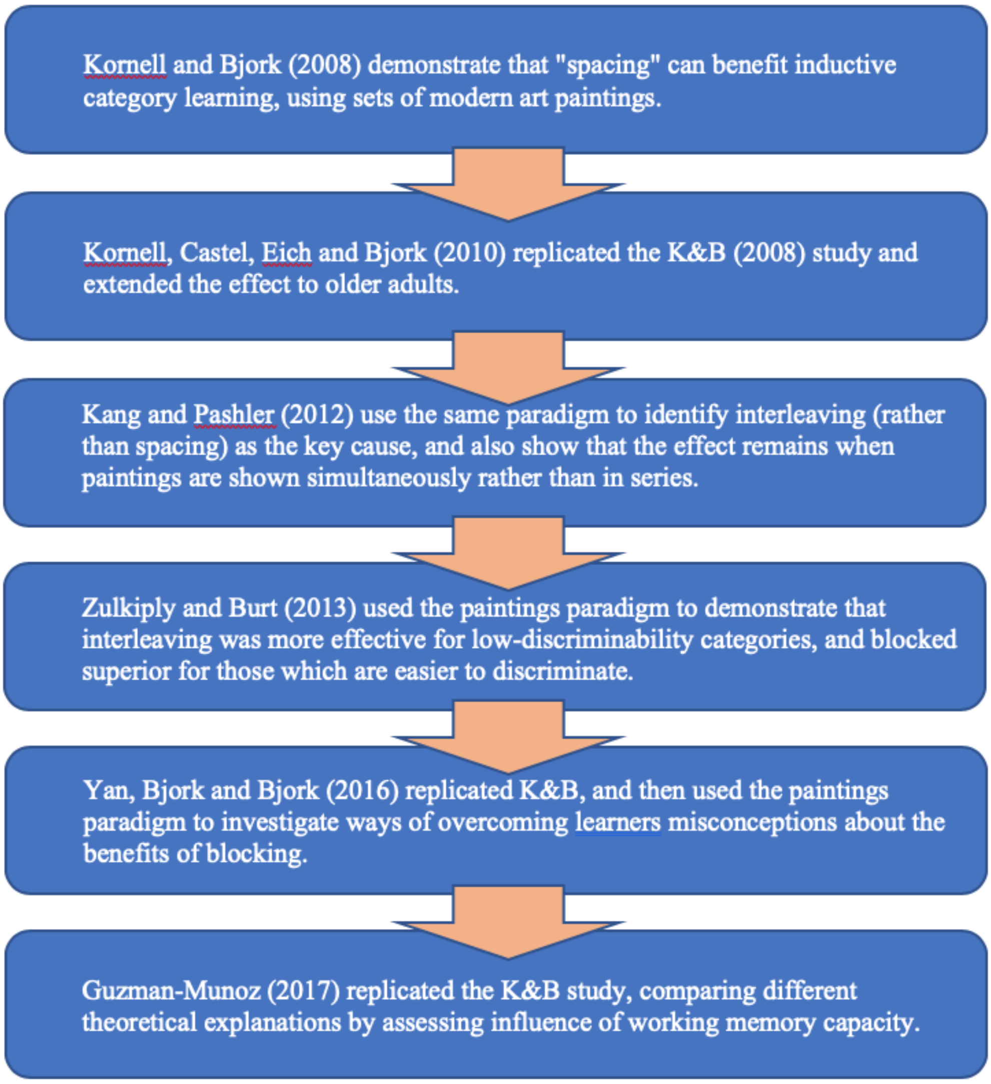


Figure 4.1: Interleaved vs. blocked presentations in either a simultaneous set or a series. Dashes represent a time delay; the delay between individual items or sets of items can vary.

The implications of the set vs. series options are discussed by a number of researchers. Showing items as a series potentially provides a greater challenge to learners, because they have to retain previous items from the series in working memory if they are to engage in discriminative contrast between current and older items, i.e. by mentally noting and learning from differences between these items (Noh et al., 2016). If items are provided together in a set, that task becomes easier. As such, increased spacing of a series could moderate or eliminate the interleaving effect. This was the conclusion of Birnbaum et al. (2013), who found that spacing and interleaving were not additive in effect, and that greater spacing actually reduced the benefits of interleaving despite its potential to boost memory. An important design decision for researchers, then, is how to time the delays between one interleaved item and the next. It is worth noting, in this context, that Guzman-Munoz (2017) found no correlation between WM scores and performance on the massed examples used, but a weak correlation was found with performance on the interleaved items. Overall, then, it may be advantageous to minimise the load on working memory, and to ensure that items are presented as close together as is possible.

The interleaving study that has been most influential in terms of methodology is Kornell and Bjork's (2008) study of memory for sets of paintings by modern artists. The researchers showed participants sets of 6 paintings as a series, with each painting was presented separately for 3 seconds rather than in a set (Experiments 1a & 1b). Figure 4.2 (overleaf) summarises key variations of the Kornell and Bjork paintings paradigm.

This paradigm has also found a key boundary condition of the effect: Zulkiply and Burt (2013) found that items that where category members were less similar and therefore easier to discriminate it was more beneficial to show them in a blocked rather than interleaved order, presumably because differences were obvious even without the items being adjacent in the series. Guzman-Munoz (2017, Experiment 2) also modified the timing, using 5-second presentations, varied to 10-second presentations to increase overall spacing within certain conditions.



Kornell and Bjork (2008) demonstrate that "spacing" can benefit inductive category learning, using sets of modern art paintings.

Kornell, Castel, Eich and Bjork (2010) replicated the K&B (2008) study and extended the effect to older adults.

Kang and Pashler (2012) use the same paradigm to identify interleaving (rather than spacing) as the key cause, and also show that the effect remains when paintings are shown simultaneously rather than in series.

Zulkiply and Burt (2013) used the paintings paradigm to demonstrate that interleaving was more effective for low-discriminability categories, and blocked superior for those which are easier to discriminate.

Yan, Bjork and Bjork (2016) replicated K&B, and then used the paintings paradigm to investigate ways of overcoming learners misconceptions about the benefits of blocking.

Guzman-Munoz (2017) replicated the K&B study, comparing different theoretical explanations by assessing influence of working memory capacity.

Figure 4.2: Key examples and variations of the Kornell and Bjork 'paintings' paradigm.

4.3.2.2 Test phase.

The other key element of any experiment into learning or memory is a test phase. Typically, in memory experiments, these involve either *recognition* or *free recall*, or both. One test of recognition that is widely used in the literature is the old–new recognition test (e.g. Mulligan & Osborn, 2009), where two or more stimuli are presented to the participant and they must judge which one they have seen before. This is a standard technique in the eyewitness memory literature to identify faces, and also works in the paintings paradigm; it was used by Kornell and Bjork (2008, Experiment 3). Accuracy can then be judged on the basis of hit rate minus false alarm rate. For example, if 10 pairs of pictures are shown and the participant gets 6 correct and 4 wrong, the 4 would be subtracted leaving an accuracy rate of 2/10 or 20%.

Testing usually involves free recall in experiments testing the spacing effect, though a recognition task can be used. In interleaving experiments, it is typical to present items and ask participants to categorise them, and these items can be either novel, repeated from the learning phase, or a mixture of the two. Studies of interleaving can thus test both memory and transfer of learning, and focus on learning and categorising new concepts rather than retrieving specific items (see Chapter 3 for more about the role of memory and transfer in this research area). This is because the interleaving effect, at least as widely demonstrated, relates to inductive learning of categories or skills.

As such, the educational implications of spacing and interleaving are different; the former has more potential for practice of fundamental facts and elemental knowledge (for example spellings, place names, dates, number bonds), while the latter has more potential for learning cognitive skills that might transfer to new situations or even new domains. This is important, because as Barnett and Ceci (2002) have argued, the ability to transfer learning outside of the academic context (sometimes termed 'far transfer') is the fundamental justification of educational efforts and spending as a society.

In any test phase, it is important to remember that retrieval or practice is also a learning event, which can add to or change prior learning (Bjork, 1975; Kornell & Bjork, 2008; see Section 4.2, above). For this reason, the experimental design needs to avoid confounds such as participants in one condition being tested more than those in another condition.

As noted earlier, some studies (e.g. Kornell & Bjork, 2008; Zulkiply et al., 2012) add a *question phase* after the test phase, to determine metacognitive processes such as beliefs about whether spacing/interleaving was effective or not, and some also use a *pre-test phase* to determine prior levels of skill or knowledge at the relevant task. For metacognitive comparisons between interleaving and blocking in a question phase the experimental design generally needs to be within-participants, although in an exception to this rule, Eglington and Kang (2017) asked participants to predict how they would do on a future test on the basis of their performance, and compared this between participants.

4.3.2.3 Fillers, distractors and delays.

The delay between task and test is a critical issue in the memory literature. For studies of the spacing effect, the typical methodology involves a delay between task phase and practice phase, with a constant delay from both practice phase, and then the same test phase for both conditions (e.g. see Figure 4.3).

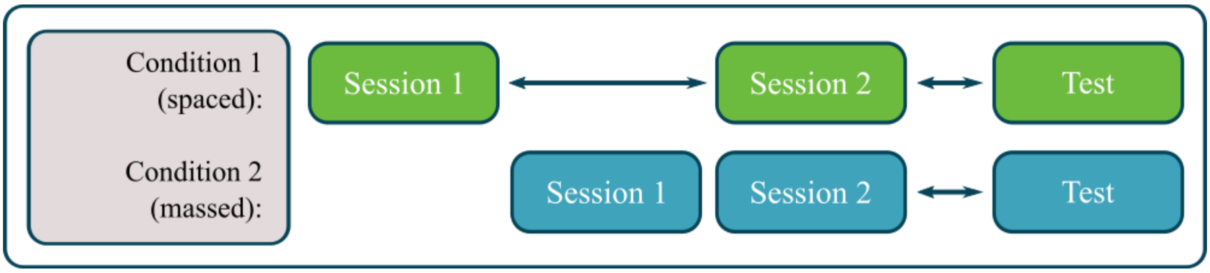


Figure 4.3: Structure of a typical study of the spacing effect.

If the delay is not kept constant between a spaced and a massed condition, there is a confound in favour of one condition. For example, imagine that a group of students all study on Day 1, and then a ‘massed’ group re-study on Day 1 while a ‘spaced’ group re-study on Day 5; all are then tested on day 10. In this scenario, the spaced condition benefit from a much shorter retention interval. This flaw is seen in a number of studies such as Grote (1995), but it is easily avoided by following the outline shown in Figure 4.3, with the gap between the two initial sessions varied, but the pre-test retention interval remaining constant.

A confound between spacing and interleaving can also occur, whereby the greater delay between examples from one category in an interleaving condition results in a spacing effect for that condition. A possible design feature to overcome this is to use filler tasks, thereby modifying the spacing between one target item and the next. For example, in Birnbaum et al. (2013), images were interleaved with unrelated trivia questions. Kang and Pashler (2012)'s study of interleaved vs. blocked paintings also used filler tasks; their study phase took 6.5 minutes for the standard interleaved condition, but almost 20 minutes for the spaced-plus-interleaved condition of their study as a result. Presentation time of the images was kept constant (5 seconds per painting/set of paintings).

Distractor tasks are widely used to reduce the immediate recall of the content working memory and therefore increase the validity of measures which are assumed to tap into WM. These can be quite brief; in Kornell & Bjork's paintings study, participants were given a 15s distractor task involving counting backward by 3s from a three-digit number (i.e. the distraction task from the *Brown–Peterson test* of short-term memory). Other studies that used word lists have included unrelated/untested words within a series as buffer items (e.g. Cho & Neely, 2013). In general it would seem wise, bearing in mind that working memory is a multi-modal and flexible system (see Chapter 2), to use verbal distractors in studies of verbal learning (in order to occupy the verbal working memory store), and to use visual distractors in studies that use visual stimuli (in order to occupy visuo-spatial working memory), and consideration should also be given to ways of occupying other components of working memory (e.g. the 'episodic buffer' as defined by Baddeley, 2000) where appropriate.

In any study of memory or transfer, it may be necessary to incorporate a longer time delay than is possible within a typical laboratory study or lesson-based experiment. Participants may therefore have to return after a specific period of time. Such designs can lead to attrition of participants, and so oversampling may be necessary. In the interleaving literature, surprisingly few studies use educationally-relevant timescales such as studying one week with a test the following week (see Chapter 3 for some exceptions). This is interesting given that the desirable difficulties literature has revealed important interactions between delay and memory processes, with the benefits of spacing appearing after tests which are scheduled days or weeks after initial study (e.g. Bird, 2010; Roediger & Karpicke, 2006). While this is an area which warrants further study, the evidence available so far has not found significant interactions between interleaving and a time delay (Carvalho and Goldstone, 2014b; Carvalho and Goldstone, 2017; Zulkiply, 2013), suggesting both that interleaving is not an artefact of the spacing effect and that any advantages found in immediate study should persist over longer spells.

For this reason, it is acceptable for classroom studies of interleaving (at least in the initial, exploratory stage) to use only short time delays between task and test.

4.3.3 Experiments – materials and procedure.

4.3.3.1 General procedures.

As mentioned above, several interleaving studies have used modern art paintings as stimuli, following the work of Kornell and Bjork (2008). Another popular option is to use images of animals such as birds or butterflies, as first introduced by Wahlheim et al. (2011). Several other studies, including the work of Carvalho, Goldstone and colleagues (e.g. Carvalho & Goldstone, 2014a), made use of abstract shapes with verbal labels. The verbal labels in such studies should be carefully chosen to equate the number of syllables and final sound, and to be English-like without strongly resembling real words (Carvalho & Goldstone, 2014a).

A comparable approach was taken by Clapper (2015), whose experimental stimuli were lists of verbal characteristics supposedly describing fictitious species of insects or trees, for example habitat type (forest, wetland, etc), Latin name and country of origin, and year of discovery. This study therefore eschewed images in favour of verbal information about the organisms studied.

However, many such materials will clearly lack face validity to students. One option to overcome this and avoid motivational problems biasing results is to provide a 'cover story' to participants, explaining why they are being asked to categorise the items. For example, Carvalho and Goldstone (2014b) told participants that a recent expedition to Mars had recovered cells from alien organisms, and that these 'cells' (i.e. the abstract blobs used as stimuli for interleaving) could be categorised solely on the basis of their physical features. However, it is worth noting that such cover stories may not be required if in studies of students engaged in syllabus-relevant learning tasks.

In addition to natural species, many other types of item or task have been investigated in studies relevant to interleaving (see Table 4.1), though not all of these are relevant to concept learning and some were therefore excluded from my systematic review (Chapter 3).

Table 4.1: Variety of task materials used in experiments relevant to interleaving.

Stimulus type	Example study
Modern art painting	Kornell & Bjork (2008)
Music performance	Stambaugh (2011)
Chess board recognition	Lavis & Mitchell (2006)
Social science concepts	Rawson et al. (2015)
Psychiatric diagnosis	Zulkiply et al. (2012)
Medical diagnosis	Hatala *et al*. (2003)
Abstract shapes	Carvalho & Goldstein (2014)
Maths problems	Rohrer & Taylor (2007)

Statistical concepts	Sana, Yan & Kim (2017)
Foreign language vocabulary	Carpenter & Mueller (2013)
Foreign language grammar	Phun et al. (2017)
Kayak skills	Smith & Davies (1995)
Throwing skills	Pigott & Shapiro (1984)
Golf skills	Porter *et al*. (2007)
Physical shape of people	Kalish et al. (2011)
Animal conditioning	Blair & Hall (2003)

As with the spacing effect, the serial position effect and many other memory phenomena, the scope of interleaving is broad (note that the motor learning research typically refers to the effect as ‘contextual interference’, but the experimental design of each of these studies constitutes a blocked vs. interleaved schedule). It remains to be established whether there is a common psychological or neurological basis to these commonalities. It seems possible, for example, that interleaving of concept learning has a different cognitive basis than interleaving of motor tasks (see Chapter 2) and would therefore be associated with different epiphenomena and boundary conditions, though it would be parsimonious to assume that brain areas representing metacognitive phenomena such as beliefs about learning are common to different types of learning.

One limitation inherent in many of the studies of category learning mentioned above is that they focus on the intrinsic features of category members. However, as discussed by Goldwater and Schalk (2016), education is often more concerned with *relational categories*. As they put it, relational categories:

“*…are defined by the relations among entities. For example, the category catalyst classifies molecules by their role in effecting changes of state in other molecules; force is defined by the multiplicative relationship between mass and acceleration*.” (p. 730).

According to Goldwater and Schalk's analysis, such categories present a potential area of common ground between the category research of cognitive psychology and the school-based research of educational psychology. Arguably, though, another such commonality is the area of skills. A number of the studies listed above relate to motor skills, and a few relate to mathematical skills. There is a gap in the literature, as noted in Chapter 3, when it comes to research into higher-order cognitive skills such as analysis, evaluation, application and analogy, despite the fact that these feature prominently in many academic courses. Such skills are undoubtedly of interest to the education sector, but are a major area of interest to cognitive psychologists as well (e.g. Agarwal, 2019; Gick & Holyoak, 1980; Holyoak & Thagard, 1989; Koedinger & Roll, 2012). They are often termed 'transferable skills' in education, but among psychologists the extent to which they can transfer from practice to use is debated (e.g. Anderson et al., 1996; cf. Lave, 1991).

Skill learning and transfer are potential areas where the research into interleaving can be productively applied to school-relevant tasks, and will be a priority for the current series of investigations.

4.3.3.2 Task difficulty and competing options.

In many tasks, particularly those where items have to be identified from among a set of alternatives, the number of items presents a computational problem whereby the greater the set of options, the more difficult the task (Miller, 1956). In many of the interleaving experiments, therefore, it is important to consider how many items (or categories of items) are interleaved. For example, the Kornell and Bjork (2008) study interleaved the work of 12 modern artists, while the follow up study by Kang and Pashler (2012) used only 3 in order to distinguish between the effects of spacing and inductive learning by reducing task difficulty while keeping spacing constant.

In a task that involves both categories and category members, both of these could vary in number. For example, Singer (2009) uses 10 categories (colours, US states, types of bird, etc) together with forty common examples of each. Zulkiply et al. (2012) used 6 psychopathological categories, with 6 case studies in each. In terms of category members, the number of possible options in a natural category is likely to be at least as important as the number of examples presented, as this will determine the *problem space* of the task (Newell & Simon, 1972). However due other invariant influences on task difficulty (such as working memory capacity), increasing the competing options is likely to make a task harder, but only up to a point.

Also, as with any task and test, difficulty can be determined in part by the similarity of the competing options. If two examples are very alike, participants will, on average, find the task harder. It may be important to use pilot testing in order to determine a level of difficulty that avoids floor or ceiling effects (which also interacts with participant characteristics - see Section 5). At least two studies – Zulkiply and Burt (2013), and Carvalho and Goldstone (2014) – have shown an interaction between interleaving and the discriminability of category members. The difficulties of verbal exemplars/category members can be controlled by using category norms where words are arranged in order of frequency (e.g. Battig & Montague, 1969), or by referring to real world features such as the size or population of a country (with larger countries tending to be better known than smaller ones). Where words are used (e.g. Mulligan & Osborn, 2009), new-old tests can be weighted by word frequency. However, determining discriminability of visual items such as paintings or conceptual items such as examples of social science concepts may be more subjective, as well as depending on participant prior experience (see Section 4, below). Some studies (e.g. Di Vesta and Peverly, 1984) have created artificial concepts to test learning in a way that can’t be distorted by prior knowledge, but such an option does need to be weighed against the desire for mundane realism (see below). A further possibility is to interfere with the use of prior knowledge by using information which is distorted in some way, so as to be unrecognisable. This approach was used in the vignettes/case studies used by Zulkiply et al. (2012).

Despite looking at this issue in the systematic review (Chapter 3), the effects of category size remain unclear, in part because few studies have directly compared larger versus smaller category sizes. As noted in that chapter, there are pragmatic reasons for wanting to reduce the number of examples in a school classroom while still achieving learning.

On the basis of transfer, the overall size of the category in terms of the number of category members (e.g. all amphibians) needn't matter much on the assumption that learners become able to transfer their learning (having studied the features of some amphibians, they can correctly categorise other ones). Therefore, on a practical level, teachers may decide to select the smallest subset of a category that was still sufficient to allow for discriminative contrast. Again on a practical level, the number may be chosen on the basis of professional judgement of what students are likely to mix up. In the example above, learners may be in danger of mixing up amphibians with reptiles, fish and insects, perhaps, but not with birds or mammals. Teachers may therefore interleave examples from those four easily-confused categories.

A key piece of evidence is that some studies (Dobson, 2011; Carvalho & Goldstone, 2017) have interleaved examples from as few as two categories (clearly, we can't interleave examples of just *one* category). This suggests that there is no classroom obstacle to using interleaving even when the number of options is small. From a research design point of view, though, interleaving items from just two categories would lead to a test phase with a 50% chance of success by guessing – reducing the sensitivity of the test phase, while very high numbers of categories will increase the study time required.

These considerations lead to the conclusion that for a field study of interleaving, a number of categories in the region of 3–5 would be optimal. With authentic learning materials, though, the precise number will be constrained to some extent on the material itself.

4.3.3.3 Feedback.

Typically, learners are given feedback on their performance, and the trials thus act as a learning situation. Only a few studies (e.g., Billman & Heit, 1988; Clapper, 2015; Clapper & Bower, 1991; Michalski & Stepp, 1983), have focused on spacing and/or interleaving sequencing effects in the absence of feedback. In a school setting, however, there is typically little feedback to a learner during a presentation of information or during a test itself.

In order to main authenticity of the learning situation, it would be useful for the present study to minimise item-by-item feedback. This is not generally provided during classroom learning.

If student attention is focused on contrasts, an advantage of interleaving should still be apparent. If so, the findings will have implications for a great many situations (e.g. lectures. written materials, slideshows, verbal remediation to students) where examples might be interleaved (or blocked) without eliciting feedback from students.

4.3.3.4 Apparatus/media for presenting learning stimuli.

All experiments that involve memory and learning must involve some way of presenting authentic or simulated learning materials. There are various media through which such items can be presented. Some experimental tasks, such as reading a text in a classroom setting, maintain a high level of mundane realism when compared with common classroom practices.

Others, such as the use of on-screen electronic flashcards, may not be associated with the typical, teacher-led classroom, but may relevant to alternative educational experiences such as revision apps or blended learning.

Table 4.2 (overleaf) breaks down the key apparatus and stimulus combinations that have been used in interleaving and spacing research. As can be seen, computerised presentations predominate, in part because they allow greater control of timings and sequencing. Popular options include text, word pairs and categorisation of images with/without labels.

One option that does not appear in Table 4.2 – due to its absence from the memory literature – is the use of virtual environments. Some other fields have used *Minecraft* or similar computer games to teach electronics (e.g. Short, 2012), and these have potential to test or demonstrate memory principles, too. Toscano et al. (2015) argue that sandbox-type games can achieve both high external and internal validity, as they can be controlled but yet stimulate authentic behaviours and use of objects and language. In terms of categorisation, Minecraft features natural objects and animals, although the range of exemplars is limited by the program itself (for example, there are fewer than thirty types of animal overall, and no sub-species), although it is possible to modify the game via coding. Minecraft also features various 'biomes' i.e. types of natural environment (desert, tundra etc), and these could be presented either interleaved or blocked to a participant.

For now, though, it would be more straightforward to stick to one of the well-established approaches, such as presenting images or short text on screen. Although particularly well suited to the laboratory, such an approach can also be applied to classroom settings.

Table 4.2: Major procedure options in interleaving and comparable studies.

Main type	Subtypes	Example 1	Example 2
Laboratory studies	*OBJECTS/ IMAGES*	*Kornell & Bjork (2008) – modern art*	*Eglington & Kang (2017) – images of molecules*
	TEXTS	*Linderholm et al. (2014) – textbook sections*	*Sana et al. (2017) – textbook sections*
	SHORT VERBAL EXAMPLES	*Rawson et al. (2015) – psychology definitions*	
	WORD LISTS	*Karpicke & Bauernschmidt (2011) – Swahili-English word pairs*	*Kornell (2009) – English synonyms*
Field studies	*OBJECTS/ IMAGES*	*Wilson (2016) – unusual objects on screen*	
	VIDEO LECTURE	*Risko et al. (2012) – video in classroom*	*Mueller & Oppenheimer (2014) – TED talk*
	REAL LECTURE/ CLASS	*Küpper-Tetzel et al. (2014) – real language classes*	*Rohrer et al. (2015) – maths teaching followed by interleaved or blocked practice*
	TEXTS	*Blunt & Karpicke (2014) – texts in classroom*	
	PHYSICAL ACTIVITY	*Cook et al. (2008) – use of gesture in maths classes*	*Bahrick (1979) – real world locations and the spacing effect*

4.3.3.5 Mundane realism.

Although the research studies in Table 4.2 are broadly divided into lab vs. field categories, in practice the distinction is often subtle when an educationally relevant task is done under controlled circumstances in an authentic or realistic setting. For example, the study by Sana et al. (2017) was conducted individually under controlled circumstances and is therefore listed in Table 4.2 as a lab experiment, but the task had high ecological validity (learning from a textbook) and the mundane realism of the setting (studying alone) is sufficiently high that the same learning processes could easily have taken place within a school or library. A similar point can be made about many of the computerised and video tasks: they simulated real classroom experiences and maintained high levels of external validity, including face validity to learners.

Having said this, there are cases where the level of control and/or the options possible experimental manipulations in a field experiment are considerably reduced compared to the laboratory alternative. The following quote from McDaniel et al. (2013) indicates some of the compromises that can be incumbent on field researchers working in partnerships with schools:

> "*As guests in the classrooms, we were unable to substantially alter the normal classroom practices...doing so would have required the teacher to significantly change her lesson plans and daily lecture content...unlike in the laboratory, the amount of target material could not be increased to accommodate additional within-subjects conditions*" (p. 362).

In terms of mundane realism, a key issue is not just how materials are presented (as learners in real classrooms could experience any of the media shown in Table 4.2), but how authentic the learning items themselves are. Are they, for example, derived from real school syllabi, from tasks and worksheets used by teachers, or from authentic textbooks? Such materials have high levels of authenticity, contributing to the overall mundane realism of the experimental situation. Thus, the mundane realism of the task could be seen to be a function of both medium and content. A silent reading or teacher-lecture task may be realistic in one sense, but if the learning material so delivered is highly artificial and contrived, the situation as a whole will still lack mundane realism. For this reason, some studies (e.g. Sana et al., 2017; Rawson et al., 2015; Mueller & Oppenheimer, 2014) have explicitly chosen to use freely available materials that are already useful for learning, such as TED talks and textbook extracts. As well as increasing face validity (with a potential impact on participants' motivation, attention and effort), such choices make it easier to make strong recommendations for practice on the basis of experimental findings.

A high priority for the current research, then, is to employ school-relevant materials as task materials in studies of interleaving or other desirable difficulties.

4.4 Self-report Methods

4.4.1 General issues with self-report.

Arising from the earlier discussion of the 'three problematic assumptions' (see Section 4.2), one immediate issue with direct, non-experimental methods of investigation such as interviews and surveys in educational contexts is that they are open to bias on the basis of the learners' (or teachers') flawed assumptions about how learning works. All questionnaires and interviews depend on insight, honesty, and accurate recall on the part of the participants. However, some forms of self-report may be more prone to error than others, due to the fact that learners don't always have insight into their own mental processes, even if they are aware of the products or endpoints of those processes (Nisbett & Wilson, 1977). Possible sources of error that arise from the research literature into this problem include:

- Memory. People are prone to forgetting, and therefore participants who are asked to report what they did on past occasions may be prone to unintentional errors. In addition, people's memories are subject to distortion based on their present beliefs and goals (Ross, 1989), and the questioning process itself may corrupt earlier memories or cause false memories (Loftus, 2000). This may be a barrier in many forms of self-reported research into education, such as questionnaires which ask learners to recollect and self-report how or for how long they studied on a task. Memory can also interact with identity; Ward and Wilson (2015) conducted a study where students were asked about their moral conduct "*all the way back at the beginning of the term*" (distant past condition) or alternatively "*in the recent past, at the beginning of the term*" (near past condition; p. 1169). This simple manipulation asked for the same information, but varied the sense of psychological distance from one's current identity. More negative judgements of participants' own moral behaviour were found in the distant past condition.
- Honesty. There may be occasions where research participants may be motivated to modify their answers due to their reading of the research task as a social situation. There can be *demand characteristics* inherent in such situations; teachers may be prone to being "good participants" in the sense of trying to help a researcher achieve the (assumed) desired outcome (Orne, 1962). Closely related to this issue, participants may give inaccurate information even when motivated to be honest due to a combination of possible cognitive biases. Most notable is the *self-serving bias* - a social motivation to present the self and one's actions in a good light (Heider, 1958). This could lead, for

example, to a teacher reporting that written work had been set in class for the purpose of consolidation, when there may have been other factors influencing the decision, such as a desire to keep learners quiet. However, Miller and Ross (1975) have argued that self-serving bias is not a major factor in cases of attributing one's errors, and self-report protocols which involve anonymity and confidentiality may also reduce the impact on any self-serving bias by removing a perceived audience or the possibility of negative consequences.

- Insight. Even when participants can be assumed to be reasonably accurate and honest in reporting on their learning, they may lack insight into the processes being studied. For example, learners may be asked 'How do you learn best', and may offer a response such as 'I learn best via watching videos', but such opinions may be more a reflection of cultural assumptions than evidence of real insight into their own learning (see Section 4.2, above). Learner attitudes often seem to be affected by misconceptions such as the notion that every learner has their own 'learning style' that should guide optimal instruction methods, despite no reliable supporting evidence for this assertion (e.g. Kirschner, 2017; Pashler et al., 2008), and such ideas are prevalent among teachers and instructors, too (Howard-Jones, 2014; Morehead et al., 2016; see Chapter 6). The lack of participant awareness of their own cognitive processes is one reason that introspection is no longer prominent as a research methodology into many areas of psychology where it previously featured heavily, such as perception and creativity. Arguably, similar concerns hold in education.

In the context of desirable difficulties, problems with accurate insight are common even when learners have had a chance to try out learning techniques. This is well illustrated by a number of research studies, such as:

- Zechmeister and Shaughnessy (1980) demonstrated that learners who were allowed to learn via both spacing and massing considered the latter to be more effect, even although they had learned more via spacing.
- Kornell (2009) provided learners with small and large decks of flashcards to study from. He found that although the large decks increased spacing and were associated with better performance, learners viewed the smaller decks as superior.
- Hartwig and Dunlosky (2012) found a preference among learners for ineffective strategies such as highlighting and re-reading, rather than self-testing or spacing, and furthermore found that student report of using such strategies correlated with a poorer grade point average.
- Zulkiply et al. (2012) found that the large majority of research participants judged the interleaved presentation of examples as less effective, when the opposite was the case.
- Yan, Bjork and Bjork (2016) found that neither a theoretical explanation nor a practical demonstration was sufficient to induce learners to prefer an interleaved schedule to a blocked schedule.

Indeed, the fact that interventions that promote more effective learning are not always intuitive is a central part of the question investigated in this thesis. If the best course of action – in terms of more effective teaching and learning – could be worked out subjectively via introspection, trial and error, or professional reflection, there would be little need for systematic research.

The studies cited above (and others in a similar vein) have shown not only that learners tend to make errors in their choices, but have also shed some light into the reasons why they do so. For example, Kornell and Bjork (2008) have argued that the increased fluency of massed or blocked presentations makes learning feel easier, thereby causing learners to mistakenly think that they are learning more effectively. A similar pattern helps to explain the preference among students for re-reading over strategies that require more effective active retrieval (Roediger & Karpicke, 2011; see also Chapter 1 for a discussion of storage strength vs. retrieval strength, and desirable difficulties).

Overall, self-report is likely to be a poor method of investigating the specifics of learning processes, but could be more useful if asking for recall of specific learning events, such as asking teachers what tasks they set or ask learners about their habits and preferences (as was done in the Hartwig & Dunlosky, 2012, study).

As highlighted earlier, self-report can also be very useful when it comes to asking for opinions about learning from specific tasks, not least to contrast flawed assumptions with empirical evidence on learning. In particular, this kind of approach has allowed researchers to investigate flawed study habits among students (e.g. Kornell & Bjork, 2007), and could potentially be applied in a similar way to investigate the choices and preferences of teachers.

4.4.2 Types of self-report.

There are, of course, several types and formats of self-report methodology. Some of the principal ones are as follows:

4.4.2.1 Likert scales.

A Likert scale is a type of forced choice or closed question which gains numerical data from a questionnaire by labelling responses with numbers. A typical format is to provide a set of statements, with participants prompted to label choices such as strongly agree to strongly disagree, usually numbered 1–5 or 1–7, though sometimes with a narrower range of options.

Although this clearly maintains some of the problems of insight, memory and honesty discussed so far, it also has an advantage of efficiency, and is a straightforward means of summarising overall learner responses.

Likert scales are prone to ‘acquiescent (or acquiescence) response bias’, where some participants tend to agree more than disagree to any statement – a criticism of the classic F-Scale used by Adorno et al. (1950) to test for authoritarian views. A generally accepted fix for this problem is that items are typically reversed or flipped, such that agreement with an item could indicate either agreement or disagreement with the associated concept (Watson, 1992).

Likert scales lend themselves to summation across items, but there is disagreement as to whether the numbers in response to the scale constitute interval data. Some (e.g. Norman, 2010) recommend a more conservative approach, though there are numerous instances in the research literature where Likert scales relating to memory and learning have been analysed with statistics that assume at least an interval level of data (e.g. McCabe, 2011), and there could be an argument that the score on a scale represents a test score, and test scores are usually construed as interval. A possible resolution is to say that the analysis of Likert scales as interval data depends on the circumstances; it can at times be treated as the score on a test, and at other times the items are separate and not scalar, and cannot be meaningfully averaged.

A further matter that affects how a scale can be interpreted is the internal consistency of its items. Since the 1950s, Chronbach’s alpha has become a routine way of checking the validity of sub-scales, with the aim of ensuring that items are tapping into the same construct – and possibly removing them if they do not. However, Cronbach’s alpha has very strict assumptions which tend to be violated in most studies, and as a result can lead to flawed estimations of reliability. It is also problematic in education settings, when items might reflect different iterations of the same general problem, and might therefore be expected to yield different scores. As Vehkalahti et al. (2006) put it, “Most empirical problems are multidimensional, and it is difficult to develop items that measure only one dimension.” (p. 16). The common practice of deleting items that do not score highly for internal consistency would therefore be problematic. Finally, Chronbach’s alpha and similar measures are likely to face difficulties when participants are simply unsure of their own knowledge (see Section 4.2). It will therefore not be considered further here – but other checks on validity will be discussed in Chapter 6 (see section 6.4).

In the memory research, Likert scales tend to suffer from the insight problems discussed above, but are widely used to gauge attitudes towards teaching and learning. Indeed, outside of research, many institutions make regular use of the scales to gain course or lesson feedback. In terms of memory, specifically, Simons and Chabris (2012) have conducted research into public views about how memory functions which used a Likert scale. In this study, validity was achieved by contrasting participant scores with responses from the views of experts in the field. Other studies (e.g. Furnham, 2018; McCabe, 2011) have used literature-based approaches to establishing the ‘correct’ answer in such scales.

Likert scales benefit from administrative simplicity. A questionnaire-based study can be distributed via an internet link, allowing access to large samples, and has a greater ease of analysis compared to verbal self-reports. The format is also very familiar to most participants due to its popularity for marketing and on social media, making it relatively easy for participants to complete such questionnaires without supervision.

These features make a Likert scale with a series of short items – similar to that used by Simons and Chabris (2011) – ideal for gaining a broad survey of teachers views about memory in the early part of the current research.

4.4.2.2 Forced choice.

Simpler still than a Likert scale, a forced choice format uses closed questions and thus induces the learner to pick between one or more options, for example ‘agree’ or ‘disagree’ with a statement. This has been productively used as a follow-up phase to a test phase, in order to ask learners if they learned better or worse via an intervention such as spacing (see Section 4.3.2.2, above). It can also be used within an experimental protocol, as in tasks that require participants to state whether a stimulus is the same or different, or to pick among multiple-choice options.

A forced-choice metacognitive question will form part of the design of the current experimental tasks.

4.4.2.3 Vignettes or case studies.

Vignettes are short scenarios – for example a brief story featuring an interaction or event – which participants read and respond to in some way. One of the best-known examples from psychology is the moral reasoning scenarios used by Kohlberg (1963). These described situations where legal and moral obligations conflicted, for example a scenario where a man breaks into a pharmacist to steal an overpriced drug that can save his ill wife. The vignettes were used to assess people’s moral reasoning.

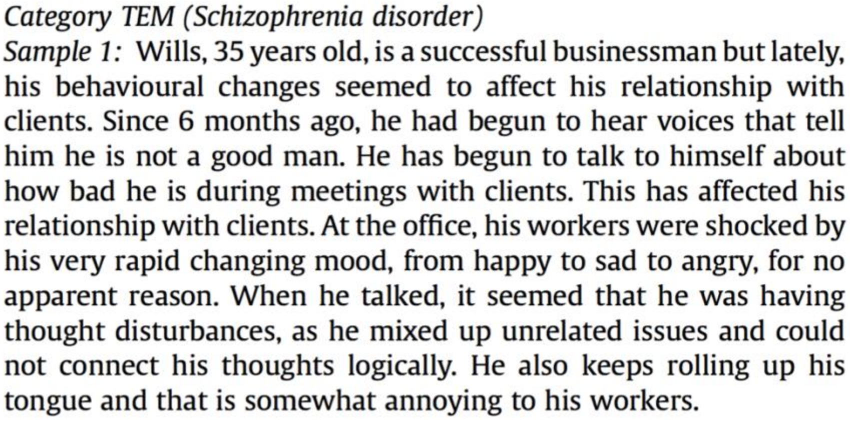
An educational example which also used moral dilemmas is the work of Underwood (2003), who showed vignettes to student teachers in order to gauge attitudes to group cooperation and exam malpractice. More complex versions can have two or more ‘stages’ to the story, with participants also responding to the follow-up sections (e.g. Finch, 1987).

There are many ways in which research participants may be asked to respond to a vignette. Kohlberg used semi-structured interviews, asking participants open questions about why they did or did not think a person’s actions were morally acceptable. Participant responses can take either a 1st person or 3rd person perspective, e.g. "I would..." or "She should..." (Senior et al., 2002), and variations can be made to wording, for example to assess differences in participant responses based on the described characters’ gender, status or ethnicity.

It is important for vignettes to have relevance to the participants and to have a high level of plausibility taking account of their level of knowledge (Neff, 1979). Compared to the brief and often abstract statements typically used with Likert scales they have the advantage of providing a context that makes a statement easier to relate to, helping to tackle the widely-acknowledged difference between expressed intentions and actual behaviour (e.g. LaPierre, 1934; Finch, 1987). However, they can still be combined with Likert or forced choice responses.

In the memory literature, a number of related studies by Nairne and colleagues (e.g. Nairne et al., 2007) involved presenting participants with scenarios followed by individual words on a screen presented at 5s intervals. Participants were then prompted to say how relevant the words were to the scenario. The rationale was that the scenario would affect the type of processing used to encode the words, with more items remembered in the context of an evolutionarily-threatening scenario (see Chapter 1).

Vignettes have been little used in the interleaving literature, but one relevant example is the work of Zulkiply et al. (2012). This study used “case studies” of psychological disorders, which participants were asked to categorise after an interleaved or blocked training phase. One example is as follows (Figure 4.4).



Category TEM (Schizophrenia disorder)

Sample 1: Wills, 35 years old, is a successful businessman but lately, his behavioural changes seemed to affect his relationship with clients. Since 6 months ago, he had begun to hear voices that tell him he is not a good man. He has begun to talk to himself about how bad he is during meetings with clients. This has affected his relationship with clients. At the office, his workers were shocked by his very rapid changing mood, from happy to sad to angry, for no apparent reason. When he talked, it seemed that he was having thought disturbances, as he mixed up unrelated issues and could not connect his thoughts logically. He also keeps rolling up his tongue and that is somewhat annoying to his workers.

Figure 4.4: Example case study from Zulkiply et al. (2013, p. 220).

There are many other possible uses of this method that could be pursued, particularly given the mismatch between learning and intuition that have been highlighted in Chapter 2. For example, learners or teachers could be given scenarios concerning learning and asked to discuss the most effective way to proceed, allowing their responses to be compared to experimental evidence on effective learning. An example of this procedure was used by (McCabe, 2011, 2018), in a set of studies where participants were shown descriptions of spaced or massed learning, or other relevant interventions, and asked to judge which was superior. However, they have not, at the time of writing, been used to judge teachers' views of optimal classroom practice.

This method would allow for a relatively rich view of how teachers respond to desirable difficulties in context. The findings of such a methodology are more likely to reflect classroom practice than responses to briefer and more abstract statements. Having established a broad basic view of memory among the teaching profession, then, I would hope to use well-chosen vignettes as part of a more in-depth study.

4.6 Populations, Samples and Data

Psychology experiments are predominantly conducted on university undergraduates, and there are corresponding biases in the findings (Sears, 1986). However, as discussed above (see Section 4.2), some psychological traits and educational behaviours are justifiably viewed as more universal than others. While researchers can make a strong case that concepts like LTM and WM hold across all samples, there may nevertheless be group differences due to factors such as age and education level. This section will consider the challenges involved in obtaining good quality samples of participants, and in generalising any findings from these samples to broader populations.

4.6.1 Sampling.

There are many relevant studies that have used small, easily-accessed samples of undergraduates, and an opportunity sample of student participants can still be considered the default position. Kornell (2009, Experiment 1) used just 20 participants in a study of interleaved practice, but other studies are larger scale; Carpenter et al. (2015) sampled 311 students, although attrition led to an eventual sample of 275, highlighting the need to oversample when spacing and follow-up conditions are involved. Some studies (e.g. Hamilton, 1990) have excluded participants who show too much prior familiarity with test material – a potential issue with any experiment which make use of authentic learning materials.

The diversity of student samples can be a problem when generalising results, and some researchers have aimed for more diverse samples using online recruitment e.g. via MTurk, often resulting in a wide age range. For example, Kornell (2015) recruited 800 such participants to study the effects of feedback on test answers. The same technique was used by Simons and

Chabris (2012) to follow up on their initial survey into public attitudes towards memory, and was combined with *stratified sampling* in order to gain a representative sample in terms of factors such as educational background, occupation and age.

In the present research these matters must be taken into account, but it is also essential that the samples used are authentic and relevant to the educational processes under investigation. For example, gaining access to carry out research on students who are doing exam-based courses in secondary schools and to current teacher education students will be a higher priority than gaining a highly diverse sample in the current research.

4.6.2 Individual differences.

Developmental changes affect an individual's immediate processing abilities (Gathercole et al., 2006) as well as their metacognitive functions (Blakemore & Choudhury, 2006) and can therefore impact on their learning concepts via examples and practice. Working memory (WM) is multi-component system involved in immediate processing and which includes verbal and visual elements (see Chapter 2). It is also subject to developmental change, both physiological and due to changes in prior knowledge which affect cognitive load (see Chapters 1 & 2).

The benefits of spacing and interleaving have been evidenced with young children (see Chapter 2), but it is also important to remember that such students may encounter new concepts in quite different ways from older students – less frequently from lecture formats, for example, and more often via inductive discovery learning. Developmental changes are also more rapid at younger ages. For these reasons, this thesis will focus on learners in the teenage years and adulthood, and evidence on spacing/interleaving on child participants will therefore be less prominent in the discussion that follows. Such studies will occasionally be discussed where they illustrate relevant methodological points, and it is also important to consider that among teenage and adult populations, too, working memory abilities vary.

Importantly, any multi-tasking of visual and verbal (or both) depend on the attentional demands of the task and this in turn depends partly on task novelty, as new tasks are less likely to be automated (and able to be completed with reduced attention). It’s also important to consider that ostensibly verbal tasks (such as reading a passage) may involve visual-spatial processing in the form of visual imagery in order to comprehend or respond to the text. The connection between images and language in multi-media tasks is well summarised and modelled by Mayer (2003, p. 129).

In addition to biological and psychological development, learning is a cumulative process. Learners with prior experience of a concept are likely to be at a marked advantage, having already established the schema knowledge. As noted above (Section 4.3), reducing task authenticity by using artificial stimuli can help to minimise the influence error of prior learning as an extraneous variable, but will also typically reduce generalisability. Where authentic classroom material is to be used, one way to reduce the influence of prior learning would be to select a novel learning task in a school subject that is nevertheless new to learners. This is an authentic situation, as learners regularly tackle new areas of scientific/social-scientific learning when they begin a new school year or new set of modules. For example, many school pupils begin to study disciplines such Politics, Psychology or Sociology in the latter years of school, when they have already completed their mandatory schooling and after most aspects of biological development are complete.

Another approach is to conduct prior testing or questioning. This was done by Jaeger et al. (2016) who asked psychology students how much they knew about El Niño, and then chose that as a topic for experimentation because ratings indicated little prior knowledge (mean self-report rating of 2.51 out of 10, where 10 indicated the highest level of knowledge). Again however, testing can constitute a learning event (see Section 4.2) and even prior testing in the absence of knowledge runs the risk of affecting later scores (this is known as the *pre-testing effect*; see Richland et al., 2009). Therefore, the increased certainty that comes with prior testing has drawbacks to be weighed up against the uncertainties that come with assuming a lack of prior knowledge.

A hybrid between the two would be to test peers from the same population on a self-report scale, and generalise their self-reported prior knowledge to the tested sample. It should also be borne in mind, though, that a person's self-assessment of their own knowledge is subject to the biases discussed above (see Section 4.2) and that generalising even within a population in the same educational establishment is difficult due to the highly idiosyncratic nature of general background knowledge.

Some of the key individual differences studied in the prior literature are summarised in Table 4.3 (overleaf), together with the action required. Overall, a combination of using older learners and those who are beginners in a relevant school subject is hereby recommended for the coming empirical studies in this thesis.

4.6.3 Allocation to conditions.

Although non-random samples are commonly used throughout the cognitive psychology literature, random allocation of participants to conditions is standard practice in experiments because it is necessary to avoid confounding variables that may exist if pre-existing groups are used (a key flaw with natural experiments).

However, it more typical to use non-random allocation in field studies, as discussed by Küpper-Tetzel et al. (2014). Different pre-existing classes may be selected as the participants for different conditions, given the difficulties of varying conditions within a single classroom.

It may also not be possible to fully control conditions or implement a thorough control condition, as discussed by McDaniel et al. (2013; see Section 4.3). These limitations will often need to be weighed against the benefits of field experimentation, which is often the logical next step after effects have been demonstrated and replicated under controlled conditions.

Table 4.3: Summary of methodological actions required to account for key individual differences among participants.

Issue	Action Required
AGE	Participants will be teenagers or older; replication of findings with different age groups.
VERBAL WM DIFFERENCES	Prior testing where necessary; ensure that any distraction tasks occupy verbal WM
VISUAL WM DIFFERENCES	Prior testing where necessary; ensure that any distraction tasks occupy visual WM
BACKGROUND KNOWLEDGE	Use authentic materials that are likely to be novel OR use prior testing/questioning about subject knowledge.

Again, a balance must be struck between authenticity and control. While random allocation is desirable, removing pupils from their usual class to allocate them to experimental conditions also has a cost, and may not be deemed acceptable to host schools.

One methodological option that may preserve many of the advantages of both scenarios is to allocate separate groups of students within a class to separate tasks. This will generally involve lesson structures which are not teacher led, as teaching the material for one condition and then the other to the whole group would result in unacceptable order effects. In a freer setting such as via project work or computer-based testing, students can be allocated to different tasks while working in a familiar and comfortable setting, particularly with within-subjects designs whereby pupils will get the opportunity to study all conditions at some point.

This option can potentially gain the best of both worlds – authenticity and randomisation, and will therefore be used in the present research

4.6.4 Incomplete data.

Regardless of the research methodology used, incomplete data is the rule rather than the exception in research (Dong & Peng, 2013). Field research is especially prone to participants who drop out or are otherwise unable to complete the tasks provided, as exemplified by the Carpenter et al. (2015) study referred to above. In a classroom, for example, students frequently leave the room for a range of reasons. Online questionnaires may be started but not completed.

If data sets are partially complete but some scores are missing, there are various options for handling this. The simplest is to leave cases that have any missing scores out of the analysis entirely. This is known as listwise (or ‘casewise’) deletion. However, it is typically seen as a flawed approach, because the missing cases may be non-random, and deletion of these records could bias the results (Allison, 2001). It also leads to data loss overall.

A straightforward fix for this problem is to replace each missing score with a mean score. This preserves the participants’ responses to other variables without unduly biasing the overall results in the variable where the missing score was replaced. However, adding mean scores does lead to some noise in the data, and this technique is generally not recommended if the number of missing scores exceeds 5% of the sample.

Another possibility is the multiple imputation (MI) technique. While little used in education, it should be considered the default technique due to its better performance and consequently higher levels of internal validity and generalisability of results (Manly & Wells, 2015). MI involves replacing missing values with values that are imputed – i.e. inferred – based on the existing data. A calculation similar to multiple regression is used to generate several imputed values, and software typically pools these and replaces the missing datum with an averaged value.

In some cases, however, there may be an argument against using imputed scores, particularly if the lack of response is meaningful in some way.

There is also a judgement that must be made about whether complete data sets should be included or not. In Clapper's (2015) study of category learning he applies a technique of removing the lowest-scoring 25% of participants from each condition to correct for inattentive responding. This could be valuable for difficult tasks in educational settings, helping to remove the random noise which would result from demotivated students picking random answers (see also Meade & Craig, 2012) and thus increasing the power of an experiment.

In the present studies, however, an assumption will be made that data should be used rather than discarded where possible. Data sets will be analysed to assess whether missing data shows a pattern of bias. Data will only be excluded listwise where technical approaches to replacing missing data are unavailable, or where attempting to replace missing data appears to be at risk of adding distortion. When data are replaced, for example using the MI method, original data will also be reported.

4.6 Ethical Considerations

As this research project draws on both Psychology and education, I will follow both British Psychological Society (BPS) and British Educational Research Association (BERA) ethical guidelines throughout.

The previous section (Section 5) raises one of the key ethical issues for any research that I conduct on students – it should only be conducted in a way that does not harm or interfere with their studies. This means, for example, the learners can't be allocated to conditions where they are given a deliberately inferior learning experience or misleading information, or where parts of the lesson contents are missed out. It would also be unacceptable in many contexts for learners to spend a substantial about of time on irrelevant distraction/filler tasks.

Following from the above, there must at least be ethical doubts about the use of superior learning practices with one group of learners, where another group/class is used as a control. If it is practical to apply a manipulation that is already known to be superior, it could be asked why this is not offered to all. Often, however, experiments will concern manipulations where there is still uncertainty about the outcome, for example when applying findings from the laboratory to the classroom. In such a scenario, a 'business as usual' control group needn't be seen as having been unduly disadvantaged. Another option would be to make participation in a task voluntary, with those who do not participate acting as a de facto control group.

Further ethical considerations relate to the time and demand placed on student learners. Their time is limited, and the purpose of their being in school or university is to learn. Fortunately, applied experimental tasks can navigate around this consideration by only using tasks which are pertinent to the learners' interests, but this requires careful selection or materials due to the additional risk of prior learning acting as an extraneous variable, as discussed previously.

As with any study, there may be a need to avoid sharing full information about aims and hypotheses at the outset. Here, BPS guidelines (BPS, 2014) indicate that full information should be provided to participants at the earliest possible stage. An acceptable and commonly used approach is to gain informed participant consent on the basis of general information (e.g. that participants are taking part in an experiment/survey on learning and memory), and furnish them with fuller information about the aims during a debriefing stage. An additional step that may be appropriate in some circumstances – for example when questioning teacher participants about learning – will be to provide full information about the processes under investigation via a web link, leaflet or video, as part of the debriefing. This will not only provide fuller information, but has the potential to aid their professional practice. Participant data must always be kept secure, and strict confidentiality maintained.

Any ethical decisions relating to the subsequent sections of this thesis were taken in consultation with the PhD supervision team, with University ethics board oversight, and with the proper permissions from the authorities responsible for any external learners involved such as school boards or local authorities. Where appropriate, teachers chose to participate as independent adults without referring to their institution/employers.

4.7 Interim Summary

4.7.1 Overview.

The previous sections have outlined several considerations that will be taken into account. Overall, experimental methodology has been the primary vehicle for research into interleaving and other desirable difficulties in laboratory settings. Taken together with the absence of interleaving research at high school level (see Chapter 3), it makes sense to used field experimentation in school settings to find out more about the technique.

It also makes sense to do so with authentic materials and to following procedures that maintain ethical standards, while ensuring that extraneous variables are kept to a minimum. Some trade-off between authenticity and control will be balanced in favour of authenticity in the present research, in part because the existing literature tends to lack evidence that clearly pertains to real classroom processes, and its external validity is therefore questionable.

When it comes to teachers, it will be useful to survey their beliefs via self-report methods. Although some flaws are to be expected, this will be informative. After all, if teachers' beliefs about learning (and specifically, desirable difficulties) are sharply out of line with the research consensus, there will be implications for practice and for professional learning.

To do so, a broad-ranging survey with short items and a Likert scale will be a good starting point, after which time – having identified the main areas of confusion – a follow up with extended vignettes can be applied.

4.7.2 Alternative methodology options.

Of course, the foregoing sections have not outlined all of the possible ways that interleaving and desirable difficulties could be investigated – just some of the most popular and thus well-established. Some other possibilities are briefly discussed below, and will be returned to when analysing the current research studies.

4.7.2.1 Written tasks.

A number of experiments have asked learners to freely recall everything they can from a lesson or text, and have then analysed this output. Roediger and Karpicke (2006) asked participants to read a text once or several times, and then attempt to retrieve it by writing down everything they could remember. Similarly, Jaeger et al. (2016) used a writing task for their research into learning from analogies, with participants asked to write an essay on paper in an academic format. Such written tests are typically scored according to content, with the presence or absence of key material assessed. Essentially this is quite similar to some forms of content analysis used with unstructured or semi-structured interviews, with content being analysed on the basis of key themes and points. Due to the risk of subjectivity in any such written test, two or more coders can be used, and their ratings can be scored for *inter-rater reliability* (e.g. via Krippendorff's alpha; Krippendorff, 2013).

4.7.2.1 Numerical responses.

As noted above (see Section 4.3), recognition memory can be assessed using various types of forced choice test. Responses to these are easier to score than free response tests, due to there being less room for errors, particularly options are presented via closed questions rather than written. A similar option is to give a direct, short answer test, such as the 18-item *concept verification test* used by Braasch and Goldman (2010), which asked participants to state whether statements are true or false on the basis of studied information. These data gathering options provide results which are quantitative at the outset and therefore cannot be biased by the subjectivity of a rater or judge. A similar advantage obtains from the use of Likert Scales. An interesting variation on the recognition list is the *Deese–Roediger–McDermott paradigm,* where participants' recall of associated (but not presented) 'lure' words is assessed. For example, when shown a list of words which contains many medical terms and concepts such as 'hospital' and 'nurse', a participant may falsely recall seeing 'doctor' if that latter is not presented. This technique can be useful to identify false memories caused by meaningful association, and participants often lack insight into the errors (i.e. their episodic memory of the list is not accurate enough to distinguish lure errors from list items, even when questioned; Roediger & McDermott, 1995). The technique has potential to be used for investigation of spacing or interleaving procedures.

4.7.2.1 Qualitative research.

Other prominent research methodologies such as naturalistic observations, open interviews and case studies have obvious relevance to learning and memory research, but have been little used. In part, this may relate to the insight problems with self-report (see Section 4.2), but it would be useful to conduct interviews with teachers as well as naturalistic classroom observations to evaluate current practice and assess how prominent real-world concrete examples are among practicing teachers, and how they are used (notwithstanding the memory errors and biases that can plague classroom observations; Firth, 2020, see Appendix 7).

The case study method could be useful for getting more detailed and qualitative data on individual learners and/or teachers. For example, one or more pupils/students could be investigated to determine whether a typical research process is efficient in its use of examples for concept learning, and to find out whether (as seems plausible) features such as spacing and interleaving occur naturally as part of an independent research process. If so, it would provide a useful piece of evidence for the ongoing debate over independent projects and groupwork in education (for example see Leat, 2017, for a supportive view, although a report by EEF, 2016, was largely critical). Alternatively (or in addition), such a scenario could be presented to teachers as a vignette.

4.7.3 Priorities established for empirical research

On the basis of the methodology options and issues identified above, a plan for empirical research to be undertaken was established. The studies undertaken are described in the following sections of this thesis, and can briefly be summarised as follows:

- A pilot study of student learning using interleaving, to establish an appropriate classroom-based methodology using computer-based tasks, and to address the combined role of spacing and interleaving in a task that could take place within a single school teaching period.
- A larger-scale study of learning using interleaving, which can be conducted on older school learners or university undergraduates including student teachers, and will use authentic learning materials which are of relevance to the participants' interests.
- A pilot study of teachers' beliefs about memory using a Likert Scale. This will offer participants a broad range of memory-relevant teaching situations. A similar methodology to that of Simons and Chabris (2011) can be used.
- An in-depth study of teachers' beliefs about memory using vignettes, focusing in on items from the pilot study which are most relevant to the specific aspects of memory under investigation, and presented in recognisable classroom scenarios. If possible, this will be done with a range of experience levels, to give an indication of how metacognition develops in line with experience.

These studies are reported in Chapters 5 and 6, next.

5

Classroom-Based Investigations of Interleaving

5.1 General Introduction

Given the strong standardised effects sizes and applicability to a range of different types of educational materials reported in Chapter 3, interleaving would appear to have huge potential for use as a classroom technique. As such, it joins a group of techniques which can form the foundation of evidence-informed classroom practice.

Although some experiments into the phenomenon have used abstract stimuli, recent demonstrations of interleaving with classroom-relevant material include the categorisation of chemicals into types (Eglington & Kang, 2017) and the learning of psychological concepts via examples (e.g. Rawson et al., 2015).

However, much of the evidence relating to the technique thus far comes from laboratory studies, and as such was produced under very specific conditions. The field studies that have been done (e.g. Hausman & Kornell, 2014; Rohrer et al., 2015) largely relate to revision and consolidation of material (such as by using interleaved flashcards) rather than new learning. How exactly can interleaving be applied to everyday classroom practice? What barriers are likely to arise? And will the now well-established interleaving vs. blocking effect replicate with groups of students studying for exam-based courses, such as those in Scottish secondary schools? These are the questions tackled in this chapter.

5.1.1 How should interleaving be applied in the classroom?

Interleaving appears to be beneficial because it prompts learners to contrast different problems or examples, and/or because they pay more attention (they mind-wander less) compared to seeing repetitive types of problems or examples. If a student needed to learn the difference between hawks and falcons, for example, it would be preferable for them to see interleaved examples of hawks and falcons (HFHFHFHF) rather than blocks with several examples of hawks (HHHH) followed by several examples of falcons (FFFF; Eglinton & Kang, 2017). In practice, a blocking of topics across separate lessons is more likely to reflect what happens in school classrooms.

Interleaving, at least as it is defined in the research literature, relates to short items, such as images and one-sentence examples. Longer examples are rare, and even the interleaved case studies of types of psychological disorders used by Zulkiply and colleagues (e.g. Zulkiply et al. 2012) are each only around one paragraph in length. The larger the grain size of each item the fewer the contrasts, and so for interleaving it is a case of the smaller the better. The term (at least as supported by the evidence outlined in Chapter 3) therefore does not refer to interleaving entire lessons or topics – something that may already be commonplace due to the intermingling of subjects/courses in a typical school day.

Furthermore, the research evidence has not provided much insight into blocking that occurs across different study sessions (e.g. hawks on Monday, falcons on Tuesday), as might commonly happen in schools. It is possible that such a schedule would make it even more difficult for learners to contrast different types of item than is the case in most lab studies (which tend to occur in a single session), and therefore make interleaving even more advantageous than blocking. There is, as yet, a shortage of classroom empirical evidence on this, but laboratory findings have shown that inserting delays can make things worse – it makes that contrast between one item and the next more difficult, meaning that interleaved schedules perform no better than blocking (Birnbaum et al., 2013; Kang & Pashler, 2012). Spacing, in this case, is unhelpful, at least in terms of its interaction with interleaving; as Monica Birnbaum and colleagues put it, “two desirable difficulties are not always more desirable than one” (p. 401).

5.1.2 Combining interleaving with spacing.

This raises an important and complex problem regarding how interleaving can be implemented in real learning situations and applied to school plans and timetables in such a way that it works in harmony with the spacing effect, such that strategies based on the two effects do not interfere with one another.

As noted in Chapter 2, interleaving can be clearly distinguished from the spacing effect in terms of purpose, practice and underlying mechanisms, even though the two have been conflated by researchers in the past. While both involve altering the order in which items (for example textual examples) appear, the latter refers to the benefit of incorporating time delays between learning and practice, leading to improved performance over educationally relevant time periods (Cepeda et al., 2008) compared to ‘massed’ items, where practice sessions occur close together. Review and practice after a delay is harder, but it is also more effective than reviewing too soon (before any forgetting occurs), meaning that spacing is a desirable difficulty. It is worth noting that spacing is typically studied over long timescales, such as from one lesson to the next (e.g. Foot-Seymour et al., 2019; Kapler et al., 2015; Mazza et al., 2016).

Interleaving, however, has the effect of *reducing* delays between contrasting items, eschewing categorisation by type in favour of alternated or shuffled orders such that learners do not see ‘like with like’, but rather ‘like with unlike’. This may be because such scheduling puts different items side by side, improving perception of the differences between them – the discriminative-contrast hypothesis (Kang & Pashler, 2012) – and helps students to develop a meaningful understanding of the differences between two categories, especially when such differences are subtle. Seeing contrasting items in different study sessions would instead require learners to contrast a memory of the past item with the current one – cognitively much more demanding to do.

An interleaved presentation of new material is inevitably spaced to some extent (see Chapter 2). Initially, that is why researchers thought that interleaving was helpful (e.g. Kornell & Bjork, 2008), but this turned out not to be the case; a number of studies have used filler activities to increase the spacing of blocked items, meaning that spacing remains the same in both conditions. If spacing between examples is incorporated into a blocked schedule by adding filler items with the overall study time spent viewing target items kept constant (see Figure 5.1), interleaved examples still lead to superior performance.

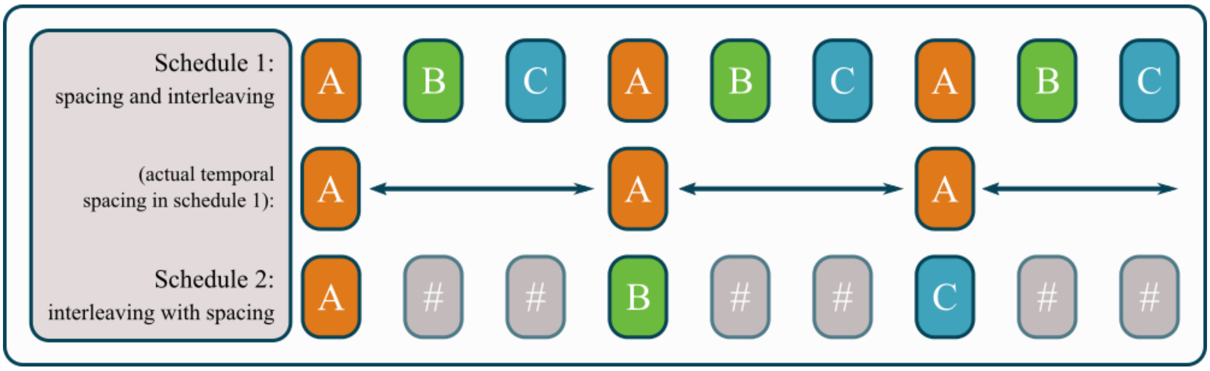


Figure 5.1: Testing interleaving while keeping spacing constant. A, B, and C represent examples of three different categories. The # symbol represents unrelated filler items. Arrows represent a time delay.

However, this benefit disappeared if the two types of presentation (blocked vs. interleaved) both had their temporal spacing increased, due to the delays interfering with discriminative contrast (Birnbaum, 2013; Birnbaum et al., 2013). That is to say, without items being presented consecutively, interleaving ceases to be helpful. This may be because the two strategies differ in their primary benefits. The spacing effect boosts memory – practice or restudy of material is more effective if spaced out over time, with forgetting reduced – while interleaving boosts inductive category learning and later transfer. This inductive learning of categories is boosted by discriminative contrast, helping learners to notice and retain the meaningful boundaries between confusable categories.

5.1.3 Barriers to applying interleaving in the classroom.

What particular benefits might interleaving have for the teacher? And what processes and considerations should guide the application of interleaving to the classroom, such that it is used when it is most beneficial and avoided when it could be detrimental?

Both interleaving and spacing are *desirable difficulties* in that they make learning more difficult, but in a way that is beneficial (Bjork & Bjork, 2011). They are both widely recommended among those who aim to apply cognitive psychology to education, appearing, for example, among the "*Six Strategies for Effective Learning*" recommended by the popular Learning Scientists blog (www.learningscientists.org).

However, as difficulties, they pose certain challenges to the educator. Learners may find the strategies more arduous to use, increasing a short-term sense of confusion or a lack of progress. Indeed, learners may actually make more errors and achieve slower progress when these strategies are used (Soderstrom & Bjork, 2015).

Davis et al. (2017) have found that frequent switching between studying and test questions can be detrimental, and there may be a motivational impact if students feel that they are being over-tested, especially when the material is new and confusing. This is one reason that Kang

(2016) reasons that a hybrid approach can be beneficial, with new learning done via blocked practice and interleaving used in a practice or consolidation phase.

In school, there must also be a face validity to the tasks used (Loyd & Kern, 2005). When faced with a new topic or sub-topic, students expect to experience some confusion and difficulty, but do not expect to be given filler or distractor items (as can be done within lab studies). Maintaining the integrity of school-relevant tasks is therefore important for the application of the technique (see Chapter 4). In practice, this may mean applying spacing and interleaving to different elements of the same classroom task; the former where brief delays in the study process can help with later recall of factual information, and the later where categorisation of easily-confused information may be helpful.

Another factor that has to be borne in mind is that school pupils work at different paces, in part because of different levels of relevant prior knowledge. Some may grasp a new concept after looking at a small number of examples, while others may need to see further examples or to revisit earlier ones. These differences imply that there might be a place for individualised (rather than teacher-centred) forms of learning.

Prior knowledge affects how well students will understand what they are learning, and how easily they will be able to retain and transfer new learning. As discussed in Chapter 1, long-term memories develop in part by making connections between new information and existing knowledge. The structure of observed learning outcome (SOLO) taxonomy presented by Biggs and Collis (1982) states that to understand learners' developing competence, it is better to look at the developing depth of the connections among their knowledge, rather than seeing skills as fundamentally separate (as is the case with Bloom's taxonomy). At an early stage, as the model explains, learners are likely to miss the point of new information, and fail to understand. They then move on to associating new learning with one particular thing, and may be able to complete simple procedures or answer questions, but it is only when they learn several independent aspects of the knowledge (i.e. more well-integrated schema knowledge) and are able to respond flexibly and in ways which demonstrate critical thinking skills such as analysis and evaluation.

Following the (SOLO) taxonomy, school learners on a new course must to establish new factual knowledge, but will initially show a lack of competence. Spacing of initial practice may be useful in that it helps with the retention of new facts prior to schemas being fully developed, but spacing alone may not lead to transfer in the absence of complex understanding. Interleaving, however, is well-placed as a learning strategy to tackle the early lack of a complex knowledge structure. By boosting comparisons between confusable concepts, it can help learners to establish an understanding of the new material (with more subtly sub-divided schemas), and thus to be able to correctly categorise new iterations of the same concept – a novel member of the same category, for example.

Interleaving therefore has the potential to help learners quickly establish schema knowledge in a new subject, while spacing has the potential to consolidate their learning with a store of factual information due to its mnemonic benefits. Together, they may help teachers to their key goal as described by Soderstrom and Bjork (2015): “the primary goal of instruction should be to facilitate long-term learning – that is, to create relatively permanent changes in comprehension, understanding, and skills of the types that will support long-term retention and transfer” (p. 176).

Application of interleaving must also target the right material – it cannot be applied to every set of facts or examples. As discussed in Chapter 3, the benefit of interleaving seems to depend on the mixing of related items such as examples from similar categories. This idea is supported by research which has shown that the interleaving of unrelated categories has no benefit, such as Hausman and Kornell’s (2014) study which interleaved anatomy terms with Indonesian vocabulary. Additionally, interleaving seems to be more effective when differences between items are subtle, and may even cease to be effective at all with items that are very different (Carvalho & Goldstone, 2014).

Smith and Scarf (2017) note that if spacing of learning across days is to be helpful, a minimum initial level of experience is required in the student. This fits with the idea that forgetting will be rapid if there is a lack of foundational knowledge in the form of relevant schemas. If forgetting is too severe, then the re-study phase will resemble a second attempt at initial learning, rather than practice of previously-studied material. To put it another way, the initial study session needs to be effective with new material well practiced prior to a delay. Dunlosky and Rawson (2011) emphasise this point: “our prescriptive conclusion for students is to practice recalling concepts to an initial criterion of 3 correct recalls and then to relearn them 3 times at widely spaced intervals” (p. 283).

In addition, as pointed out in the previous section, combining difficulties does not necessarily result in a cumulative benefit. Spacing and interleaving are both desirable difficulties, but they appear to work best when applied separately. On the other hand, testing (retrieval practice, another desirable difficulty; see Chapter 2) can be easily and successfully combined with other techniques. Bahrick (1979) describe the technique of spacing out practice tests as ‘successive relearning’, and Rawson et al. (2013) provided evidence that it can boost exam performance in a real-world psychology course over educationally-relevant timescales. However, as Rawson and colleagues also note, surprisingly little field research had been done into this area. An important direction as research moves from the laboratory to the classroom, then, is to consider how individual desirable difficulties can be combined, and how they interact in real classroom situations.

The studies discussed in this chapter looked at how real students could take on board new psychology concepts during their studies. It comprises two studies – a pilot study done in a school classroom, followed by a larger-scale study carried out across several secondary schools. Implications for other educational settings are also discussed.

5.2 Study 1 – A Pilot Study of Interleaving and Spacing

5.2.1 Introduction to Study 1.

The best way to apply spacing and interleaving to school-based learning of new concepts is still little understood. It is unclear how these desirable difficulties interact in realistic situations, and what effect interleaving and spacing have when combined within a single classroom task which presents new concepts to be learned. Away from the artificial tasks used in the lab, can the two techniques be productively combined to establish both factual recall and an understanding of new curriculum material? The aim of Study 1 was to address this question.

To do so, I used the medium of psychology learning. Psychology is a popular but niche school subject course in Scotland. *Niche* because only some secondary schools offer it, and pupils can typically only study it during their final two years of school (during which most are around the ages of 16–17). *Popular* because when it is offered, the uptake tends to be high (over half of the eligible pupils in some schools; see Appendix 8). As a subject that is relatively new to learners, relatively concept-based, and where the learners are of a similar age to the undergraduates studied in much of the previous research, it offers a good starting point for applications of interleaving to the secondary classroom.

The main psychology courses taught in Scottish schools are *Higher Psychology* and *National 5 Psychology*, both provided by the Scottish Qualifications Authority. Both constitute optional courses within 'senior phase' (i.e. final years of secondary) of a broader national curriculum which outlines aims and content for young people aged 3–18, the *Curriculum for*

*Excellence (*or *CfE*). Each course lasts for one academic year, and collectively around 5000 candidates per year study the two courses (see Appendix 8).

The psychology course content has changed periodically, but the most popular topics covered over recent years across the two courses include Sleep, Conformity, Prejudice, Psychopathology, Aggression, Memory and Non-verbal Communication. The specific content is outlined by the exam board, but there is some freedom for schools/teachers to interpret and flesh out this outline (for example, the Higher Psychology topic of Prejudice states that 'social identity theory' must be covered, but does not say what specific research studies or concepts this should include; SQA, 2018).

Prior knowledge among pupils at the outset is very limited because psychology is not widely taught during the 'Broad General Education' phase of the Scottish curriculum (which stretches from pre-school to the middle of secondary school), and psychological content does not explicitly feature among the 'Experiences and Outcomes' around which this part of the curriculum is based[13]. A few topics (for example mental health, stress, sleep & dreams) either have an element of common knowledge (most students will know that people dream and get stressed and have some idea of what these things involve) or are touched upon more technically in other school subjects (most will cover the 'fight-or-flight' response in Biology, for example). However, at least in my experience as a classroom teacher, other concepts will be entirely new to learners, as will the broader approaches and methodologies used in the subject (pupils often anticipate a largely discussion-based course, rather than one that explores and analyses scientific evidence). They may also have misconceptions, for example thinking that psychology mainly covers forensic and clinical issues.

[13] It is implicitly present, however; for example, particularly in the Health and Wellbeing curriculum area (one of eight curriculum areas), in which pupils are asked to consider their own behaviour and feelings and those of others.

Given the fact that school students tend not have much in the way of useful prior knowledge in most psychology topics at the outset of their studies and that what they do have is quite variable and uncertain, it would be very useful for teachers if there was a way of rapidly establishing and understanding of basic concepts in a whole class at the outset of teaching.

In the pilot study, I attempted to address this problem. Working with a single new cohort who had begun to study psychology in June[14], I developed and piloted a computer-based task that provided school-relevant learning of new topics while allowing for research data to be gathered. The focus of the task was the new Psychology topic of phobias.

The research questions and predictions were as follows:

(a) How will interleaving and spacing affect the learning of new concepts in psychology?

Both techniques are (desirable) difficulties, and therefore it might be assumed that over the short-term at least, students will find new concepts harder to grasp if examples are spaced and/or interleaved. On the other hand, spacing has the potential to arrest forgetting and thus make learning more secure, meaning that newly-constructed memories will be easier to access at the point of a criterial test. Interleaving may boost understanding due to discriminative contrast, potentially counteracting the harm caused by the increased difficulty that it adds to a task.

(b) What interactions will be apparent between the two concepts?

On the basis of previous research such as that by Birnbaum et al. (2013), as well as the inevitability of spacing occurring in an interleaved presentation (irrelevant filler tasks cannot reasonably be used in a real academic context where class time is strictly limited) there is good

[14] In Scottish secondary schools, a short period between the May/June national exams and the summer holidays is usually devoted to making a start on new courses.

reason to suppose that spacing and interleaving will interact. The nature of the interaction is difficult to judge in advance, however, due to the lack of prior research which tests the concepts in realistic classroom contexts.

In addition to these two questions, a broader aim of this study was to pilot the methodology of the research task itself – an online presentation of psychology examples – in terms of its suitability for larger-scale research and for teaching purposes. It was essential that the learning task was authentic and maintained face validity, even if this meant compromising on experimental control in some aspects of the methodology.

5.2.2 Method.

5.2.2.1 Participants.

An opportunity sample of 31 school pupils was used. They were drawn from a large, co-educational independent school in Glasgow, Scotland, where the researcher worked at the time. All participants were aged between 16–17 years of age. Data were gathered during an end-of-year taster session during which pupils sampled several new subjects. They had not previously studied the topic being learned.

5.2.2.2 Design & materials.

In order to make the tasks as authentic as possible, all materials were based around the National 5 Psychology specification (see above), which features the concept of phobias. The experiment aimed to reproduce the range of activities in a typical school class, and so learners were taught facts about key concepts, and they also spent time reading and evaluating relevant supporting research evidence. The types of phobia covered included social anxiety, specific phobias and agoraphobia. Tasks were delivered via an online protocol on the PsyToolkit website (www.psytoolkit.org; see Stoet 2010, 2017), a free set of software for programming and running experiments and surveys via a web browser.

The independent variables in this study were interleaving and spacing. The dependent variable was the score on a test, assumed to be multidimensional, and analysed as a percentage.

Interleaving was operationalised as presenting information on three types of phobia, such that information on three types (specific phobia, agoraphobia and social anxiety) were presented either in a mixed format (one fact on each type of phobia per screen) or blocked (three facts about the same type of phobia per screen). For example, in the blocked condition, a participant would view three items relating to agoraphobia, while in the interleaved condition they would view a key feature of each of the three types. Definitions were based upon criteria in DSM-5 (American Psychiatric Association, 2013). Types of phobia were defined with key diagnostic information given; concepts and information were either presented together (blocked) or mixed with information about different types of phobia on the same screen (interleaved).

Spacing was applied to information about a research study pertaining to phobias which was shown on two screens, the first with a description of the study and the second with evaluation points, with the latter either presented immediately (massed condition) or after the second phase (spaced condition). The precise time delay for each participant therefore depended on their reading speed during the second phase (reading 353 words on screen); pilot testing had indicated a delay of two to three minutes. Spacing was therefore operationalised as immediate study of a related set of information versus delayed study of a related set of information. This differs from experiments which have spaced repetitions of *identical* information. However, as noted above, it was important for the task to be delivered in a way that resembled a nomal classroom information, and the view taken here was that second part of the task will have led to a spaced re-activation of the previously-studied material.

These phases are summarised below (Figure 5.2).

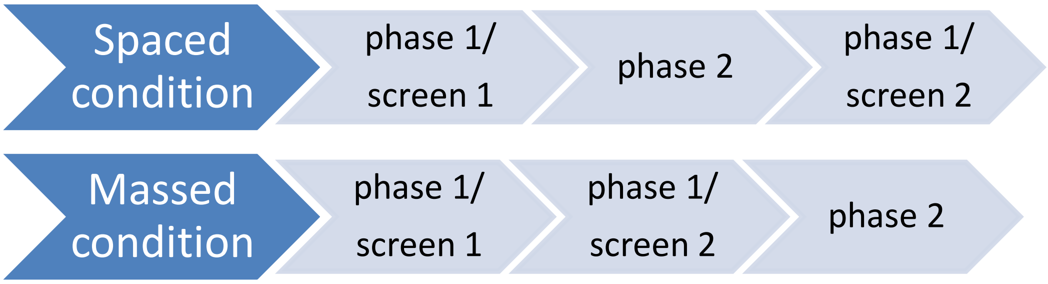


Figure 5.2: Schedule of tasks across the conditions of study 1.

The online tasks also featured a criterial test, comprising multiple choice questions about the research studies as well as a categorisation task featuring novel examples of each type of phobia. The questions on the research study linked to the spacing IV, while the categorisation task related to the interleaving IV (types of phobia). The choice of items from this test reflected standard classroom practice, with ecological validity prioritised over concordance with prior research procedures. It aimed to provide an educationally authentic measure of how well participants retained the previously-displayed information about the research study and the factual information about the phobias. Each subset of questions (spacing/research and interleaving/categorisation) led to a score, which was converted to a percentage for each participant.

5.2.2.3 Procedure.

Participants sat at individual desktop computers, and the researcher (who was also the class teacher) oversaw the session. After a general briefing, each student completed an on-screen consent form, followed by viewing the material presented in an order which depended on allocation to experimental conditions, which was done via random numbers. The main learning task overall took a mean of 7.9 minutes. As soon as they had completed the task, the software automatically initiated the test phase.

Ethics approval followed the school's own in-house framework; as a research-focused independent school with its own research centre, it had set up its own in-school ethics board with an academic panel providing oversight.

5.2.3 Results.

I used IBM SPSS Statistics for Mac OS, Version 25, to replace four missing values (replacement based on single imputation) which had been registered due to one participant stopping before the final four questions on the test; as less than 5% of the data were missing (0.76% of the total data set) the missing data were not analysed further; this method performs acceptably well for small quantities of missing data (Graham, 2009; Scheffer, 2002).

The mean percentage scores on the final categorisation test were broken down for the interleaved and blocked conditions as shown in Table 5.1.

Table 5.1: Table of mean scores in Study 1.

	Spaced (n = 15)	Massed (n = 16)	Interleaved (n = 15)	Blocked (n = 16)
MEAN	56.64	72.06	63.97	65.27

SD	16.60	13.33	12.86	20.56

A Levene’s test of homogeneity of variance was non-significant across both the dependent variable and design, suggesting that the variance was comparable across conditions – one of the assumptions of ANOVA. A between-subjects 2 (spaced vs. massed) x 2 (interleaved vs. blocked) ANOVA was carried out, and revealed a significant main effect of spacing; F (1,31) = 9.95, p = .004) with an effect size (partial eta squared) of -0.269, while interleaving did not have a significant main effect; F (1,31) = 0.005, p = .946).

Importantly, there was also a significant interaction between spacing and interleaving; F (1,31) = 6.60, p = .016), indicating that interleaving had a mediating or protective effect against the lower scores associated with the spacing intervention in the study (see Figure 5.3).

The ANOVA is not always preferred with percentage data, as there can be a chance of spurious results (Jaeger, 2008). However, these concerns have been raised in terms of categorical data, analysing responses to a fixed response question (e.g. yes vs. no). The present study obtained a score on a task, with each percentage representing a participant aptitude score rather than being a binomial distribution. Discrete responses (such as responses to multiple choice questions on the present task) are not continuous variables, but this issue is commonly overcome by re-scaling the data as proportions (Dixon, 2008). While I recognise the concerns that have been raised about this, I will follow the precedent in the current literature which features the use of ANOVA and MANOVA in similar studies of interleaving (e.g. Eglington & Kang, 2017; Rawson et al, 2015.

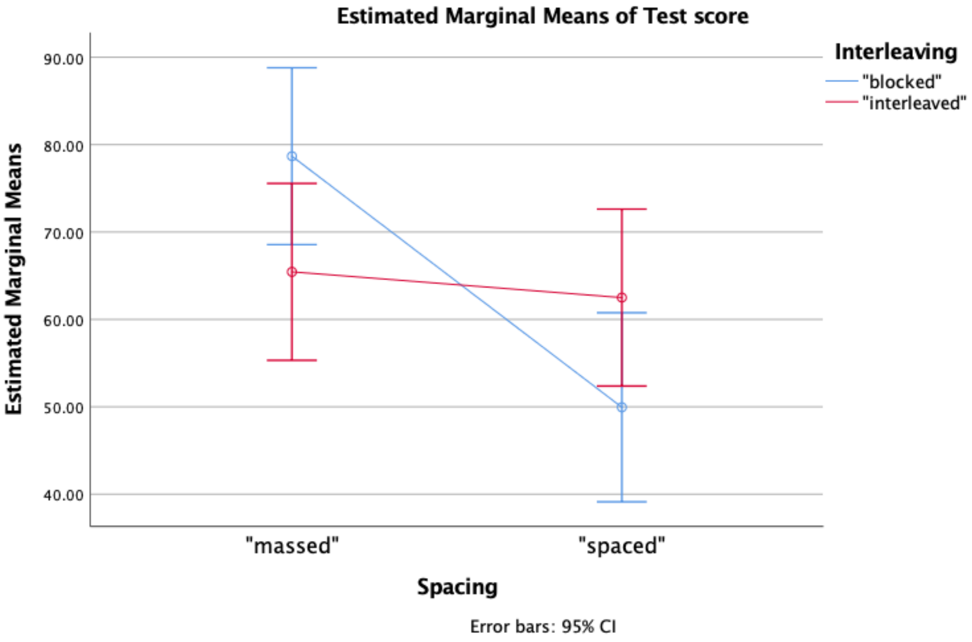


Figure 5.3: Interaction between interleaving and spacing

Another concern is that while ANOVA is a robust test, there is a risk of unreliable findings if group sizes are both small and non-normally distributed. Here, kurtosis in the dependent variable across the sample was calculated as +0.27 (SE = 0.821), well within an acceptable range (Doane & Seward, 2011), and visual inspection of a frequency histogram did not indicate either platykurtosis or leptokurtosis (see Figure 5.4). Therefore, despite the small sample size, there was no evidence of kurtosis, supporting the use of ANOVA here.

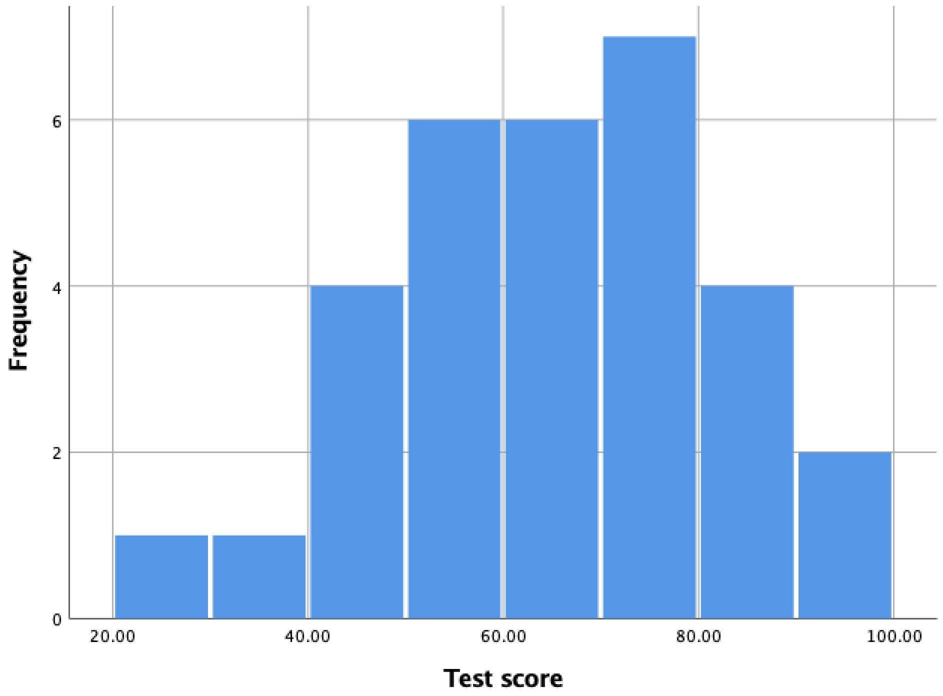


Figure 5.4: Distribution of participant scores

5.2.4 Interim discussion.

The findings of Study 1 show that spacing had a harmful effect on the immediate test, while the main effect of interleaving was neutral. The results fit with the idea that spacing is a ‘desirable difficulty’ with potential to impede learning over the short-term. Soderstrom and Bjork (2015) describe how such strategies often lead to performance being slower and more error-prone, but improve learning over longer intervals. However, with this in mind, the current study suggested that interleaving will not cause major impediments to new learning, if used by teachers to present new classroom concepts. What’s more, it appeared to mitigate some of the short-term harm caused by spacing out of material in another task that took place before and after. I will now discuss these issues in turn.

5.2.4.1 Effects of spacing.

Spacing, as discussed in previous chapters, is an evidence-based learning technique which suggests that delays between practice can be beneficial over the long term (Cepeda et al., 2006). The lack of a beneficial effect of spacing in the current findings could be explained in numerous ways.

One possibility is that the task and follow-up test were too short-term to observe the benefits, and thus the protocol simply measured the ‘difficulty’ element of spacing (a reduction in performance) without measuring the ‘desirable’ part, i.e. a boost to learning that may occur on a more long-term basis. If this is the case, teachers should be aware of the increased errors within a learning session that can result from distributed practice. These suggest inefficiency in the learning process as consolidation tasks would be more poorly done, and could have motivational impacts on teachers and learners alike. A way around this would be to follow the advice of Landauer and Bjork (1978) in beginning with very brief delays, and gradually increasing the retention interval (but cf. Roediger & McDaniel, 2011, who recommend fixed intervals between study sessions).

However, the design of the task clearly varied in some important ways compared to most laboratory studies of spacing. The spaced condition here should perhaps not be considered to be an example of distributed practice because there was no direct repetition of the initial information. Instead, the second task in the spaced condition involved further information. It may have reminded learners about the first element, but they were not directly required to retrieve or use that earlier information.

This meant that the group who experienced a delay experienced forgetting but no review, making it perhaps unsurprising that they performed more poorly on the test. This raises in important point for teachers – spaced *practice* is helpful, but a delay is not always helpful.

This idea conflicts with some evidence in the literature which suggests that simply spreading out teaching over time (without directly repeating it) *can* be helpful. Bird (2010) implemented spacing in the teaching of modern languages, with the same 9 classes being taught in either an intensive (one class per 3 days) or distributed (one class per 14 days) fashion. He found that a more spread-out study schedule led to better performance – but only over the long term: "after 60 days a significant amount of what had been learned during training was forgotten in the 3-day ISI condition but not in the 14-day ISI condition" (Bird, 2010, p. 646). This suggests that a delay between practice sessions can help to arrest forgetting. It's also notable that his study did not find a difference between the spaced and distributed conditions in a test just 7 days after the final study session. The difference only became apparent after longer retention intervals. This strongly suggests that the benefit of spacing was to slow the progress of forgetting (flattening the forgetting curve).

However, any such benefits were not apparent in the single session tested here, and it is impossible to say whether they would have occurred later. It is possible that a benefit would have emerged after a period of time. However, here, the retention interval was too short to determine whether this was the case; there was a delay between the first part of the spaced learning task and the test, but very little delay between the second part and the test.

It is also possible that the Bird (2010) task, as a study of English as a foreign language over 9 learning sessions, featured more review of material than was the case here. Learning a language is a cumulative process, and it is not possible to teach new language concepts without incidentally practising others.

The material in the current study may also have been too complex to benefit from spacing. Donovan and Radosevich (1999) found that spacing was not beneficial for complex tasks; complexity interacts with learner experience, and when learning a new concept, complexity for a learner can be high. Most tests of spacing avoid higher-order critical thinking skills, and feature extensive direct repetition e.g. of vocabulary or terminology. This raises a practical problem, for simple repetition of basic terms would not fit well with the norms of contemporary school teaching; it is also not the expectation of most school exams in science or social science subjects (though memorisation of terminology, as attempted by Hausman & Kornell, 2014, is perhaps an exception). Basic repetition would lack face validity as far as learners go, perhaps appearing boring and repetitive, and does not match what a teacher would typically cover in a real lesson.

The material used for spacing in Study 1 related to skills such as evaluation, rather than simple factual knowledge. A reduced benefit of spacing for a skills-based task fits with the idea that desirable difficulties interact with learner skill, as proposed by McDaniel and Butler (2011). However, there is some evidence that spacing can impact on critical thinking skills (see 5.3, below), at least by improving a foundation of factual knowledge, making this a less plausible explanation of the findings.

Overall, it appears that a simple delay is of little benefit to a classroom task, perhaps especially so when the task is complex and little direct repetition or retrieval is required. It is worth thinking, then, about how to overcome the mismatch between what appears to work in laboratory studies of spacing and the practicalities of classroom application. Firstly, it would be advisable for classroom studies of spacing to mimic more closely the schedules used in other research; in particular a longer-term follow-up test should be used. Secondly, if re-study of the same materials is not integral to the task (as it would be in, for example, tasks that involve use of the material from session 1 in order to complete session 2), then direct restudy should be built in. This fits with the recommendation of Rawson and Dunlosky (2011) that learners should first automate recall of concepts within a single study session, and later retrieve items three times at widely spaced intervals.

5.2.4.1 Effects of interleaving.

In terms of its main effect, interleaving appeared to be neither harmful or helpful in the present study. Given the importance of similarity discussed in Chapter 3, this could indicate that the examples used in the classroom task were insufficiently similar to be confused by learners, but insufficiently diverse for a blocked order to be superior. Here, a difficulty for classroom applications lies in the fact that it is difficult to know in advance how obvious the differences between items will seem to learners. As complete beginners to the topic, it might be assumed that the participants in the current study – and other learners like them – would find the three types of phobia confusable. More advanced students would likely find these concepts to be obviously distinct. The exact level of similarity between classroom concepts is not just a feature of the material itself, but depends on the learners' experience (see Chapter 2). This is something to consider going forward; even where benefits of interleaving are found, it may require teacher judgement to determine whether they will apply to a particular set of material with a particular class of students.

Overall, the lack of a main effect may be taken to imply that interleaving of the material was neither helpful or harmful, perhaps because the items were neither very similar not very distinct. However, again there were flaws (or at least differences compared to the lab-based evidence) in terms of the way interleaving was operationalised and tested. The way that the concept was applied in Study 1 may have stuck too closely to standard classroom practice in order to gain the benefits found in laboratory studies of the effect.

Firstly, the task used interleaved information about the three disorders – their diagnostic criteria, prevalence, and so forth. Such factual information is commonly taught in classrooms, and it would be relatively easy for teachers to interleave such information on slideshows or worksheet such that students experienced it in a mixed or interleaved order rather than categorised by concept. However, it differs from the inductive learning approach seen in many laboratory tests of interleaving, which tend to provide multiple examples of the same concept rather than providing supporting factual information of each one.

A second problem arises from the criterial test. Questions such as the following were used:

Which type of phobia varies in its frequency in different parts of the world?

1) Specific phobia

2) Social anxiety disorder

3) Agoraphobia

These items essentially form a memory test rather than a test of concept learning. It differs from the way that interleaved material has been tested in other studies; for example, Zulkiply et al. tested participants by presenting them novel case studies and asking them to categorise these into one of the previously-studied categories. Therefore, while fact-based multiple-choice tests are appealingly relevant to classroom contexts, they may fail to fully realise the benefits of interleaved learning, especially when it comes to its use in promoting transfer. For future classroom-based research, test items should focus on learners' ability to categorise example items. These could include both items (thus drawing, to a degree, on memory), and to categorise novel unstudied items (this drawing on transfer). This approach will be adopted in Study 2.

Going forward, what might be expected from future classroom-based studies into interleaving? It might be assumed, on the basis of the current study, that interleaving is neither harmful nor helpful in classroom settings. However, this conflicts with a body of previous research (see Chapter 3) which has found sizeable effect sizes in studies of interleaving. It seems more likely that the approach taken to applying the concept here has failed to tap into the potential benefits of the concept.

It is possible that the current study was underpowered. However, past studies with small sample sizes have still found an interleaving effect. Overall, it seems more likely that the way interleaving was operationalised was insufficiently focused on comparing and contrasting example items, and thus failed to fully test its utility for developing concept knowledge that will transfer to future examples. An interim conclusion for both spacing and interleaving, then, is that simply changing the order of study tasks to introduce delays (spacing) or to vary the presentation of factual information (interleaving) is not in itself likely to be beneficial. The benefits of both techniques depend on certain features (repetition of material in the same format in the case of spacing, and inductive learning of multiple examples in the case of interleaving) which constrain their application.

This latter point also implies that some skill and professional understanding will be required to apply the techniques productively, and that neither can be seen as a 'quick fix' which can be applied without an understanding of the underlying processes.

5.2.4.2 Interaction between interleaving and spacing.

In terms of the second research question, the findings of this experiment support the idea that spacing and interleaving can interact within a classroom situation as – surprisingly – interleaving one phase of the task appeared to attenuate the short-term difficulties caused by spacing the other. The interactions observed between spacing and interleaving in the study are novel, and notwithstanding the flaws in how the two concepts were operationalised, this finding raises some useful issues going forward.

On the face of it, this result does not fit well with Birnbaum et al. (2013)'s finding of a harmful interaction between spacing and interleaving, but again it is important to note the somewhat unusual way that the two concepts were operationalised in order to maintain ecological validity in a classroom setting. Here, the two interventions were used in different task phases within a short lesson, and spacing of one element therefore did not prevent discriminative contrast in the other.

Why would adding two desirable difficulties make the task easier in this case? It is conceivable that the difficulty of one task focused attention on the other; a variation of the attenuation of processing theory (see Dellarosa & Bourne, 1985) which suggests that spacing results in more attention being paid to a task. However, it is unclear why increased attention would not result in a main effect of spacing as well.

Another possibility is that the learners benefited from forming conceptual links across the two task phases, with interleaving prompting learners to seek outside meaningful connections beyond what was on a single screen in a way that did not happen when the concept learning phase was presented in blocks. Such connections could be harder to recall or perceive if the first learning phase had been massed and therefore appeared complete (an instance of the Zeigarnik effect; Zeigarnik, 1927). This explanation could be further examined in future by comparing the learning of sets of information with varying levels of conceptual similarity, on the basis that meaningful links could not be formed if items were unconnected (similarly to the findings of Hausmann & Kornell, 2014, whereby material from different domains could not be productively contrasted).

Combining these ideas, we can ask what learners were paying attention to, and why. The attention-attenuation hypothesis of interleaving discussed by Carvalho and Goldstone (2015; 2017; see Chapter 3) suggests that learners don't just pay *more* attention on tasks that feature interleaving, but that their attention is directed to *different* things – to differences in some cases, and to similarities in others. It is possible that the increased attention and covert problem solving required in an interleaved task (as learners attempt to make sense of the diverse information they are viewing) increases the extent that they focus on links between the information, helping them to categorise the new information in ways that later made it easier to retrieve.

There are other examples of how a process that modifies attention in a learning task can have a protective effect. Mueller and Oppenheimer (2014) found that students who took notes by hand rather than on laptops had briefer notes but later showed better recall. The note-writing students appeared to be summarising, prompting deeper processing of the new information. Baird et al. (2012) found that when doing a creativity test, learners who did a simple task during a short break – promoting mind-wandering[15] – subsequently did better than those who did either a demanding task or no task at all. These studies show that a subsequent task (note taking or mind-wandering, respectively) can affect an earlier one. Indeed, Peterson and Wissman (2020) found evidence suggesting that some of the benefits of review tests may be that they keep learners on task and reduce mind-wandering.

However, in terms of applications to practice it is also slightly concerning that performance on a task that is ostensibly finished (in this case the massed phase of the classroom task) is affected by a subsequent task, and it raises the important issue that tasks involving desirable difficulties cannot be seen in isolation in a classroom context. A typical lesson may last around an hour or more, and it appears that the outcome of a ten-minute activity is going to be affected not just by the design of that task itself, but also of what comes before or after it.

But to determine whether this is actually the case, further investigations might need to ask learners what they were doing during the task. Investigations that looked at the benefits of techniques such as spacing and interleaving when they are scheduled before or after demanding tasks or at different points of the day would be well worthwhile, though they are beyond the scope of the present thesis.

[15] *Mind-wandering* is the state where a learner’s attention lapses as they become lost in thought, rather than focused on external stimuli.

Perhaps most importantly for the purposes of classroom application, the results suggest that the small-scale spacing that will inevitably occur when examples are interleaved in the classroom (as outlined above – see section 5.1) is not a reason to avoid the technique. If anything, interleaving appears to have a protective effect, and it may be the case that it prompts learners to pay attention to similarities across different parts of a task, viewing the material as a problem to be solved. As such, it could also be a relatively simple method for boosting active learning – an area of teaching practice that is generally sought after.

5.2.4.3 Evaluation of methodology.

Some of the main lessons of the current study are methodological, and indeed, the broader purpose of this as a pilot study was to explore ways in which interleaving and spacing can be applied to authentic classroom tasks. To that end, the study raised several issues, some of which relate to the operationalisation of the two concepts and have already been discussed. A tension exists between the need for a classroom task to be educationally helpful and authentic, and the need for a degree of experimental control. While this study emphasised the former, it is important that tasks do not stray too far from the research literature if the benefits that have been demonstrated in the lab are to be successfully applied to practice.

In particular, it is important to analyse those established benefits and consider which of them are integral to the effects studied – in this case the interleaving and spacing effects.

Spacing is unlikely to be beneficial unless there is an opportunity for delayed or distributed practice, and what benefits do arise are unlikely to be observed within a single lesson. For this reason, it is a problematic concept to apply in the field, especially in the context of exam-based courses where the timescale is strictly limited, and when both students and teachers of those courses will already be doing a lot of work to tackle any gaps in their understanding, potentially confounding any experimental manipulation that is tested over longer timescales. The subject matter and nature of the test carried out is more flexible as the spacing effect is a very broad phenomenon, but field research into how the effect pertains to skills-based concepts on exam-based courses is very limited indeed (but see Kapler et al., 2015).

In terms of interleaving, a key concept that arises from the literature is the ability of learners to engage in concept learning by meaningfully discriminating specific examples – example images or very short texts. Indeed, such a methodology was universal in the studies reviewed in Chapter 3, and therefore cannot be considered optional in methodology as it is applied to the field. Any task which works on the basis of memory recall or of longer/larger items cannot be considered to be evidence-based – at least, not with reference to the literature on interleaving.

This is problematic, however, as the education community more broadly often perceives interleaving as a manipulation which mainly targets memory, and which can be applied to entire topics. Consider these recent and influential examples:

- The Sutton Trust report, entitled ‘What makes great teaching’ (Coe et al., 2014) states that “interleaving with other tasks or topics leads to better long-term retention and transfer of skills.” (p. 17). This both implies that interleaving mainly targets factual retention, and misleadingly suggests interleaving longer sections from a course (such as entire tasks or topics) rather than of short verbal or visual items/examples. It does usefully note the relevance to transfer, but the what exactly is meant by ‘transfer of skills’ in this context? Skills, as they are usually referred to in education, tend to relate to either language skills (e.g. reading) or thinking skills (e.g. analysis), and there is a lack of evidence in either area.

- A 2019 article in ‘SecEd’ (Secondary Education) magazine states that interleaving “can help boost students’ long-term memories and retrieval of key information”. This very much places the focus on factual retention rather than on concept learning and transfer.

- An article in the Times Education Supplement by Tsabet (2018) says the following: “One method that can be used to improve the chances of committing learning to long-term memory is ‘interleaving’. This replaces the traditional method of block learning, where students cover one topic at a time”. Again, the emphasis placed on LTM may misleadingly give the impression that interleaving is primarily a mnemonic technique, and implies that entire topics should be interleaved (perhaps by mixing together the lessons from different topics without any interleaving of specific examples or short texts). The TES article also focuses on the learning of skills such as reading and listening; the evidence for interleaving this kind of skills learning is absent from the review in Chapter 4 (but see section 5.3, below).

Such examples raise major questions about how interleaving is being (mis)interpreted throughout the teaching profession. There are more helpful and accurate examples, such as the report of the Deans for Impact (2015) which states that teachers can “interleave (i.e., alternate) practice of different types of content...it’s more effective to interleave practice of different problem types, rather than practice just one type of problem, then another type of problem, and so on” (section 2, bullet point 5). However, such examples tend to be brief, not fully exemplified, and tend to omit guidance on boundary conditions of the interleaving effect. Together with the evidence from Study 1 above, it is clear that there is a need for guidance that makes it easier for teachers to understand what is meant by interleaving, and therefore to apply it successfully to the classroom.

Interleaving does, however, have more potential for application to skills-based tasks, given its role in transfer. Indeed, it would have been interesting to see the outcome if interleaving had been used in the research-based phase of the current experimental task, where evaluation of the research study was carried out. For beginner students, the idea of evaluating a piece of psychology research is a key skill, and one that they need to understand conceptually and will come to be able to do with other studies (i.e. to transfer their learning). It would therefore be a useful context to consider interleaving, and this will be addressed in the next study in this chapter.

In terms of timescale, it seems integral to interleaving that the items are short – no more than a few sentences. A methodological benefit of this is that unlike spacing, implementation and testing of interleaved practice does not necessitate long delays by design; concept learning can undoubtedly be slow and gradual, but the benefits of interleaved presentations have been seen within a single study session in many of the experiments reported earlier. This could be because – again unlike spacing – it is not a manipulation that targets forgetting as it occurs (which spaced practice does by interrupting the forgetting process). Instead, it attempts to boost meaningful understanding, something that should lead to both immediate and long-term improvements in performance.

That is not to say that it wouldn't be helpful to study interleaving over the long term. Would classes who had been taught using interleaved examples do better in a later test or exam? Given that meaningful understanding is more durable in LTM, it might be expected that short-term advantages of interleaving would widen as time went on. Of course, it would be useful to test this, but again a barrier with exam classes lies in both the ethical and practical difficulties of manipulating factors that might interfere with attainment, as well as the possibility that extraneous study activities might lead to error in the findings.

In terms of the suitability of task for larger-scale research and for teaching purposes, most of the improvements discussed above can easily be incorporated within the PsyToolkit software used in Study 1. This task used short verbal items, which could be replaced by short verbal examples. The software is best for reasonably brief interventions, but as entire teaching tasks and topics are *not* the optimal focus of interleaving as a strategy, this does not present a barrier to this method of presentation.

Teachers may not want their pupils to be online and at a screen for an entire lesson, in which case the initial interleaved task could be conducted using the software, and a later test done in a more traditional format.

Overall, this classroom-based study found that while spacing caused short-term harm to learner performance, interleaving helped to counteract this problem. As such, it is suggested that combining interleaving and spacing may prompt learners to seek meaningful connections during concept learning. And it also found that interleaving alone was relatively benign; despite being a desirable difficulty, it did not cause any particular harm to learners over the short-term. This is useful, because learners and teachers alike may be put off from applying desirable difficulties to their study habits and teaching practices respectively.

Granted, though, such metacognitive errors may be due to *perceived* learning rather than actual progress, and student perceptions were not gauged in the present study. It will be useful, in future studies, to include a metacognitive measure at some stage in the task. For example, students could be asked to predict their own performance prior to completing the criterial task (see Chapter 3).

It will also be useful to explore the extent to which interleaving can be applied to *skills* in the classroom, and this will be the focus of Study 2.

5.3 Study 2 – Classroom Interleaving of Skills Learning

5.3.1 Introduction to Study 2.

Study 1 piloted a methodologically novel approach to investigating interleaving and spacing, and applied the concepts to a set of authentic classroom tasks. It was found that interleaving had a neutral effect over the short-term, in comparison to spacing which can cause short-term harm to performance. What’s more, interleaving appeared to have something of a protective effect against the difficulty caused by spacing. This could be because it prompts learners to seek out meaningful connections across current and previous examples, and to take a more attentive and active approach to their learning. However, there are limitations in the way that the two concepts were applied, raising the difficulty for the practitioner of navigating between classroom authenticity on the one hand and applying evidence-based techniques as they are conceptualised in the research literature on the other.

The present study followed some of the same procedures as Study 1. Again, secondary school Psychology classes were used, and authentic materials were presented to them. Interleaving was a focus of the investigation, with an intention to increase the concordance between the way it is applied in the field and the way it has been operationalised in the laboratory, and in particular to use sets of examples amenable to inductive learning, rather than supporting factual information as was used in Study 1.

However, Study 1 was limited in terms of its scope and sample. With only 31 pupils tested it served mainly as a pilot, helping to explore the methodological options available. It was necessary to follow this with more in-depth research which did not risk being underpowered. The following sections report on one such study, carried out across two Scottish secondary schools rather than a single institution.

Another limitation of Study 1 was that a baseline measure was not used. This was not deemed necessary given that students were new to their course and prior concept knowledge was assumed to be absent. However, it is difficult to be sure how much learners know about the task at hand before presenting the information. In the case of Study 1 with its information on phobias, these are (at least to some extent) known about outside of the classroom. In addition, with any study that tests learners on authentic course material – even on topics that they have not yet covered – there is a chance that participants will have done advance reading or simply show unusual levels of background knowledge and interest (given that they chose to study the course in the first place), potentially leading to bias especially with between-groups designs. I therefore decided that follow-ups should include a pre-test phase. This would both check for levels of prior knowledge and help to gauge any improvements made across the task.

Educators have also increasingly come to recognise that as well as applying the techniques associated with successful learning, it is valuable to also discuss and raise awareness of these techniques in class, so that students develop a metacognitive understanding of how and when to apply them (EEF, 2018). Doing so can help to build a classroom vocabulary around effective learning. I was aware that teachers take a similar approach when it comes to higher-order skills, with school classrooms often displaying posters that show skills in a pyramid form (based on Bloom's taxonomy; Anderson et al., 2001; Bloom et al., 1956) or with definitions. It is interesting therefore to consider how a learner's ability to make a verbal generalisation of what is meant by analysis and evaluation could link to their ability to use these skills.

What's more, self-explanation is often considered an evidence-based study strategy in its own right, and a desirable difficulty along the same lines as spacing and interleaving (see Chapter 1). In a meta-analysis, Bisra et al. (2018) concluded that self-explanation prompts are a potentially powerful intervention across a range of instructional conditions, with a pooled effect size of 0.55.[16] As an exploratory element of the study, then, I decided to include self-explanation as one of the research variables, potentially allowing for exploration of both the utility of the technique and its time efficiency within the broader task protocol.

A further issue that I considered was the selection of material. Facts about psychological disorders, as used in Study 1, are clearly amenable to interleaving as has been seen in the work of Zulkiply and colleagues (e.g. Zulkiply et al., 2012). However, establishing declarative

[16] These authors also recommend computer-generated rather than instructor-provided prompts to facilitate independent learning, which is interesting given that the current study presented the task via a computer interface.

knowledge is not only challenge facing school psychology teachers – or even the main one. The school psychology exams in Scotland include a heavy emphasis on the *higher-order skills* of analysis and evaluation. The Higher Psychology course specification (SQA, 2018, p. 49) defines these skills as follows:

- “Analysing: candidates develop the skill of analysis when they compare and contrast theories, concepts and studies, and when they provide implications, applications and conclusions based on their understanding of psychological topics and studies.”
- “Evaluating: candidates develop the skill of evaluation when they identify and explain strengths and weaknesses of theories, concepts and studies based on valid criteria.”

These can be contrasted with the *lower-order* skills of description and explanation (SQA, 2018, p. 49):

- “Understanding (explain): candidates develop understanding when they explain ideas or concepts.”
- “Describing: can the candidate recall or remember the information?”

In practice, psychology teachers often struggle to teach the difference between evaluation and description to their classes, with many students tending to state strength and weaknesses as factual information (for example by saying “the study was a lab experiment”) rather than fully evaluating a study or theory as defined above. Likewise, it can be hard for learners to discern the difference between analysis and explanation – both involve elaborating on basic facts, but in the case of analysis this is deeper and more critical, and tends to involve making links outside of the material that is under consideration (for example, understanding/explanation could involve saying how a theory works, while analysis could involve contrasting its aims with those of another theory).

Desirable difficulties are not often associated with these higher-order skills – quite the opposite. For example, Coe (2020) argues that it may be difficult for the strong lab-based benefits of retrieval practice to fully generalise to the complexities of information covered in the classroom, while Van Gog and Sweller (2015) argue that its benefits may disappear altogether with complex material (but cf. Karpicke & Aue, 2015). In terms of spacing, the bulk of the seminal research has been done on simple word lists (Fishman et al., 1968; Landauer & Bjork, 1978; Sobel et al., 2011; Zechmeister & Shaughnessy, 1980), or on isolated trivia facts (Cepeda et al., 2008), as discussed in previous chapters.

More broadly, evidence-based approaches are often seen as useful only in as far as they promote lower-order skills (describe and explain, from the ones defined above). For example, APA's '20 top principles for PreK-12 Teaching and Learning' (APA, 2015) notes that much of learning involves practice in order to get information into long-term memory[17], but also states that in doing so, "transfer of practiced skills to new and more complex problems is increased" (p. 11). This idea – that higher-order skills are improved by building up lower-order knowledge – is the essential principle of Bloom's Taxonomy (Bloom et al., 1956; see above), a skills hierarchy which states that each skill is founded on another set of more basic skills and knowledge, and develop on the basis of these as a foundation. It might therefore be assumed that the best use of desirable difficulties is to boost factual knowledge, so that higher-order skills can subsequently develop; this idea was certainly endorsed by Willingham (2009), who stated that "factual knowledge must precede skill" (p. 19; cited by Agarwal, 2019).

However, there is evidence which shows that desirable difficulties can have a more direct benefit to higher-order skills. Kapler et al. (2015), in a study that simulated a classroom

[17] "Principle 5: Acquiring long-term knowledge and skill is largely dependent on practice."

situation, found that an 8-day delay led to superior learning than a 1-day delay for both factual and higher-order learning. And Foot-Seymour et al. (2019) found that spaced practice improved schoolchildren’s ability to both remember and to critically assess the validity of online sources – a useful transferable skill when it comes to reducing the harm caused by misinformation online.

Further, Agarwal (2019) found that while retrieval practice via quizzes consisting of higher-order and mixed (both factual and higher-order) questions boosted later higher-order skills, quizzes that were purely fact-based did not. This means that the boost to higher-order thinking found in her study cannot be put down purely to learners having developed a sound factual foundation. Instead, according to her analysis, the benefit came from their having had practice of retrieving information in relevant ways and contexts, with transfer from higher-order practice to a higher-order assessment being easier than transfer from more distant material.

Although it conflicts with Bloom’s taxonomy, Agarwal’s interpretation fits with the transfer-appropriate processing explanation of desirable difficulties (see Chapter 1), and also with mainstream theories of near and far transfer whereby the more dissimilar the practised material is to a later situation, the harder it is to transfer learning to that situation (see Barnett & Ceci, 2002, for a detailed taxonomy of the key ways in which material can differ and how this affects transfer). Successful application of desirable difficulties, it appears, is not (just) about forming a stronger foundation of factual knowledge, but about practising whatever retrieval or transfer processes will later be required, and doing so in a way that resembles the later conditions of use (which in non-academic contexts will typically be varied, involve retrieval, and feature lengthy time delays from one situation to the next; Bjork, 1994; Bjork & Bjork, 2019).

However, it should be noted that the above studies interpret higher-order thinking as generic critical thinking, and that they relate to spacing and retrieval practice. To my knowledge, no previous studies have looked at desirable difficulties in the context of evaluation and analysis of experimental research evidence specifically, and none of the studies mentioned above (or those reviewed in Chapter 3) involved the interleaving of such skills.

Out of all of the desirable difficulties, interleaving would appear to be especially appropriate for developing higher-order skills in students. While it does have a mnemonic benefit, the technique is also based around building variation into a practice situation such that learners are better able to transfer what they have learned to novel examples (see Chapter 3, and note that many studies reviewed in the chapter included a *transfer* as well as a *memory* condition).

The importance of this variation is easy to see in an academic situation such as an exam. As noted by Rohrer et al. (2015) in the context of maths learning, an exam tends to feature questions that are mixed and unlabelled, meaning that the student approaches each question without knowing what they are going to be asked about, or which skills may be relevant. An interleaved practice session better resembles the testing situation than does a blocked one, leading to a difficulty which is appropriate to the processing that will later be required.

This ability to transfer learning is fundamental to the skills which teachers seek to develop, at least in the study of psychology. For example, if a student is to identify a weakness of a previously-unseen research study (*evaluating*, as defined above), then they will need to transfer what they have learned when studying other prior examples of research. Indeed, the difference between *knowing* strengths and weaknesses on the one hand and being able to *evaluate* a research study on the other is fundamentally about this kind of transfer. The fact that students so often fail to transfer what has been learned in one situation to distinct but conceptually similar situations led Perkins and Salomon (1998) to describe this as a key but neglected problem in education. As they put it:

> “A great deal of the knowledge that students acquire is ‘inert’ or ‘passive’. The knowledge shows up when students respond to direct probes, such as multiple choice or fill-in-the-blanks quizzes. However, students do not transfer the knowledge to problem solving contexts where they have to think about new situations” (p. 23).

If learners are to take what they have learned in school and use it in real-life situations, they will benefit from having experienced learning situations which don’t just teach them concepts but also prepare them to use this learning in varied, unpredictable real-world settings. Given the benefits of the technique in terms of memory and time efficiency, it is important to ascertain whether interleaving of examples has the potential to achieve these ends.

The present study, then, aimed to do the following: follow up Study 1 with a larger sample; apply interleaving to the skills of analysis and application in the context of authentic school material; and determine the extent to which past findings on the potential of interleaving to promote transfer can apply to real school classroom contexts.

The research questions and predictions are as follows:

(a) Will interleaving will be advantageous compared to blocking in a high school skills-based task?

On the basis of the research discussed above and earlier in this thesis, there is ample evidence that interleaving can be superior to blocking when it comes to learning new concepts. I predict that the advantage will generalise to higher-order skills.

(b) Will self-explanation be advantageous on the same task?

Self-explanation is also a desirable difficulty as discussed above, and so an advantage of this technique as an intervention compared to a control condition is predicted.

(c) How will interleaving and self-explanation interact?

It seems possible that the opportunity to reflect on and describe the higher-order skills under investigation will affect the advantage (or otherwise) of comparing and contrasting them in an interleaved order. I therefore predict an interaction between these two desirable difficulties, in line with the interactions previously found between interleaving and spacing. The nature of this possible interaction cannot be confidently predicted due to a lack of prior evidence.

(d) Will metacognitive awareness of interleaving be inaccurate?

Prior research has suggested that learners tend not to appreciate the benefits of interleaving compared to blocking. It is therefore predicted that interleaving will be less frequently chosen than blocking when participants are asked which strategy is more effective, and furthermore that predictions of performance will be inaccurate.

5.3.2 Method.

5.3.2.1 Participants.

99 participants took part, all school pupils and with a modal age of 17 and a mean of 16.8 years. They were all pupils at schools within two local authority areas in Scotland. All were taking a one-year course in Higher Psychology, and were a few weeks into the course at the time of data gathering. A power analysis had indicated that eighty or more participants would be required. Nine participants did not complete the task beyond the demographic questions and the pre-test phase (which in most cases was due to technical difficulties getting online, given

that the task was supervised and none announced an intention to withdraw); their data were discarded listwise, leaving 90 complete records for analysis.

The task took place in school classrooms during normal lesson time, by arrangement with the teachers involved and following ethics approval from the researcher's university, the local authorities involved, and the specific schools. Each session was supervised by both the researcher and the normal class teacher. The researcher waited until all pupils had arrived (minus any absences) before beginning. After a briefing and explanation of the participants' rights, each was given a link to the online task. They accessed it using school computers or tablets, or (in a few cases) via their own internet-enabled mobile phones.

5.3.2.2 Design and materials.

The study used a 2x2x2 mixed, fully-crossed factorial design (Judd, Westfall & Kenny, 2017). The between-participant factors were self-explanation vs. no self-explanation (Factor A), and task order (Factor B). The within-participant factor was interleaving vs. blocking (Factor C).

In addition to the main independent variables of interest (interleaving and self-explanation), Factor B was included to investigate possible main effects or interactions caused by the counterbalancing of the order in which different groups of participants engaged in interleaved or blocked practice.

The dependent variable was the participants' scores on the criterial task (see below). Task performance was sub-divided between skills which had been practiced via interleaved and blocked examples, with each part of the task resulting in a score out of 12, and an overall possible score of between 0–24.

The software randomly allocated participants to conditions. In the case of Factor C, interleaving vs. blocking, it randomly allocated participants to two groups (which I will henceforth refer to as Group 1 and Group 2), and this affected which condition (blocking or interleaving) they did first. Group 1 did the first part of the learning task in an interleaved sequence (different skills alternated) and the second part in a blocked order, while Group 2 studied the same learning tasks in the same order, but for this group the first part was presented in a blocked sequence (different skills grouped together), and the second in an interleaved order. By dealing with Factor B, task order, in this way, there were no nested factors in the analysis.

The task involved seeing statements which had been chosen as examples of the skills description, explanation, evaluation and analysis. One example is as follows:

“As all of the participants in this study were male, it is impossible to generalise the results of the study to women, limiting the value of the original findings.”

(One example of evaluation used in Study 2)

Each example related to a classic psychology study widely taught of the Higher Psychology course. The full set of examples is shown in Appendix 9.

In the main (learning) phase, the task order varied depending on which group participants had been allocated to (i.e. Factor B). Group 1 first saw examples of the skills *description* and *evaluation* (labelled 'skill set 1' in Figure 5.5) in an interleaved order, and Group 2 saw examples of these skills in a blocked order. Five examples of each skill were shown. The same procedure was then repeated with the skills *explanation* and *analysis* (labelled 'skill set 2' in Figure 5.5) with the allocation to condition counterbalanced, such that Group 1 saw examples of the skills *explanation* and *analysis* in a blocked order, and Group 2 saw the same examples of the same skills in an interleaved order (see Figure 5.5). This way, all participants saw the same twenty examples, and all participants experienced both the interleaving and blocked orders (Factor C) but with different examples. After the main learning task but before the criterial test, participants were randomly allocated to either the self-explanation intervention, or to the control condition of Factor A.

The design ensured that every participant benefited from seeing every example skill – a feature that was deemed important for ethical reasons.

It was then possible to gain an overall score for both interleaving and blocking for each participant, based on the questions in the criterial test that linked to the skills that had been interleaved (12 out of 24) and those which had been blocked (12 out of 24).

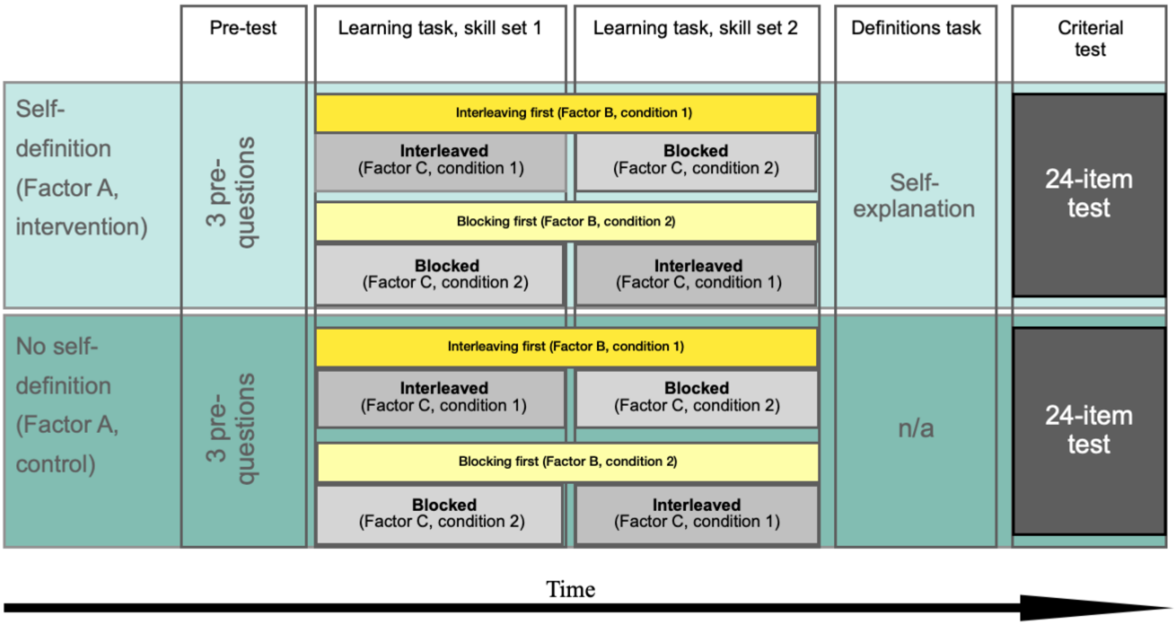

Pre-test
Learning task, skill set 1
Learning task, skill set 2
Definitions task
Criterial test
Self-definition (Factor A, intervention)
3 pre-questions
Interleaving first (Factor B, condition 1)
Interleaved (Factor C, condition 1)
Blocked (Factor C, condition 2)
Blocking first (Factor B, condition 2)
Blocked (Factor C, condition 2)
Interleaved (Factor C, condition 1)
Self-explanation
24-item test
No self-definition (Factor A, control)
3 pre-questions
Interleaving first (Factor B, condition 1)
Interleaved (Factor C, condition 1)
Blocked (Factor C, condition 2)
Blocking first (Factor B, condition 2)
Blocked (Factor C, condition 2)
Interleaved (Factor C, condition 1)
n/a
24-item test
Time

Figure 5.5: Graphical summary of Study 2 design.

5.3.2.3 Procedure.

The experiment was based around a task designed on the PsyToolkit website, as referred to in Study 1.

After an ethics screen and online consent form, all participants were asked to judge three example statements according to which skill it exemplified (description, explanation, evaluation or analysis) using a multiple-choice test. This was the pre-test phase of the study protocol. I avoided using wording such as “this was a strength...” so that participants would have to judge the statements on the information included rather than surface features.

Participants then saw examples of skills as they related to key studies, as described in the previous section. For each example, participants were asked to name the study described. This was done to ensure that the students paid attention to the examples, and differs from other research into interleaving which tests participant after every example, but is more in line with standard teaching practice where testing and review tends to occur subsequently to the initial presentation of material, and therefore maintains high ecological validity. The studies used were related to the Higher Psychology course, most deriving from mandatory content, and the examples were devised by the researcher, based in part on the textbook by Firth (2019b; see Appendix 9 for the full set of examples used).

After all of the examples had been shown, participants in the self-explanation condition were prompted to make a generalisation between the examples via the following instruction: “as best you can, explain what two or more examples of [skill] have in common”, followed by a text box. This was repeated for all four skills (describe, explain, analyse, evaluate). The remaining participants acted as a comparison group and did not do this stage, making the whole task slightly shorter overall for those allocated to that condition.

Next came two metacognitive questions. Participants were prompted to make an estimate of which strategy was better – interleaving or blocking – as they had tried both. This was worded as follows: “You might have noticed that for some examples you have seen, the skills/command words were alternated, while other examples appeared in a block of five of the same type. On reflection, which order of examples do you think was more helpful to your learning process?” They were also asked to estimate their later overall test performance on a scale of 25% (guessing only) to 100% (perfect performance).

Finally, there was a test phase. A further set of multiple-choice test items were presented, each with examples for the participants to categorise. Some of the examples had been shown in the main phase of the experiment and others were novel. After 12 items, participants were given an interim total score, though they were not given item-level feedback on their answers. In the test phase, half of the items overall drew on the same skills that the participants had practised in an interleaved order, and half drew on the skills that they had practised in a blocked order.

At times, some participants had questions, and these were answered by the researcher. The most common question was that they did not recall the name of the researcher who had done a particular study (e.g. Asch, Milgram; see below). As this was not essential to the task, they were told that it was fine to guess, or to type the first letter of the researcher’s name only, or write ‘don’t know’.

At the end of the task, participants were directed to a download link for a free ebook about how to study (Firth, 2018b), which was intended both as an incentive for participation and to ensure that their time spent on the task had a long-term educational benefit.

5.3.3 Results.

5.3.3.1 Learning task.

The three pre-test questions were used to establish a baseline competence, which was important given that the students were already a few weeks into their course at this stage. Each took the form of a 4-item multiple choice question, and the correct response rate overall for all

90 participants who completed the learning task and criterial test was 38.1%, n = 90 (if 9 additional participants who stopped before completing the task are included, the pre-test percentage becomes 39.7%, n = 99). Given that a value of 25% would represent picking answers at random, this indicated a relatively low level of relevant prior knowledge and skill. No further analysis of the baseline scores was undertaken.

Descriptive statistics for each experimental condition are shown in Table 5.2.

Table 5.2: Descriptive statistics for each condition of Study 2.

Type of Practice (Factor C)	Self-Explanation (Factor A)	Task Order (Factor B)	Mean	SD	N
Blocked Practice	No Self-Explanation	Blocked First	5.06	2.01	18
		Interleaved First	4.68	2.03	22
		Total	4.85	2.01	40
	Self-Explanation	Blocked First	5.53	2.24	17
		Interleaved First	5.03	1.91	33
		Total	5.20	2.02	50
	Total	Blocked First	5.29	2.11	35
		Interleaved First	4.89	1.95	55
		Total	5.04	2.01	90
Interleaved Practice	No Self-Explanation	Blocked First	5.33	2.45	18
		Interleaved First	5.09	1.80	22
		Total	5.20	2.09	40
	Self-Explanation	Blocked First	6.71	2.69	17
		Interleaved First	5.70	1.78	33
		Total	6.04	2.16	50
	Total	Blocked First	6.00	2.62	35
		Interleaved First	5.45	1.79	55
		Total	5.67	2.16	90

Although skewness and kurtosis were found to be within acceptable bounds for both the interleaved group and the blocked group (all z-scores for skewness < 0.890 and all z-scores for kurtosis > -1.295 and < 0.483; Doane & Seward, 2011), a Shapiro-Wilk test of normality of distribution (Shapiro & Wilk, 1965; Razali & Wah, 2011) was borderline (p = 0.051) for the blocking condition and significant (p = 0.016) for the interleaving condition. Although this test is commonly used for smaller sample sizes, it is sensitive for larger samples and is the most powerful test of normality (Razali & Wah, 2011). It was thus judged that the data were not normally distributed, and bootstrapping would be appropriate as a sensitivity analysis (Efron & Tibshirani, 1993; see below) for Factor C.

Using the original data set and with IBM SPSS Statistics for Mac OS, Version 25, I carried out a mixed ANOVA. This revealed yielded a significant main effect of Factor A, self-explanation F(1, 86) = 4.28, p = .042), with participants who had attempted to define the skills (M = 11.12, SD = 3.11) scoring higher on the criterial test than those who had not (M = 9.87, SD = 3.17). The main effect of Factor B, task order, was nonsignificant, F(1, 86) = 2.46, p = 0.12. However, the main effect of Factor C, interleaving vs. blocking, was significant, F(1, 86) = 4.46, p = .038, indicating that participants scored higher on tasks where they had engaged in interleaved practice (M = 5.67, SD = 2.16) than where they had engaged in blocked practice (M = 5.04, SD = 2.01) of the relevant skills. No significant interaction effects were found (all *p*-values > .336). Specifically, there was no interaction between definitions and task order (Factors A and B); F(1,86) = 0.434, p = .512, or between task order and interleaving (Factors B & C); F(1,86) = 0.1, p = .753, or between definitions and interleaving (Factors A & C), F(1,86) = 0.932, p = .337, or between all three factors (A, B & C), F(1,86) = 0.287, p = .594.

The difference between self-explanation (Factor A, intervention) and the control group is illustrated in Figure 5.6. I further separated out those participants who had completed a satisfactory attempt[18] at defining the skills in order to find what effect, if any, defining the skills (had on overall task time. For this subgroup (n = 31), the average overall task time was 26.35 minutes, compared with 23.58 minutes for all other participants (n = 59); this was found to be a significant difference using a one-tailed t-test (t (88) = 1.77, p = 0.04).

[18] 'Satisfactory' here means a comprehensible and fully written out definition, but it need not have been perfect or even correct. By contrast, some in this condition wrote 'don't know' or similar, which was considered an unsatisfactory attempt.

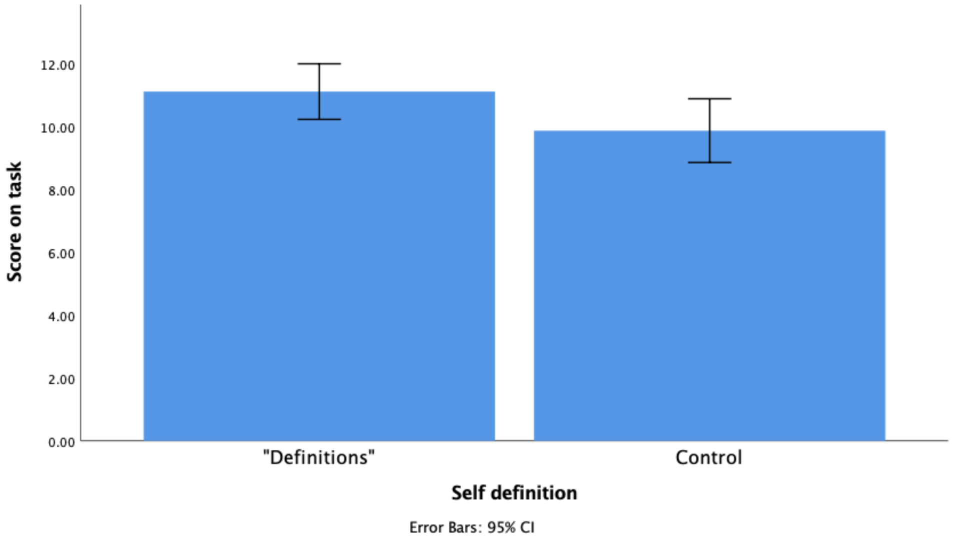


Figure 5.6: Difference between self-explanation group and control group (Factor A).

5.3.3.2 Sensitivity analysis.

Bootstrapping involves resampling on the basis of the observed samples, and supports more reliable generalisation from the sample to the population (Efron & Tibshirani, 1994). In effect, it treats the existing sample as if it were a population, and then derives a large number of further samples from this population (Field, 2018). This allows a standard error to be calculated in a way that does not violate the assumptions of parametric tests, and it is therefore a straightforward way to deal with non-normally distributed samples. It allows use of parametric tests in situations where there may otherwise be concerns about meeting the assumptions of those tests (Efron & Tibshirani, 1993; Field, 2018). There are a variety of methods of bootstrapping with different advantages and disadvantages (see Wright et al., 2011, for a discussion); I chose to follow the bias corrected and accelerated method as recommended by Puth et al. (2015). As there were no significant interactions, I carried out sensitivity analyses on the main effects of Factor C using a t-test.

Using this bootstrapping (1000 samples) sensitivity analysis for Factor C (blocking vs. interleaving), I again found that scores in the interleaving condition were again significantly higher than scores in the blocking condition, t (89) = -2.17, p = 0.032. These results are in line with the findings from the original data, supporting the robustness of those findings.

5.3.3.3 Metacognitive questions.

In terms of the metacognitive questions, two participants did not enter a predicted score, and the missing values were replaced using multiple imputation using the Monte Carlo Markov chain (MCMC) method (see Lin, 2010). Multiple imputation is a process that uses existing data to predict the most likely values for missing data, and is less likely to lead to bias than discarding data sets due to the possibility that the distribution of missing data is non-random. Participants' predicted score and actual total score were compared. The two variables were found to be positively correlated, albeit weakly; r(88) = .25, p = .016. That is to say, predictions had low but not chance-level predictive value.

Participants were also prompted to judge the better item order (after viewing all examples but prior to the criterial test). The question in the task was phrased as follows: "You might have noticed that for some examples you have seen, the skills/command words were alternated, while other examples appeared in a block of five of the same type. On reflection, which order of examples do you think was more helpful to your learning process?" This was followed by a choice of 'mixed examples' and 'examples in a block of the same type'. In response, 41.6% judged interleaving to be superior, while 58.4% judged blocking to be superior.

While the preference for blocking may be seen as a metacognitive error (see Chapter 2), it should be noted that some were correct in that they did go on to do better in this condition on the test phase. Overall, 44 participants did better in the interleaving condition, 32 did better with blocking, and for 14 their scores were equal in both conditions.

Analysing this in terms of predictions, 36 participants made a judgement that was in line with their later performance (e.g. judging interleaving to be superior and then doing better on test items based on their interleaved practice), 38 participants made a prediction that was in the opposite direction to their later performance (e.g. judging interleaving to be superior and then doing better on test items related to their blocked practice); the remainder (n = 16) did equally well on both conditions. It was essentially impossible for the latter group to make a correct prediction, given that equality between the two orders was not presented to them as an option, and this limits the extent to which the accuracy of predictions overall can be analysed. Nevertheless, a chi-square test of independence was performed to examine the relationship between predicted and actual superiority among those participants who scored higher in one condition or the other (i.e. excluding those who scored equally well in both), and this was not significant (X^2 (1, n = 74) = 1.9, p = .168). This suggests that participants are no more likely than chance to correctly judge which learning sequence would lead to better performance.

5.3.4 Interim discussion.

5.3.4.1 Interleaving versus blocking on a skills task.

A key question raised at the outset of this study was whether the factual-learning effects of interleaving found in the literature as a whole can generalise to the learning of higher-order skills as typically presented in school classrooms. The main effect of interleaving found above supports the prediction that the technique would indeed be useful for this kind of material.

Granted, the statistical significance was at the 5% level, and there could be some concern that this was not a strong enough effect to recommend changes to classroom practice. However, it is worth considering that interleaving is not an educational intervention per se. Teachers have to present examples of concepts, and are faced with an inevitable choice over when to do so and in what order. This being so, they must at times present examples in either a blocked or an interleaved order. Given the present finding together with the evidence presented so far in this thesis regarding the benefits of interleaving and its utility for transfer, it seems reasonable to conclude that presenting examples of skills in an interleaved order is preferable to doing so in a blocked fashion.

Again, it is worth recalling the previous research which suggests that item similarity matters, with more of an interleaving benefit for similar/easily confused examples. In the present study, skills which (at least according to the researcher's own teaching and examining experience) tend to be confused by students were contrasted. The benefit might be reduced if different skills were contrasted, or if the learners themselves found the skills clearly distinct and easy to understand.

More broadly, this finding suggests that interleaving can improve comprehension of complex material in the classroom, and can prompt a more active interpretation of examples which helps them to make links and to better compare and contrast different types of skills as they are used in a curriculum-relevant examples.

It could reasonably be pointed out that the task as a whole had relatively little impact on student learning overall; even when examples were interleaved, the performance on the criterial test was not radically higher than the performance on the pre-test; as can be seen from Table

5.2, the pre-test mean percentage was 38.1% compared to the interleaved mean of 5.67, which equates to a percentage of 47.2%[19].

Performance would, however, be expected to deteriorate slightly across the duration of a lesson as attention levels fall, and the final total score might therefore underestimate improvements in understanding[20]. In addition, it would be surprising if a 20-example activity which occupied less than a single school period was sufficient to solve a problem that teachers of the subject highlight as one of their main headaches for the year, and which is frequently cited by the exam board SQA as a major cause for underachievement among students of the subject (for example the exam board report on the 2019 Higher Psychology assessment notes that "Generally, candidates showed little evidence of skills development; particularly the skills of evaluating, analysing and applying"; SQA, 2019, p. 2).

A few percentage points of an increase is not without value, and if a similar task were to be repeated multiple times, one might expect a cumulative benefit. Indeed, most of the studies cited in Chapter 3 report several study phases. Such delayed practice would also fit with the recommendations of research into spacing and retrieval, Rawson and Dunlosky (2011), for example, recommend that items should be practised until students get them correct three times, and then practised a further three times at widely spaced intervals. Such practice and consolidation could have item-level feedback, too – an element which was absent in Study 2 but which might be expected to improve any gains observed.

Multiple practice and feedback sessions are challenging to achieve within a field experiment, especially with exam classes when class time is so constrained. But if a task were developed which proved genuinely useful to teachers, they might be expected to make use of it multiple times across a school year (just as practice exam questions are typically tackled

[19] Although the examples were not randomised across the pre-test and test phases, so it is possible that the pre-test simply had harder items.

[20] It was apparent to the researcher that some learners were getting restless by this stage, and they may have rushed the final test. A further distraction is that some of the other students in the room had finished.

multiple times by students). In addition – and more importantly – if the general procedure of interleaving rather than blocking were adopted by teachers, it could be built into a number of elements of classroom practice, such as the way they present their oral descriptions and lecture slides, or the format of written materials. This again could lead to a cumulative benefit.

But will teachers have enough faith in the technique to do so? Chapter 6 will return to the issue of whether teachers' beliefs about memory are a barrier to the adoption of desirable difficulties. For now, it can be concluded that while developing higher-order skills and the ability to use these skills when faced with novel examples is not an easy matter, the data here suggest that interleaving of initial examples is an appropriate method for developing such skills.

5.3.4.2 Self-explanation on a skills task.

Moving on to the data on self-explanation, two of my research questions related to whether this desirable difficulty would be beneficial – question (b) and whether it would interact with interleaving – question (c). The results supported the hypothesis that self-explanation of key skills would improve performance, a finding which (taken together with prior research on the strategy such as Bisra et al., 2018) suggests that a more prolonged intervention based around self-explanation could have tangible benefits.

However, teachers and students must in part make educational decisions based on practical factors such as the time available to them. When the task time of those participants who wrote complete definitions was taken into account, this added around two and a half minutes on average, or around 10% of task time overall, compared to those who did not – a significant difference. This must be a consideration when teachers are considering a strategy such as self-explanation[21]. In contrast, interleaving examples should not, in principle, occupy any additional class time given that it only involves modifying the order of learning experiences (see Chapter 2).

Relating to research question (c), no interactions between interleaving and self-explanation were found, suggesting that the relative scheduling of these two desirable difficulties, if used, should not be a concern to teachers. However, due to the constraints of the design used, participants only attempted self-explanation at the end of the current practice task and without explicit instruction about what the skills involved. It therefore remains currently unknown whether more direct teaching of the skills might have affected the findings for self-explanation, or led to an interaction. If I am correct to conclude that interleaving promotes meaningful understanding and transfer, then it is possible that it might at some point impact on learners' ability to explain verbally how a skill or concept works, particularly since the findings of Study 1 suggested that participants may seek out meaningful connections before and after the information presented in an interleaved task. On the other hand, numerous studies have shown that learners can profit from desirable difficulties while remaining unaware of their benefit, and their metacognition can reflect a persistence of flawed study habits (see Chapter 6 for a more detailed analysis of this issue). This suggests that interleaving could work on an implicit level, and supports the idea (as per the current findings) that interleaved examples of skills and self-explanation of skills are unrelated.

5.3.4.3 Metacognitive findings.

[21] By virtue of the fact that they were able to provide definition, these participants probably represented an academically stronger subset of the 'define' group, and therefore it wasn't meaningful to compare their scores on the criterial test.

There are also some implications for students' independent learning from the current findings. As can be seen from the metacognitive questions in this study, 41.6% of participants judged interleaving to be the better strategy, while 58.4% judged blocking to be superior, and there was no indication that this judgement predicted later performance. Overall, these findings suggest that there was a lack of insight into which item order was more beneficial, and that their judgements were little more than guesses. In as far as there was a trend towards students favouring a particular item order, it was in support of blocking – the less successful order in the later test.

5.3.4.3 Evaluation of methodology.

There were certain limitations in the methodological approach used in this study, from which lessons can be learned and advice given to future investigators. As noted earlier, examples were presented without direct feedback. Instead, a cue question – "which study is this describing" – was used in an attempt to ensure that participants focused their attention on each example rather than clicking through the task rapidly. However, this may have had unwanted effects. From observation in the room, it was apparent that many participants found it quite difficult to name the studies in question. As such, this cue question may have unhelpfully occupied much of their working memory, reducing the extent to which they could think about the similarities and contrasts between the examples. A simpler version of the cue question – perhaps presenting a forced choice between two researcher names with a simple checkbox – would have been preferable.

Nevertheless, it should be noted that this study is relatively novel in not providing item-level feedback to learners, and nevertheless found an advantage of interleaving over blocking.

This has implications for external validity, and particularly for how easily the findings can be applied to a range of unsupervised classroom tasks.

5.4 General Discussion of Studies 1 and 2

5.4.1 Review of key findings.

The two studies reported in this chapter have explored interleaving in a school context, alongside other desirable difficulties spacing and self-explanation. The present studies are novel in that prior research into interleaving has mainly taken place in laboratories and with undergraduate students, with the exception of a small number of studies of maths practice (e.g. Rohrer et al., 2015) and of concept learning in very young children (Vlach et al., 2008).

Taken together and in the context of the literature reviewed across previous chapters, the two studies in this chapter present a number of issues and considerations for any educator who wishes to apply interleaving and other desirable difficulties to the classroom. Most importantly, they provide some initial evidence that interleaving can be effective as a classroom teaching strategy. Adding to the lab-based research evidence, the present studies demonstrated ways that short texts can be interleaved in the classroom in a way that is accessible to both beginners and to those doing exam-based courses.

Nevertheless, the technique does need to be applied with care and to relevant sets of information. It remains the case that interleaving can only be considered an evidence-based strategy if key aspects of the strategy resemble the evidence. In particular, it appears that interleaving should be used with multiple examples of easily-confused concepts or skills, rather than as a mnemonic device for factual information.

Another major implication of the findings of Studies 1 and 2 is that despite the status of the technique as a desirable difficulty, interleaved presentations do not present much of a risk to performance even over the short-term (given that neither study found a harmful effect). In this it contrasts with the spacing which significantly reduced performance in Study 1. Desirable difficulties which result in subjective difficulties and high error rates could have motivational impacts, and avoiding these problems in the case of interleaving would be a genuine strength of the technique. On the other hand, the metacognitive data showed that the majority of students felt, incorrectly, that blocking was the superior order, and this falls in line with previous studies (e.g. Eglington & Kang, 2017; Kornell & Bjork, 2008). It suggests that although interleaving did not harm performance, the task was subjectively more challenging perhaps because of the cognitive effort involved in paying attention to differences.

Compared to other interventions, interleaving is also a low-cost strategy, and one that is entirely within the control of the practitioner. Any teacher, anywhere in the world can walk into their classroom tomorrow and present examples to their class in either an interleaved or a blocked fashion. Unlike some other interventions it does not require any complex or expensive external input or expertise, and it does not require more time, either.

More broadly, I have explored the way that three desirable difficulties are applicable to classroom practice. It is clear that each comes with its own benefits and drawbacks. Spacing is easy to plan and schedule (it is largely a matter of timing and planning) and the delays between one practice session and the next can be used for other tasks. However, delays make practice subjectively harder, harming short-term performance, which has potential to affect motivation. Self-explanation is also straightforward to implement and can lead to improvements in performance, but has a cost in the time required which may affect its efficiency as a classroom choice. Interleaving requires careful thought in terms of exactly which items should be interleaved and which should not.

5.4.2 Reflections on methodology.

The methodology used in studies 1 and 2 – a computer-based task – has direct implications for online learning. The whole concept of learning as being inherently classroom-based is arguably changing, with a more flexible, partially home- or library-based model emerging. Indeed, many school pupils learn entirely online (a trend which is growing rapidly; Levinson, 2015), while some conventional classrooms have adopted a 'flipped' model whereby traditional homework-type tasks are done in class with the aid of the teacher, and exposition-type tasks (e.g. watching a lecture or video) are done at home (Lo & Hew, 2017). Online access to schools and universities alongside traditional formats is also becoming standard. In this context, it is important to understand and support the choices that learners make during autonomous modes of study (Rice & Carter, 2016), and the present task provides one way of doing so.

One option for future research would be to build in a degree of choice into the task, so that rather than making judgements about which order is best, students could actively choose how and when (and perhaps, how much) to study examples of concepts and skills.

A difficulty with more optionality within the experimental task is that it would reduce the size of any particular condition of study, leading to a lack of statistical power. An option to address this would be to establish an online study aid or entire course with the specific purpose of tackling the research questions raised in this chapter. Such a course could be offered free to students or members of the public who were happy for their learning progress to be analysed, and potentially result in large data sets. This would allow course designers to try multiple variations of the order in which items were presented, and judge the outcome via a series of brief progress tests.

A similar approach to this was trialled by Feddern et al (2018) in a randomised controlled trial of over 1000 GCSE pupils, although this was not the principle aim of establishing the courses (which were primarily a study aid). Their software applied four key variables (retrieval, interleaving, spacing and visuals):

> "Interleaving and spacing are achieved with an algorithm that monitors each student's performance on each piece of content and presents new question modules in accordance with the pieces in need of further practice. Each piece of information is tested in one or more question formats to promote retrieval practice. The use of visual cues is implemented through the presentation of the same images during both the introduction and testing modes."

A possible follow-up to this approach could be to present examples of skills based around study habits. The metacognitive benefit would be to teach students how to learn, but such a course could, again, be used as a research tool.

Further research could also address some of the limitations of the computer protocol used in the two studies in this chapter. While there were improvements from Study 1 to Study 2, there were still a number of questions that were left unanswered. In particular, the use of immediate feedback on performance is worth exploring. While this is often absent in teaching situations (e.g. lectures, reading), the online presentation would make it possible to give participants feedback on their response to each task. Presumably this would improve performance, and it would be useful to determine whether such an approach would be a viable teaching option.

Linked to this, a further flaw was the lack of a follow-up after a delay. Given that learning can only truly be assessed after a delay, the current studies measured performance rather than learning. There is a research basis to suggest that interleaving does not interact strongly with a time delay (see Chapter 3), and here, the short-term criterial test in Study 2 revealed an improvement among the interleaved skills learning condition. Nevertheless, concept learning can be a slow process (see Chapter 1) as can skills development, and the improvements shown over the task in Study 2 were modest. As Soderstrom & Bjork, 2015, p. 193) put it, “we recommend that experimenters include both short- and long-term measures in their studies”, and this is particularly important for research that aims to look at the application to attainment on real academic courses.

A follow-up study which tracked students over several intervals – once per month, say – would add useful information about the sustainability of interleaved examples as a means of skills learning, and provide an insight into how the technique might interact with the general progression of learning over the duration of a one-year academic course.

5.4.2 Implications for future research.

The current findings could be extended by replicating the procedures with school subjects other than psychology. In terms of the scope of items that are amenable to interleaving, the systematic review (Chapter 3) showed that the use of the technique with meaningful verbal examples and individual images from either art-based or science-based subjects is well established. It would seem reasonable to suggest that interleaving could extend to a range of other scenarios – images used medical and dental education, for example, or examples of logical arguments used in philosophy, categories in social sciences, or geographical images. Beyond the scope of the systematic review, interleaving has also been applied to music and to sports.

While the main focus of the present chapter has been on interleaving, I have also found some initial evidence which suggests that interleaving and spacing may interact, even if they are applied to different phases of a learning task. It was already established (Birnbaum, 2013; Birnbaum et al., 2013) that adding delays between examples can be disadvantageous when the delay interferes with discriminative contrast. The current evidence from Study 1 suggests that the interactions between the two techniques can be broader.

This raises a methodological point; again, most of the research literature into interleaving is lab-based, and spacing and interleaving are most often studied in isolation. There is a lack of information about how use of the techniques may pertain to longer study periods such as an entire school day, and the present studies provide some intriguing initial evidence suggesting that interleaving could have beneficial effects in this regard. Finding out more about any costs or benefits when they are applied to longer periods is an important priority for further applied study.

5.4.4 Implications for practice.

The findings on self-explanation in Study 2 suggest that this technique can be effective in the classroom, but they come at a cost in terms of learning time. Both the effect and the cost of the technique have practical implications. However, as the present research did not compare it to other techniques that might be tried during the same time period (for example reading about the skills), I will not, in this thesis, give any further consideration to whether this cost may be worth paying. Suffice to say that under the right conditions, there is evidence that engaging in self-explanation of skills is more beneficial than not doing so.

It is apparent from Studies 1 and 2 that interleaving does depend on the choice of material to be interleaved – simply reordering items will not be helpful in all cases. As has been found in the prior literature, for example, there is no interleaving benefit from interleaving unrelated facts with vocabulary items (Hausman & Kornell, 2014).

The data presented by Carvalho and Goldstone (2015b) present an interesting dilemma in this regard. According to their analysis, items to be interleaved need to be highly similar; internally diverse categories fail to promote the same cognitive processes as learner attention is instead directed towards cross-category similarity. This consideration might imply a sliding scale from identical to entirely unrelated items, with a point upon that scale at which interleaving ceases to be helpful and blocking becomes preferable. The reason for this is that it may start to become more advantageous to notice similarities than it does to notice differences (Carvalho & Goldstone, 2015b; MacKendrick, 2015). A principle then, might be to say the following: similar and easily-confused examples should be interleaved to help learners notice, while examples where it is hard to discern important commonalities should be blocked.

However, in the classroom, difficulty and similarity are not always clear cut. Instead, they depend on the knowledge level of the learners themselves. Beginners may find apparently obvious commonalities hard to notice because they lack the background schema knowledge to make the relevant connections.

Given this fact, teachers are probably in the best position to determine what target items are commonly confused, and to therefore determine when they should or should not interleave examples. As such, teachers may find themselves in the unique position of knowing both the learning situation (comprising of the material and their own students) and the technical craft involved in applying teaching techniques. However, to do so optimally, they may require a fairly sophisticated understanding of human memory. A knowledge of memory and of learning techniques could therefore be seen as part of professional competence, underlying successful classroom decision making. My own article (Firth, 2017, reproduced in Appendix 10) makes this case directly, and the present evidence (in particular, Study 1) shows that there are risks involved in applying desirable difficulties without following some of the key principles that arise from the research literature.

The research in this chapter strongly suggests that the scope of interleaving can be extended to cognitive skills as well. The particular focus in Study 2 was on the higher-order thinking skills that feature in many school and university courses. Not only are such skills integral to success in many courses, they are also typically awarded credit that cannot be achieved by demonstrating factual knowledge alone. In some cases, especially at higher levels of academic study, it is impossible to achieve top grades without demonstrating a proficiency at skills such as analysis and evaluation. It is therefore important that students are able to distinguish between examples of these skills.

While the exact nature of what these skills will look like varies by subject – and there are plausible doubts that skills can fully develop in the absence of domain-relevant expert knowledge (Baer, 2016; Bransford et al., 2000) – higher-order skills are an important and under-researched target for classroom application of the desirable difficulties framework (Agarwal, 2019). A general principle established here is that interleaving rather than blocking of examples has the potential to boost student learning of these skills.

An area that links closely to higher-order skills is students' metacognitive awareness of effective study habits (see 5.3.1). The idea that students tend to find spacing and interleaving subjectively harder has implications for independent learning, given that harder tasks may be avoided. Perhaps more concretely, the findings which suggest that skills development can be improved via interleaved examples (Study 2) have potential to be applied to metacognitive skills. The idea of metacognitive competence in learning – of learning *how to learn* – has gained increasing attention in recent years, and many. A useful area to extend and apply the current findings would be to interleave metacognitive examples in order to promote better understanding of learning processes (including desirable difficulties) among the students themselves. The role of beliefs in learning is discussed further in Chapter 6, and I will return to the practical applications of the findings from this thesis in Chapter 7.

5.4.5 Implications for professional learning.

Finally, it is important to consider how teachers would realistically use and employ interleaving. While it has been argued that the present studies better reflect classroom practice than does the research literature as a whole, there is a shortage of empirical evidence about how teachers actually present examples in the classroom. It seems at least possible that examples are presented quite infrequently. There is also some evidence that students often misjudge examples (Zamary et al., 2016), and it is even possible that some of the same errors apply to examples that teachers spontaneously generate during a class. Finding more about how teachers use examples when explaining new concepts as well as their approach to testing students on examples would help to guide the application of interleaving to real classrooms.

If some training is required in order to help teachers use these techniques optimally, this need could itself be amenable to the use of desirable difficulties; as noted in Chapter 2, if interleaving and spacing are effective for students, there is no reason that they should not be effective for teacher professional learning, too. Professional learning could be spaced and interleaved – and perhaps it already is. Three first steps to investigating this could include:

- Investigating current professional learning opportunities such as short courses and practitioner enquiry projects to assess the extent to which learning is spaced and interleaved or prompts.
- Running experimental tasks to test whether the current findings generalise to skills that have a direct applicability to teaching, such as judging/grading student work.

- A metacognitive investigation into how teachers judge the value of techniques such as spacing and interleaving, on the basis that underestimating their utility or dismissing them could lead to suboptimal classroom practice.

It is to the latter that I now turn. There is increased recognition across the education sector that a deep, evidence-based understanding of human learning and memory processes is a part of the modern teacher's professional toolkit. In England, the Chartered College for Teaching runs a 15-month Chartered Teacher programme the aims of raising the status of the profession, developing teacher expertise, and teacher engagement with research evidence (Chartered College of Teaching, 2020). Meanwhile in Scotland, a widespread and growing emphasis on practitioner enquiry (e.g. Wall, 2018; Wall et al., 2018) could provide a professional learning environment for teachers to semi-autonomously explore the use of desirable difficulties in their own classroom. To better understand the context for such efforts, Studies 3 and 4 investigate teachers' beliefs about memory, and about spacing and interleaving in particular.

6

Teacher Beliefs About Memory

6.1 General Introduction

6.1.1 Role of desirable difficulties in learning.

What can be drawn from the findings presented and discussed so far is that interleaving has a powerful potential as a learning strategy, especially when combined with the broader evidence base around desirable difficulties, and with some cognisance of key factors such as the types of tasks to which it best applies.

More broadly, learning (as defined in Chapter 1) can be boosted by the use of several well-established strategies relating to the order and manner in which learning takes place. As well as interleaving, these include spacing (incorporating a delay before restudying leads to more durable learning compared to more immediate restudy) and retrieval practice (tests and recall activities are more effective than passive listening or re-reading. Dunlosky and Rawson (2015) provide a useful and accessible summary of these issues as they apply to education in their ‘teacher-ready review’, noting that the evidence for spacing and retrieval practice is ‘strong’, and the evidence for interleaved practice is ‘moderate’ (and several of the studies of interleaving reported in Chapter 4 were conducted subsequent to their 2015 paper, adding to the evidence base for the technique which they cite[22]).

However, as *desirable difficulties*, all three of these techniques stand to boost long-term learning but can present significant challenges to shorter-term performance. As discussed in

[22] It’s also worth noting that Dunlosky & Rawson primarily refer to interleaving as a revision/practice technique rather than one that can be used for presenting new material. This may be in part due to the influence of early studies of mathematics by Rohrer and colleagues (e.g. Rohrer & Taylor, 2007).

the review by Soderstrom and Bjork (2015), performance and learning are often negatively correlated (an immediate practice session, for example, would be quite easy for a student, but not the most effective way to review and consolidate). A practical implication is that the benefits of such strategies may not be obvious to learners during learning, or for some time afterwards. As found in Study 1 (see Chapter 5), learners may actually do *worse* when material is spaced – such interventions are, after all, difficulties.

When students engage in independent learning activities such as quiet classroom study or revision outside of the classroom, they have to make a series of choices about what to study and for how long. The *region of proximal learning (RPL) model* presented by Kornell and Metcalfe (2006) suggests that instead of working towards a particular goal (such as mastery of a sub-set of the course content), students tend to persist in studying target material until they feel they are no longer making progress. If they feel they are making gains, they will continue. If progress slows, they will switch to different target material. This model provides the important insight that students do not necessarily focus on what they most need to learn or on areas where they are weakest, but instead spend their available time on material where progress feels fastest. If they feel that the item is so hard that they are learning little or nothing, then they will stop or switch.

Taken together, these two ideas – desirable difficulties and the RPL model – allow us to predict that students will make some very flawed study decisions (flawed in the sense of being suboptimal when judged by the long-term outcome of their study behaviour). And indeed, examples of this are evident from the literature on student study strategies. In laboratory studies, metacognitive judgements have demonstrated a widespread belief that blocking is preferable to interleaving (e.g. Eglington & Kang, 2017; Kornell & Bjork, 2008; Yan, Bjork, & Bjork, 2016; and see also Study 2) and that massing is preferable to spacing (e.g. Zechmeister

& Shaughnessy, 1980). Learners do not appear to be able to judge for themselves which techniques are more effective, even when given a chance to try them out.

In authentic settings things appear to be very similar. A survey of 472 American psychology undergraduates by Kornell and Bjork (2007) found that 59% selected their target study material on the basis of "whatever is due soonest/overdue", while 86% responded 'no' to the question "Do you usually return to course material to review it after a course has ended?" (p. 223)[23]. A replication and extension by Hartwig and Dunlosky (2011) with 324 participants found very similar percentages – 56% and 78% respectively – and also revealed that flawed study behaviour correlated with poorer grade point averages (GPAs). What's more, such GPAs typically only reflect scores on modules lasting a few weeks or months; if learning over a timescale of years had been measured, the benefits of desirable difficulties may have been greater still (Bahrick, 2005, notes that in studies of real-world learning situations, forgetting does not plateau until 3–5 years after initial learning).

Both of the surveys mentioned in the previous paragraph were conducted on undergraduates. As such, the student participants had already engaged in a lot of independent study; even 1st year undergraduates are closer to the end of their years of formal education than they are to the beginning. Together with the experimental evidence, this strongly suggests that experience alone is insufficient for learners to begin to understand their own memory processes, and thereby to appreciate the benefits that strategies such as interleaving and spacing could bring to them.

If learners are unable to fathom the depths of human memory for themselves through the medium of experience, it would appear to fall to teachers to guide the learning process. However, for this to be achieved, teachers themselves need an accurate understanding of human

[23] A corollary of these findings raised by Carpenter et al. (2020) is that students' teaching evaluations may be flawed; if they are unable to recognise effective learning when they experience it, they may give higher ratings to instructors who make classroom experiences easier but less effective.

memory processes. They need to know how to present and explain new concepts in ways that will lead to lasting representations in memory. They need to model spaced and interleaved practice, retrieval, and other effective techniques. And they need to provide some kind of instruction – whether implicit or explicit, oral or written – about how learners should approach classroom tasks.

A major question to address in this chapter, then, is the following: how well equipped are teachers to do so?

6.1.2 Teachers' beliefs about memory.

Flawed study decisions and habits such as cramming the evening before a test are to an extent inevitable, as such behaviour depends not just on learners' judgements of how memory works but also on practical matters such as the amount of time available to study and the role of deadlines. Nevertheless, teachers may wish to guide their learners closer to optimal decisions by equipping them with the skills and knowledge required to engage in successful learning strategies. As such, teachers can act as metacognitive role models (Wall & Hall, 2016), either by demonstrating good learning habits during classroom activities or by explicitly explaining them (as recommended by several authors, e.g. see Pintrich, 2002; EEF, 2018).

Indeed, it would surely be preferable to teach learners *how to learn* in general – so that they can then make good decisions autonomously – rather than having to instruct them on how to proceed with each specific task (perhaps without ever developing a metacognitive understanding of what they were doing and why).

However as raised in Chapter 1, memory is complex and its workings are not intuitively obvious (Bjork, 2011). Misconceptions about memory are rife; surveys by Simons and Chabris (2011, 2012) found numerous deviations from the scientific consensus among members of the general public (see also Chapters 2 and 4). For example, 82.7% of their sample strongly agreed or mostly agreed with the statement “people suffering from amnesia typically cannot recall their own name or identity” (Simons & Chabris, 2011, p. 3), but 0% of memory researchers surveyed did so.

Flawed understandings of memory can be found among professionals too, perhaps most notably in legal settings. In a study that focused on the recollections of eyewitnesses to a crime, Melinder and Magnussen (2015) found that professional psychologists and psychiatrists serving as expert witnesses in court often endorse memory myths, following up on an earlier study which found that judges make similar mistakes (Magnussen et al., 2008; both studies were conducted in Norway). Neither sample were any better than the general public overall in terms of their understanding of human memory, although there were some differences in the pattern of results, with (for example), a higher proportion of psychologists correctly opining that a witness’s confidence is a poor indicator of the accuracy of their recollections. However, psychologists scored lower than the other samples on some items, including the idea of the forgetting curve.

It is perhaps surprising that psychologists didn’t score better given that they must have been exposed to accurate information about these concepts at some point during their prior studies or training. One possible explanation for this phenomenon comes from an experiment on American undergraduates by Will et al. (2019) which looked at techniques for tackling misconceptions. Their task items included culturally-widespread misconceptions such as the beliefs that ostriches bury their heads in sand and that Napoleon was unusually short. A key finding of the study was that *directly refuting* flawed beliefs is more effective than simply providing contrary evidence. That is to say, it is important to state clearly that “x is wrong, and here is why”. Direct refutations led to more accuracy and reduced circular reasoning in texts written later by the experimental participants; simple exposure to the correct information (as the psychologists may have experienced during their careers) had less of an effect.

Melinder and Magnussen (2015) did find that psychologists associated with academic institutions rather than private practice did slightly better on some items such as the myth of recovered memories, and female participants and younger participants scored better overall. But on the whole, these results suggest a very restricted benefit of professional knowledge and experience when it comes to an understanding of memory.

Education level also appears to have rather limited benefits in this area. A survey by Furnham (2018) presented adults with the items listed by Lilienfeld et al. (2010) in their book '*50 great myths of popular psychology',* and found that found participants' education level did not correlate with accurate identification of myths, and that background study specifically of psychology was also unrelated. However, Furnham acknowledged that his work did not delve especially deeply into the level and manner with which participants had previously studied psychology; the questions also did not focus just on memory. McCabe (2011) also didn't find a difference between psychology majors and non-psychology majors in her study which looked just at learning strategies. Ost et al. (2017), using questions more akin to those used by Magnussen and colleagues, found that psychology undergraduate students in the UK did somewhat better than previous samples, suggesting that a deeper theoretical understanding of human cognition can lead to more accurate memory beliefs. The students scored better than UK hypnotherapists in the same study, but worse than chartered clinical psychologists (interestingly, the hypnotherapists had the highest self-reported knowledge of the memory literature). Graduates scored better than non-graduates in the Simons and Chabris survey, too, suggesting that education more broadly could play a role in attenuating misconceptions about memory.

In as far as the findings above can be generalised to the teaching profession, they paint a picture which suggests that misconceptions about learning and memory are not self-correcting simply through exposure to experience or to contrary factual information. There is at least some evidence that it will be beneficial to have been educated to a high level (as nearly all teachers have been), but psychology education and training will only help to combat myths if it is quite specialised – perhaps beyond what teachers are likely to have experienced during their preparation for the role. However, much of the research discussed so far has focused on the legal setting, and – perhaps surprisingly – evidence relating specifically to the teaching profession is harder to come by.

One closely-related area that is very well evidenced is that many teachers are prone to myths about learning. *Neuromyths* are popular ideas about learning, memory and the brain which are not in line with the consensus among researchers. They include the concept of 'learning styles' – the notion that each learner has a specific sensory modality in which they learn best, for example visual or auditory learning (a common formulation is 'VARK' – visual, auditory, reading/writing and kinaesthetic 'styles'). Despite significant doubts that all learners can be divided among such simplistic and absolute categories, this idea has been endorsed by 74–97% of teacher participants across numerous international surveys of the profession (Betts et al., 2019; Howard-Jones, 2014; Morehead et al., 2016). Despite a lack of evidence that learners can be so simplistically categorised, teachers and learners alike may choose to refer to themselves as 'visual learners', and so forth. Reviews (e.g. Kirschner, 2018; Pashler et al., 2008) and field research (e.g. Husmann & O'Loughlin, 2018) have demonstrated that allocating particular individuals to particular modes of study does not improve attainment, and it is widely agreed among psychologists that attempting to cater pedagogically for learning styles is therefore at best pointless, and at worst counterproductive. There is also evidence that teachers and their students do not agree on which learning style each student should be allocated to (Papadatou-Pastou et al., 2018).

In a study by Morehead et al. (2016), 91% of university teachers endorsed the 'learning styles' myth. More positively, many of the same sample did endorse some more effective strategies such as self-testing (a strategy which affords an opportunity for retrieval practice). However, their reasoning for endorsing it did not indicate an understanding of how new memories are formed. And, pertinently to the present research, Morehead and colleagues also found that students' misconceptions appeared to link to those of staff. That is to say, the students may have been gaining flawed information about how to learn from their teachers.

In their review of performance and learning, Soderstrom and Bjork (2015) state several times that teachers are likely to misunderstand memory, but they do not report empirical evidence of this; the evidence they cite is instead focused on *student* behaviour. While I agree with the principle of statements in their paper such as the following...

> "research in metacognition suggests that fleeting gains during acquisition are likely to fool instructors and students into thinking that permanent learning has taken place, creating powerful illusions of competence" Soderstrom and Bjork (2015, p. 193)

...it is important that any lack of metacognitive understanding of memory processes is established empirically, and in a sample of teachers specifically. Despite the many studies conducted on neuromyths, studies of this issue are very rare in the literature (although both Morehead et al., 2016, and Betts et al., 2019, included some questions on memory-related issues).

The purpose of the studies in this chapter is therefore to better establish the level of accuracy of teachers' beliefs about memory, and, if misconceptions are uncovered, to analyse the implications of those misconceptions.

To this end, I now describe two further studies. In Study 3, a broad-ranging survey with short statements on the functioning of human memory was circulated to teachers from various sectors. In Study 4, a more focused survey with extended vignettes about three desirable difficulties – interleaving, spacing, and retrieval practice – was circulated to three groups: student teachers, practising teachers, and retired teachers.

6.2 Study 3 – Teacher Beliefs about Memory

6.2.1 Introduction to Study 3.

As noted in the introduction to this chapter, teachers' beliefs about memory are a particularly under-researched area. Previous studies into beliefs and misconceptions about memory and learning have tended to be either very broad-ranging studies into psychological myths (e.g. Furnham, 2019), studies pertaining to eyewitness memory (e.g. Magnussen & Melinder, 2015), or studies of memory which sample the general public (e.g. Simons & Chabris, 2011). Those studies that have been done into beliefs about learning and memory in educational contexts have either sampled students (e.g. Kornell & Bjork, 2007) or have focused on learning myths (e.g. Morehead et al., 2016).

As such, there is not a clear scientific consensus about the level of teachers' professional knowledge in this area. While it goes without saying that both teachers and the systems within which they work will vary, it would be very useful to better understand whether the misconceptions about memory that have been uncovered in other professions and among students will also be found in the teaching profession.

The present survey therefore aimed to establish whether teachers' beliefs about human memory are subject to similar misconceptions as have been found among other populations. It took a broad-ranging approach, including questions not just on the desirable difficulties discussed throughout this thesis such as interleaving, spacing, and retrieval practice, but also questions on the nature of long-term memory in general, with some items drawn directly from previous research. Accuracy of responses was operationalised in terms of concordance with the scientific consensus, based upon a reading of the evidence as reported in previous chapters of this thesis and, in some cases, norms established in prior research into misconceptions.

The study investigated three main questions, and made the following predictions:

(a) Are teachers' beliefs about memory accurate?

In the absence of evidence to the contrary, the parsimonious prediction is that teachers' responses to questions about memory will, like those of the general public or other professionals, show signs of being out of line with the scientific consensus. I therefore predicted that participant responses would be out of line with the scientific consensus.

(b) What role does years of classroom experience have on accuracy of beliefs?

On the basis of memory being fundamentally counterintuitive in its functioning (Bjork, 2011), as well as the findings reported above which showed that learning myths do not correlate with years of experience in the classroom and that age and experience have not played a major role in the accuracy of beliefs among other populations, it can be hypothesised that any misconceptions which are held early in a teacher's career will not be self-correcting. I therefore predicted that there would be no correlation between years of classroom experience and the level of accuracy of participant responses.

(c) Is a psychology background related to more accurate memory beliefs?

The study also aims to look at participants' teaching sectors and subjects (primary, secondary English, secondary Biology, higher education, and so forth), to investigate whether background psychological knowledge would affect accuracy of responses. Past research (see section 6.1) is mixed on this issue, but if a professional knowledge of psychology is beneficial in reducing misconceptions, this would have important professional learning implications. It can also be assumed that people who presently teach psychology have more of a working familiarity with topics such as memory than do people who studied it many years in the past, as may have been the case with participants in the Furnham (2019) study, and may therefore have a deeper understanding of memory processes in line with the participants in the Ost et al. (2017) study. I therefore predicted that participants who teach psychology would give more accurate responses than those who do not.

6.2.2 Method.

6.2.2.1 Design and materials.

A survey was undertaken, again using the software provided on the PsyToolkit website (see Stoet 2010, 2017). The instrument used featured 24 statements relating to memory, each with a 5-point Likert scale labelled "strongly agree" to "strongly disagree". The first set of questions related to memory in general, with two drawn directly from Simons and Chabris (2011). The remaining questions focused on aspects of memory relevant to teaching and on which there is broad scientific agreement, for example the spacing effect, retrieval practice, and interleaving (see Appendix 12 for a full list of questions as well as notional correct responses and summarised responses from the sample).

Granted, the scientific consensus is open to change, and not every researcher would necessarily agree that the option selected in the current study is the correct one. However, a great effort was made to ensure that was indeed an accurate representation of what is generally accepted by mainstream cognitive psychology[24]. Indeed, most aspects of the psychology of memory as it applies to education are less controversial than those issues presented in studies of eyewitness testimony. In that sub-field, matters such as false memories, memory under hypnosis, repressed childhood memory and the emotional state of a victim were subject to a prolonged scientific debate so fierce that it is often now referred to as the 'memory wars' (see Loftus, 2019).

The source of the items used included a small number of statements used in the Simons and Chabris (2011) survey, which I will henceforth refer to as 'general memory questions'. These particular items were selected from among those used by Simons and Chabris because they appeared more relevant to teaching than did the other statements in that research.

The other items used were newly devised for the present study. The broad issues covered included desirable difficulties, the functioning of LTM (e.g. forgetting curve, connection of memories into meaningful schemas) and metacognitive beliefs about memory; the supporting literature to the matters included has already been outlined in this thesis, and Appendix 12 also summarises the issues included and refers to relevant sources.

[24] Granted, this mainstream view is somewhat out of line – or out of *phase* – with the mainstream of education. For example, the idea that new memories form mainly be repetition in the short-term store (as presented by Atkinson & Shiffrin, 1968) is widespread in education but is no longer commonly accepted by cognitive psychologists.

6.2.2.2 Sample and procedure.

79 participants were recruited from a Scottish secondary school (n = 45) or via weblink shared on Twitter using the author's account, together with a tweet inviting teachers to take part. Six participants dropped out before completing the survey. Listwise deletion was used. While this method can be seen as inferior to the imputation methods of data replacement (Graham, 2009) used in other studies in this thesis, these concerns largely apply to longitudinal studies, in which there is a significant likelihood of bias due to theoretically important and non-random reasons for participants dropping out of a study. Here, the most likely cause of attrition was participants being unwilling to answer long lists of questions or being put off by a lack of optionality – phenomena commonly observed in both paper-based and internet surveys (Andrews et al, 2003). In addition, the number of cases missing was 7.6%, only slightly above the 5% for which Graham (2009) describes the negative effects of listwise deletion as 'inconsequential' (p. 554). It should be noted, however, that there remains a slight loss of power and increased risk of bias; these limitations will be discussed below (see 6.2.4).

In terms of the cases prepared for analysis (n = 73), 41 participants were recruited from the secondary school and 32 via Twitter. Of these, 51 participants were from the secondary sector, and the remainder from primary (n = 8), FE (n = 5), HE (n = 8) or "other" (n = 1).

The online survey featured an ethics statement and a consent form. At the end of the task, participants were provided with a unique code and contact details, allowing them to withdraw later if they chose to do so.

6.2.3 Results.

Prior to analysis I reverse-scored each item where disagreement was the 'correct' answer (i.e. the answer best in line with the research literature), such that '5' always represented an optimal response.

Following the practice used by Simons and Chabris (2011; 2012), I combined 'weakly agree' and 'strongly agree' responses and the equivalent for 'disagree' responses. The general memory questions revealed misconceptions about memory, but at lower levels than had been found in prior research (see Table 6.1).

Table 6.1: Percentages of participants *out of line* *with the scientific consensus, combining 'weakly' and 'strongly' responses.*

	Present Study	**Simons/Chabris (2012) online survey**	**Simons/Chabris (2011) survey**
"People suffering from amnesia typically cannot recall their own name or identity."	17.5%	81.4%	82.7%
"Once you have experienced an event and formed a memory of it, that memory does not change."	11.3%	28%	47.6%
Other general memory questions (mean)	20.05%	48.2%	72.5%

The remainder of this section will focus on how often participants were *concordant* with the scientific consensus on aspects of memory relevant to teaching, combining correct weakly agree/strongly agree responses and the equivalent for ‘disagree’ responses. These will be reported as percentages of the mean from the Likert scale after reverse-scoring responses such that a high score on the question always reflected the notional ‘correct’ answer. This mean score thus takes into account all responses; 5 would represent maximum accuracy (every participant strongly agreeing with the scientific consensus). The data can also be represented in terms of the percentage of participants who agreed or strongly (dis)agreed in line with the scientific consensus and therefore gave a response that can be considered ‘correct’. These values are shown in Table 6.2, below.

Table 6.2: Table of items used and findings from Study 3.

ITEM NUMBER AND TEXT	% of Ps correct*	Mean from Likert scale (5= maximum)
1. People suffering from amnesia typically cannot recall their own name or identity.	51.9	3.49
2. Sometimes people who have committed murder have no memory for the crime because they have repressed the memory.	24.1	2.72
3. Very high stress during an event can harm a person’s ability to remember the event accurately at a later date.	93.7	4.42
4. When small children describe events they have experienced, their accounts are usually more accurate than those of adults.	31.6	3.11
5. A person’s perception and memory for an event may be affected by his or her attitudes and expectations.	94.9	4.58
6. Once you have experienced an event and formed a memory of it, that memory does not change.	81.0	4.2
7. Improvements in learning always require spending more time studying.	54.5	3.32
8. Most learners have a good idea of how practice/study will impact on their memory.	70.1	3.79

9. When reviewing a topic, it's best to give learners open questions rather than multiple-choice questions or verbal summaries.	37.7	3.01
10. The best way to learn something is to go over it repeatedly within the same hour.	84.4	4.17
11. Learners benefit from mixing up lots of different types of problems, rather than doing one type of task at a time.	63.6	3.73
12. The majority of information taught during a class will still be retained by learners 2–3 weeks later.	77.9	4.065
13. Learners are in the best position to judge what and how they should study.	64.9	3.58
14. It's always best to simplify things for learners in some way, because making something easier helps it to be processed into long-term memory.	44.6	2.96
15. As a teacher, it is wise to wait until learners have almost forgotten things before you go over them again.	12.2	1.87
16. Multiple re-readings are more useful for learning than doing lots of tests.	55.4	3.45
17. Good study advice for learners should include telling them to find a place where they are comfortable and to do all their revision there.	32.4	2.75
18. A learner's current performance on a task is not a reliable guide to their long-term learning.	67.6	3.605
19. One of the best ways to remember something over the long term is to focus on its meaning and how it links to other things.	97.3	4.395
20. If a learner guesses and is not correct they may remember the wrong answer, so it's best to avoid guessing/predictions during lessons.	75.3	3.84
21. Including extra information or examples in a written passage makes it harder for learners to remember the main points.	17.8	2.34
22. Ultimately, learners form new memories through frequent repetition.	26.0	2.50
23. It makes sense to do a homework task soon after the material is done in class.	27.4	2.41
24. Once learners have got a question wrong and then been corrected, they will be able to predict whether they will get it right in future.	41.1	3.09

As can be seen from Table 6.2, the highest level of accuracy was found in responses to items 5 and 19, both of which related to the role of meaning in learning, with 94.9–97.3%

agreement among participants respectively (and mean scores of 4.58 and 4.395 respectively). Items 8, 12 and 20 (on student metacognition, forgetting and the role of inaccurate guessing, respectively) also prompted over 70% correct responses across the sample.

Questions to do with the spacing effect were associated with more inaccurate responses; in particular, most did not agree that “it is wise to wait until learners have almost forgotten things before you go over them again” (item 15; 12.2% agreement), a principle derived from Bjork (2011). However, there was agreement that repetition within a single hour (i.e. cramming; item 10) was a poor strategy (84.4%) agreement.

Another area where beliefs differed from the scientific consensus was retrieval practice. Only 37.7% were correct on item 9 – the benefits of open questions that require retrieval – while only a slight majority correctly rejected the idea that “multiple re-readings are more useful for learning than doing lots of tests”.

I next calculated an overall score for each participant, again based on their responses to individual items ranked from 5 (closest to correct) to 1 (furthest from correct); the maximum score was 120 (5/5 on 24 Qs). Years of experience were self-reported on the survey in bands of five years (0–5 years; 6–10; 11–15; 16–20; over 20). The normality of the distribution was analysed; both as a whole group and divided by experience level, skewness and kurtosis showed z-scores within an acceptable range. The Shapiro-Wilk test likewise suggested a distribution that did not differ significantly from normality, as did visual inspection of histograms and Q-Q box plots of the data (Ghasemi, & Zahediasl, 2012). They were therefore accepted as suitable for parametric analysis. The mean scores by years of experience are presented in Table 6.3, below:

Table 6.3: Participant scores on memory survey stratified by years of experience.

Years of experience	n	MEAN	SD
0–5	14	82.36	8.48
6–10	20	86.15	10.58
11–15	12	79.25	9.73
16–20	6	80.33	7.61
over 20	21	83.93	8.16
Total	73	82.93	9.24

The homogeneity of variance between the different subgroups was non-significant (p[4, 68] = .685) indicating that the variance was comparable. To compare these groups, a one-way between-subjects ANOVA was conducted (F[4,68] = 1.233, p = .305). This did not indicate a significant difference; it was therefore concluded that there was no difference in accuracy on the memory questions depending on teachers' experience level.

Finally, teachers with psychology as one of their main subjects scored higher on average, with a mean score of 91.0 (SD = 11.3, n = 13) compared to 77.74 (SD = 8.94, n = 61). A two-tailed independent means t-test was carried out, and it was found that this difference was significant (p < .001), supporting prediction (c).

6.2.4 Interim discussion.

6.2.4.1 Overall accuracy of teacher beliefs about memory.

An accurate understanding of human memory is likely to affect successful planning and teaching. The findings from Study 3 suggest that teachers hold a better understanding of such issues than other professionals or the general public, perhaps because initial teacher education features input on learning theories, or because of ongoing professional learning, or an identity-based motivation to engage with and learn about such issues. Perhaps most notably, the sample were less likely than participants of prior surveys to agree with the statement, "Once you have experienced an event and formed a memory of it, that memory does not change". This suggests that teachers have an understanding that new memories are moulded gradually through a range of experiences, and that they are aware of how easily their students can misremember events. Overall, then, the data conflict with prediction (a), above.

Among the new items used in this survey, some other areas also showed notably accurate responses, and although these cannot be compared to the general public, they compare favourably to the research consensus. The highest level of accuracy in terms of number of responses (rather than mean response) was found in response to a statement about meaning: "*One of the best ways to remember something over the long-term is to focus on its meaning and how it links to other things*". 97.3% of the sample agreed. Items relating to metacognition, forgetting and the role of inaccurate guessing also elicited a high rate of correct responses.

However, significant inaccuracies were also present, which showed a lack of understanding of how desirable difficulties such as the spacing effect and retrieval practice might apply to classroom practice. The statement *"As a teacher, it is wise to wait until learners have almost forgotten things before you go over them again"* fits with the spacing effect, but only 12.2%

agreed/strongly agreed, the lowest for any question (see Table 6.2). Similarly, just 27.4% disagreed/strongly disagreed with the statement "*It makes sense to do a homework task soon after the material is done in class*" – a statement that implies that spacing should be avoided/delays should be brief, and which therefore conflicts with the spacing effect. Interestingly, though, teachers appeared to be aware of the limited benefits of cramming, given that 84.4% disagreed with the statement "*The best way to learn something is to go over it repeatedly within the same hour*".

In terms of retrieval practice, only 37.7% of the sample gave the optimal response when asked if asking open questions is a good way to review a topic (item 9), a slight majority disagreed with the proposal that re-reading is more useful than testing (item 15), and just 26% rejected the idea that memories are mainly formed through frequent repetition (a suggestion which implies a focus on input and rote rehearsal rather than retrieval). Again, therefore, there was a tendency to reject statements that reflected desirable difficulties.

Interestingly, though, teachers were more divided on items that linked to interleaving. The statement: "*Learners benefit from mixing up lots of different types of problems, rather than doing one type of task at a time*" was endorsed by 63% of participants, indicating that a slight majority recognised that interleaving can help. However, it is notable that the reference to 'problems' could indicate entire tasks, extended activities or sets of questions, rather than the short items which benefit from interleaving according to the literature. A limitation of Study 3 was that none of the question specifically asked about interleaving or alternating examples; a further limitation was that no other statements on interleaving were included; it would also have been helpful to separate blocking vs. mixing at the initial learning stage rather than only using a statement which implies interleaving for later consolidation.

Higher scores among the present sample compared to the general public may be attributable in part to their higher education level as qualified teachers, but there may also have been an effect of sampling bias, and future replications with a more representative sample would provide more certainty regarding this finding. Twitter users may not be representative of teachers as a whole, and although a school sample was also used, this institution might not have a typical staff profile. To investigate these issues, an ad hoc comparison of the Twitter and school samples was conducted via an independent means t-test,[25] and this revealed that Twitter users scored significantly higher than participants recruited via a school (F[71] = 4.53, p = .037).

This supports the idea that sampling via social media can impact on results. Twitter users may be more likely to engage with the science of learning, for example by reading educational blogs which are easily shared through online links; popular Twitter hashtags such as #UKEdResChat (chat about education research methods and ethics) #cogscisci (chat about cognitive science for science teachers) and #lrnscichat (chat about the learning sciences hosted by cognitive scientists) support this idea.

6.2.4.2 The effects of classroom experience on beliefs about memory.

The finding that years of experience had no impact on the accuracy of teachers' beliefs supported the study's second prediction – prediction (b), above. There was no obvious pattern across the 5-year bands of experience; teachers with 6–10 had the highest mean score at 86.15 and those with 11–15 years the lowest with 79.25, but these differences were not significant. Granted, these are not longitudinal data, and I cannot make definitive claims about how the newer teachers might develop, or how accurate the more experienced teachers were in the early years of their careers. Nevertheless, the data overall are consistent with the idea that an

[25] Analysis for skewness and kurtosis alongside a Shapiro-Wilk test indicated that the distribution of the samples did not differ significantly from normality, and therefore bootstrapping was not used.

evidence-based knowledge of memory may not develop simply through classroom experience alone, and that misconceptions are not self-correcting.

The exact reasons why these misconceptions do not diminish with experience remain to be firmly established, but this finding fits with the points made earlier about the counterintuitive nature of memory and the slow feedback loop that occurs in the process of teaching practice. Indeed, what feedback is gained may promote rather than discourage some misconceptions – flawed strategies are sometimes better in the short-term, and, as noted earlier, teachers may mistake current performance for learning.

As noted above, the majority of participants did not endorse a 'cramming' approach of rapid short-term repetition (item 10), and in this respect their views did align with the research consensus. This finding is an exception compared to other items relating to desirable difficulties as noted above. It could be due to teachers having autonomously observed forgetting in the wake of cramming for a test, though this seems unlikely given that teachers often move on to a new topic after a test, and also would not typically know which study strategies their students had used. Another possibility is that the belief that cramming is flawed is widespread among the profession. A third possibility is that teachers may be drawing on their own personal experience of having crammed for tests and then forgotten things.

Prediction (b) framed the relationship between years spent teaching and performance on the task in terms of classroom experience, an idea that fits with other studies (e.g. Hood, 2016; Rivkin et al., 2005) that have found that teachers' effectiveness tends to increase for a few years post-training and then plateau. However, again there are other possible explanations. Another factor that should be taken into account is that as they progress through their career, a teacher's academic learning becomes gradually less recent. Recall that Magnussen and Melinder (2015) found that younger professional psychologists scored better on their task. And Ericsson (2017) argues that professional training is in some cases followed by a gradual deterioration in expertise, due to forgetting of technical knowledge. It is possible, then, that benefits that arise from experience balance against harm that is caused by forgetting of relevant technical knowledge.

It is also possible that the findings are an artefact of other differences, such as generational differences between groups. Future research could include a longitudinal investigation via surveys with trainee/early career teachers at various stages. Alternatively, or in addition, the views of different age groups on some other aspect of teaching, unrelated to memory, could be surveyed as a means of gauging how well they have retained the professional learning covered during teacher education. The online survey proved to be a flexible means of gathering data, and this flexibility could be further exploited.

6.2.4.3 The effects of a psychology specialism.

The sample in Study 3 also showed a lower level of flawed beliefs among those participants with a psychology teaching remit, supporting prediction (c). However, it's difficult to know the exact reason for this. An obvious assumption might be that people who reported psychology as their main teaching subject have a greater background knowledge relating to concepts such as memory and metacognition, and are therefore relatively immune to myths and misconceptions. However, this conflicts with the broader research literature, as (for example) Furnham (2019) did not find a relationship between prior psychology-related education and endorsement of myths, and psychologists do appear to be prone to misconceptions in legal contexts (e.g. Melinder & Magnussen, 2015). Another interpretation, then, could be that the benefits are less to do with past learning (it is in any case possible that teachers of other curriculum areas have studied psychology at some point), and more to do with identity – teachers of psychology may see issues such as memory and metacognition as falling within their sphere of interest and expertise. If so, they may prioritise engagement with the research literature. Future studies could explore this by surveying the independent professional learning activities that teachers engage in.

Another possible explanation for the latter finding is that while a background as a psychology student makes little difference as shown by Furnham (2019; a finding which also fits with the findings of McCabe, 2011, with regard to evidence-based study strategies), being a *teacher* of the subject has a more tangible effect. But if so, why is this the case? There are multiple possibilities. A teacher may well know relevant theories in more depth that their students do, and may also have more easy access to the information (having taught it several times, it will come more easily to memory).

Again, though, limitations of the sample in Study 3 need to be taken into account; with only 13 of the teachers surveyed reporting Psychology as their main subject, it is difficult to be confident that this aspect of the findings is representative of this group of teachers as a whole. A potential follow-up to this would be to conduct a direct comparison of teachers sampled in a different way, for example through subject networks. For example, secondary school teachers recruited from the UK's Association for the Teaching of Psychology (https://www.theatp.uk/) could be compared with teachers who are members of one or more other subject-focused professional secondary association. Such a survey is beyond the scope of the current thesis, but would help to shed light on whether it is the teaching subject in particular that led to more accurate responses. In such a follow up, it would be interesting to ascertain whether any advantage found among the psychology teachers was due to their engagement with the subject in general, or reflected a particular professional interest in the psychology of memory. This could be assessed via questions about professional identity.

Flaws in teachers' knowledge have implications for their professional learning. The current findings suggest that ongoing professional learning/training undertaken by teachers does not improve their understanding of memory, perhaps because this topic is absent from such training or because an in-depth understanding is not fostered. Teachers may benefit from more thorough education in memory and cognition, a conjecture that is supported by the finding that those teachers whose remit included psychology scored higher overall.

6.2.4.4 Methodological limitations in Study 3.

While Study 3 served as a pilot study into an under-investigated area, there were several methodological flaws. Some of these have been discussed already; the sample was small and not sufficiently representative, and the sampling method open to bias, particularly through the use of social media for recruitment. In addition, the use of listwise deletions, though common practice, is not always considered to be the optimal response to the issue of missing data, for it affects the power of the study and entails a slight risk of bias (Dong & Peng, 2013; Graham, 2009). Here, these risks were deemed to be low, both due to the small scale of the missing data and the cross-sectional nature of the research.

I looked for possible sources of bias among the six participants who dropped out. Four were recruited from the school and two from Twitter. One was from the 6–10 years' experience level, two from the 11–15 years level, and three from the 20+ years group. All were secondary teachers – by far the largest sector represented in the sample. All had completed the first screen with the first 5 items, and on these their mean score on those five items was 20.3 compared with a mean of 22.71 for the rest of the sample. These details provide an indication that the six participants may have dropped out because of finding the task difficult, and may therefore have scored lower than average had they completed it.

However, there is no obvious reason to suppose that this would have affected the main conclusions of the study. Given that the six who dropped out were distributed across the age groups, their scores were unlikely to have affected the conclusions regarding experience level and overall performance on the task. Even if all six had disagreed with the items on meaning, overall agreement would still have been in the high 80s%. It also seems unlikely, given that these participants scored lower than average on the early questions, that they would have been exceptionally accurate when it came to the questions on desirable difficulties, but it should be borne in mind that they represented around 7.5% of the initial sample, and their responses could therefore have moved any of the percentages up or down accordingly. Methods such as multiple-imputation or the full information maximum likelihood method could be considered for future studies into this area, particularly if the number of participants failing to complete a task is higher, or if data is thought not to be missing at random.

The instrument itself had certain flaws in terms of how well it can generalise to classroom practice and decision making. The use of short items may lead to participants making judgements based on their academic knowledge or espousing perceived general knowledge rather than responding in a way which reflects their classroom behaviour (McCabe, 2011). As such, it would be helpful to follow up on these findings by using prompt items which show realistic classroom scenarios rather than questions about memory in the abstract.

The current study also did not provide baseline responses to the questions; this could be done by surveying beginner student teachers. If it is indeed the case that a teacher's recall of professional academic learning plays a role in their responses, then practising teachers should score more highly on questions about memory compared to beginning teachers or teaching students. The latter may respond more like the members of the general public surveyed by Simons and Chabris (2011). It may well be the case, too, that there are important differences across teachers at the very start of their career compared to the latter part of the '0–5 years' group.

Taking these points together, it would be worthwhile to establish a baseline by studying teachers' beliefs at the point of training, or (preferably) before. Differences in beliefs across different age groups could also be measured more sensitively by looking at single year cohorts rather than grouping participants into bands of five years as was done in Study 3.

Study 4 followed up the present findings, and focused particularly on areas that led to the most deviation from the research consensus in Study 3 – desirable difficulties.

6.3 Study 4 – Teacher Beliefs About Desirable Difficulties

6.3.1 Introduction to Study 4.

Although small scale, the pilot survey (Study 3) addressed an under-researched area and provided a foundation for further work. It found evidence that misconceptions about memory may be at a lower level among teachers – especially psychology teachers – than has been found in the general population in previous research. It also piloted the use of an online survey as a means of gauging professional knowledge about memory processes.

Some of the poorest performance in the pilot study (in terms of being out of line with the research consensus) was found in response to items asking about spacing and retrieval practice. Interestingly, these are two of the desirable difficulties with the strongest evidence of efficacy (Dunlosky & Rawson, 2015). This negative relationship between efficacy of techniques and their endorsement among the profession mirrors the negative relationship between performance and learning discussed in the introduction to this chapter, and hints that teachers may be being misled by a subjective sense of ease (on the part of their students) or progress over shorter timescales.

However, there are limitations with the methodology used in the pilot. The items used were broad ranging, allowing for an analysis of various aspects of learning and memory. There were only one or two items relating to each concept, and reliability would be increased by including multiple items per concept. The items were also rather abstract, perhaps causing the participants to draw on remembered academic and technical knowledge rather than focusing on what they would actually do in a classroom context (McCabe, 2011). It was desirable for this broad survey to use brief, quick-to-answer items, but further research that focuses on a narrower range of concepts could use detailed, classroom-based vignettes to tackle this problem, following the advantages of this technique discussed in Chapter 4.

The findings relating to years of classroom experience provided some tentative evidence that beliefs do not become more accurate in line with years spent teaching. This is an important finding because it suggests that misconceptions about learning are not self-correcting, and in this sense it fits well with prior research into 'neuromyths' which was discussed in Section 6.1. However, there are other factors that must be taken into account. Several things can bring about changes in a teacher's professional knowledge as they get older, including time spent in the classroom, and also professional learning undertaken and forgetting of past learning. A person's attitudes and professional identity may also change across the lifespan (Bebeau & Monson, 2012).

As Wiliam (2010) has pointed out, teachers do not necessarily continue to improve beyond the point of initial training. Indeed, it is perfectly possible – as found in other professions – for performance to deteriorate rather than improve post-training (Ericsson, 2017). Teachers as a whole do seem to become more effective for a few years immediately after the preparation/training stage and then plateau, such that the performance of a teacher with five years' experience in the classroom is much the same as that of a teacher who has spent fifteen years in the classroom (Hood, 2016). However, early improvements could be due to a number of factors – they needn't imply a better understanding of learning and memory. Some of the other things that could account for these improvements include: improvements in a professional's ability to form working relationships with students, increased confidence, a broader repertoire of classroom techniques and materials, and a deeper familiarity with syllabus content.

Overall, then, it would be useful to look further at the issue of professional experience. In particular, the pilot did not fully distinguish between teachers across these earliest stages of their careers due to the fact that the first five years of practice were banded together as a single response. Training/initial education, the probation process and early career mentoring all have the potential to improve the accuracy of teachers' technical beliefs, and as a whole these first five years form a time when some change can be expected, as noted above[26]. The pilot therefore raised the issue of how professional skills and knowledge change over time, but did not provide enough information to fully compare teachers at different stages. It would be useful to study the memory beliefs of teaching students, and to compare their performance with that of experienced teachers. It would also be useful to get a more fine-grained estimate of the number of years of practice that experienced teachers have undergone, too. At the upper end, this could include teachers who have retired, and have therefore completed the maximum number of years in the classroom that they are ever going to.

Psychology teachers scored higher in Study 3, although it was unclear whether this was because of an understanding derived from their knowledge of learning-related topics (topics such as Memory, Intelligence and Learning feature in many introductory psychology courses), or from a greater engagement with evidence-based teaching strategies and the 'what works' movement. To distinguish this, it will be helpful to gauge the level of prior knowledge that

[26] In Scotland, there is also a five-year window after the PGDE when teachers can take part in funded part-time study to achieve a Masters in Education (MEd) qualification.

participants have about in areas. Have they, for example, read books about the spacing effect or interleaving? This information will allow us to distinguish between participants who score highly because they have developed an intuitive understanding of how learning works via their practice and scholarship, and those who score highly because they have read about concepts such as spacing and interleaving. For this reason, questions asking participants to declare their familiarity with retrieval practice, spacing, and interleaving will be included in the present study.

Study 2 included a metacognitive measure asking the learners to declare their own confidence in their answers, but this was not included in Study 3. Here, again, it would be useful to gain some idea of a participant's confidence in their own answers. From a professional learning point of view, it is important to know whether more confident teachers are also more accurate ones. If not, those who believe they are doing the right thing (when it comes to scheduling their students' practice, for example) may fail to engage with professional learning, feeling that it is irrelevant to them because their practice is already optimal. A simple metacognitive measure will be added here by asking in-service teachers to declare their confidence in their own responses after completing all of the classroom vignettes.

Finally, the limited nature of the sample in Study 3 needs to be addressed. As volunteers accessed the survey via social media for that study, they will share some of the same networks and interests as the researcher and may be unusually interested in and aware of memory research. It is perhaps inevitable to some extent – given that teachers must give their consent to take part – that samples in such research studies could be biased towards those with an interest in the subject under investigation. However, there are specific sources of bias associated with the demographics of social media users (Wojcik & Hughes, 2019). In the present study, data will be gathered more systematically and from a wider-ranging sample, including staff from multiple schools accessed via their school leaders, as well as teaching students accessed by contacting their university tutors, and retired teachers.

Five specific questions arise from these points, and despite an insufficiency of past research on any of these issues, I here attempt to make an interim prediction for each one:

(a) How accurate are the participants in responding to questions about spacing, interleaving and retrieval practice when these are set in the context of real classroom scenarios?

The use of vignettes will cause participants to draw more on practice than on technical knowledge, but this is unlikely to help if their intuitive judgements are flawed. Research summarised above (see section 6.1) suggests that most people do not have accurate intuition regarding the functioning of human memory, and Study 3 provided interim evidence that experience does not serve to counteract this when it comes to desirable difficulties. I therefore predict that most teachers will be out of line with the scientific consensus on these three issues, supporting the findings of Study 3.

(b) Are in-service teachers more accurate than student teachers?

The generally superior performance of teachers compared to the general public in Study 3 suggests that some technical knowledge was brought to bear, perhaps derived from the teachers' preparation or training. At the start of their careers, student teachers may lack this technical knowledge. I therefore predict that students will perform worse than practising teachers.

(c) Does the relationship between years of classroom experience and response accuracy found in the pilot study still hold in this study?

On the assumption that teachers are misled by a false sense of ease on the three desirable difficulties, the level of classroom experience post-training should not play a major role in the accuracy of responses. I therefore predict that as found in Study 3, there will be no relationship between years of practice and accuracy of responses.

(d) Does a teacher's declared level of confidence in their own answers correlate to their accuracy in responding?

Numerous studies discussed so far in this thesis suggest that metacognitive awareness of learning is often out of line with reality, and we can therefore predict a weak relationship between confidence and accuracy (this will only be investigated among practising teachers, on the basis that student teachers do not have classroom experience and as such their confidence in their choices is less educationally meaningful, perhaps reflecting personality traits or demographics).

(e) Does a participant's declared level of knowledge of the memory-related phenomena correlate to their accuracy in responding to the scenarios?

As noted above, an insight into background knowledge can help to distinguish teachers who have read a lot about areas such as spacing and interleaving from those who have intuited an understanding through experience. In general, a positive correlation between declared knowledge and accuracy is predicted among both students and in-service teachers, and would indicate that some professional reading can help judgements on learning and memory, even if this is de-contextualised. However, if the correlation is strong this would suggest that participants are drawing mainly on background technical knowledge rather than judging the vignettes on the basis of classroom practice, so findings here can also provide a check of the validity of the methods used.

6.3.2 Method.

6.3.2.1 Participants.

Two groups were sampled. First, ethical permission was obtained to sample PGDE postgraduate students at a large Scottish university. An advantage in using this population (rather than, for example, undergraduates, or school pupils who are interested in getting into teaching) is that the age and psychological development of these participants is more comparable to practising early career teachers. Sampling was done by convenience, a sampling method which is open to bias due to pre-existing variation among the sample but which is more reliable in cases where the sample closely resembles the target population (Robinson, 2014). I approached PGDE subject tutors in all secondary subjects; those in Biology, Business Education, Psychology, Computing and English took part by distributing the study to their students, who then had the option of taking part if they chose to do so. 77 took part; the mean age of this sample was 31.2 and the median was 28; some were recent graduates and others were mature students.

The second group consisted of experienced school teachers (n = 43). This part of the sample was obtained by approaching local authorities for ethical approval, and then emailing the survey to a relevant contact, for example the headteacher. A participant information sheet was attached to the email. Two authorities participated, both located in west of Scotland, as well as one independent school. The majority of teachers in those local authorities are trained at the same institution that the student sample were attending. In addition, a small number of retired teachers were recruited. As these are not working for a specific school or employer, I approached them via personal contacts. Again, these participants were emailed the survey together with a participant information sheet. In the subsequent sections, the procedure and data analysis for retired and in-service are combined.

6.3.2.2 Materials & design.

The survey was again prepared by the author and distributed using the website PsyToolkit. It featured 3 demographic questions, followed by nine scenarios/vignettes relating to memory in a classroom context, each with a 7-point Likert scale. This was used in preference to the 5-point scale (as used in Study 3) both to allow for more nuanced decisions by teachers when judging classroom scenarios, and to concord with the work of McCabe (e.g. McCabe, 2011) which used a similar scale. There was then a question asking participants to state how confident they felt in their own answers (practising teachers only), and then three questions which asked them to report how well they felt they understood the three concepts at hand – interleaving, spacing and retrieval practice – on a four-point scale: *not at all/very little/quite well/very well*.

Scenarios were designed to be recognisable and relevant to teachers across the profession, and as such, multiple subject areas were included across the vignettes. Each vignette presented a professional choice relating to spacing, interleaving, or retrieval practice (3 of each). For example:

A senior Geography class is learning about lakes. Their teacher divides them into two teams. Team A looks at pictures of different types of lakes all mixed together (MIXED), each with a picture of the lake, its name, and a label saying what type of lake it is. Team B see the same information categorised by type (that is, GROUPED), so that they see all examples of one type of lake together, then the next type, and so on. Finally, the two teams are given the same test.

The scenarios are reproduced in full in Appendix 13. I drew scenario numbers 2 and 3 from McCabe (2011, 2018) with minor modifications to the wording in order to fit better with the Scottish educational system and to make the scenarios more recognisable to UK-based participants. The remaining seven scenarios are novel, developed to follow a similar style and length to the McCabe examples.

The Likert scale was accompanied by a prompt which asked participants to judge which of the two options in the vignette would be more effective on the basis of a specified later attainment measure, such as a score on a later test. Terminology such as ‘interleaving’ and ‘spacing effect’ was avoided, and the choices were instead presented as follows:

- “Mixed” (i.e. interleaved) versus “grouped” (i.e. blocked).
- “Spread out” (i.e. spaced) versus “intensive” (i.e. massed).
- “Test” (i.e. retrieval practice”) versus “restudy” (i.e. passive re-reading).

This design decision aimed to ensure that participants thought about the scenario at hand, rather than relying on memory for terms which they may have seen recommended in their training or professional reading (McCabe, 2011).

The Likert scale was labelled with levels phrased as much/moderately/slightly better performance for each alternative (e.g. mixing) together with a central “about the same” option (see Appendix 13). In every case, the use of a desirable difficulty was viewed as the notional ‘correct’ answer, and in particular the more extreme choice (i.e. stating that the mixed/spread out/tested group would do “much” better rather than moderately or slightly).

6.3.2.3 Procedure.

Student teacher data were collected early in the academic year, over weeks 3–4 of a PGDE (Professional Graduate Diploma in Education) initial teacher education course. This allowed the students to be questioned at the very start of their teaching career, before they had spent any supervised teaching time in the classroom, but after they had made a commitment to the professional role. Most by that stage had experienced several university lectures and classes, and had presumably in most cases spent some time on relevant academic reading. Among this initial preparation there was one lecture on memory and learning, and reading relating to learning theories had been assigned.

The survey was shared with PGDE tutors in a range of different subjects, who then shared it with their classes by emailing the link to them. Participants completed the survey on university computers if they chose to participate. Scenarios were presented in the following order: *interleaving 1; spacing 1; retrieval 1; interleaving 2; retrieval 2; spacing 2; spacing 3; retrieval 3; interleaving 3*.

As an incentive, the researcher provided a link to the same short ebook about learning referred to in section 5.3.2.2 (see Chapter 5) at the end of the survey.

The other participating teachers – in-service and retired teachers – were recruited during the same academic year, and surveyed between October and February of that year. In most cases there was no direct contact between these participants and the researcher, but they were also provided with the ebook link on completion, and were provided with contact details. All participants received a (software generated) unique participant code, allowing them to withdraw from the study in the period between completion of the task and the beginning of data analysis if they chose to do so.

6.3.3 Results.

Prior to analysis I reverse-scored four of the nine vignette items such that ‘7’ always represented the optimal alternative for every question. This process revealed that the scenarios divided opinion more than might be expected with most classroom-related decisions. All of the scenarios attracted scores from the minimum 1 up to the maximum 7, with the exception of the second interleaving example (‘Scenario 4’ in Appendix 13) which attracted answers of 1 up to 6 (that is to say, every answer was given except for the optimal one).

Table 6.4: Overall responses to each vignette from Study 4.

Vignette	n	MEAN	SD
Retrieval 1	109	4.29	2.04
Retrieval 2	103	5.49	1.64
Retrieval 3	101	4.45	1.81
Spacing 1	110	4.95	1.89
Spacing 2	101	5.22	1.64
Spacing 3	101	4.75	1.79
Interleaving 1	119	2.84	1.31
Interleaving 2	104	3.49	1.68
Interleaving 3	100	3.06	1.57

Scores were then calculated for each individual vignette for both groups of teachers combined. This was based mean scores on the Likert scale. These data can be seen in Table 6.4, above; they have here been organised by concept, so it should be noted that the order is different from Appendix 13 which presents the items in the order that participants saw them. Data from incomplete questionnaires were retained for this process (they include, at the most, 43 practising teachers and 77 students, but the total number varies because later vignettes were answered by fewer participants, as can be seen in column 2, below).In terms of overall accuracy, it can be seen from Table 6.4 that most vignettes received responses which averaged above the midpoint (4), with the exception of the three interleaving vignettes. However, only the second spacing vignette (scenario 6 in Appendix 13) averaged above 5.

I next viewed each participant's responses to all three techniques. Only participant data for those who completed the whole survey could be included in this step (student teachers n = 67; 87% of those who began the survey; experienced teachers n = 33, 75% of those who began the survey), and identified those participants who had consistently endorsed one or more of the three strategies. This was operationalised as a score of 5 or higher on all three of the relevant vignettes (in line with the procedure followed by McCabe, 2011). The percentages of these 'consistent endorsers' was then calculated for each of the sub-samples (see Figure 6.1, below).

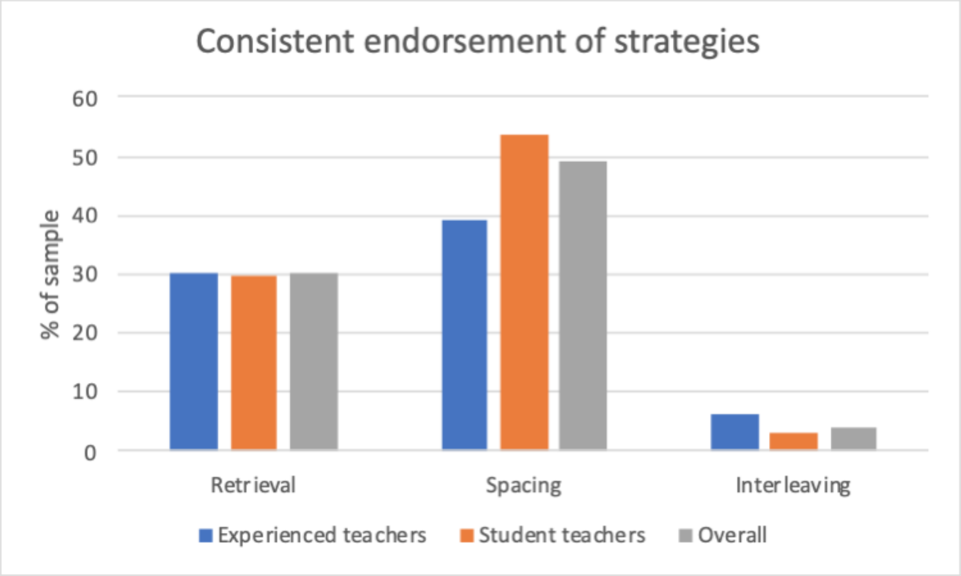


Figure 6.1: Percentage of participants who consistently endorsed the desirable difficulties presented in the vignettes.

As can be seen from Figure 6.1, all three strategies were endorsed consistently by a minority of participants overall, although the overall total was close to the halfway point for spacing; slightly more than half (53.7%) of student teachers endorsed the spacing effect on every occasion. Interleaving was rarely endorsed by either group according to this measure; only 6.1% of practising teachers and 3.0% of student teachers endorsed it across all three scenarios.

Together, these findings suggest that most teachers are not fully in line with the research consensus on retrieval practice, spacing and interleaving. There are signs that they are superior on these matters to what has been found in previous samples (e.g. student participants), but only just over 50% of the student-teacher sample consistently endorsed spacing, and in all other cases a minority of participants did so.

Multiple imputation (MI) using the MCMC method following analysis of the missing values was used at this stage using IBM SPSS Statistics for Mac OS, Version 25; it was found that 25% of cases had missing data, with 8.3% of values missing overall. Values were assumed to be missing at random. I analysed these missing values to assess for bias, and determined that the main cause of missing data was participants dropping out mid-way through the task and failing to complete the final questions. No other pattern of bias was apparent, supporting the assumption of data missing at random.

A comparison between student teachers and in-service teachers was run next. Scores were pooled across the vignettes that tapped into the same concept (e.g. interleaving) resulting in a mean score between 1–7 for each concept per participant, at least for those participants who had completed at least one question on the concept. MI was used to replace missing values for this analysis; the mean score per vignette among the subset of participants with complete data as described above was 4.42 for student teachers and 4.03 for in-service teachers across all vignettes completed. Table 6.5 shows these results broken down by concept for both students and teachers, including both the original and pooled imputed data.

Table 6.5: Descriptive statistics, student and practising teachers by concept.

	Practising teachers			Student teachers		
Concept	Retrieval	Spacing	Interleaving	Retrieval	Spacing	Interleaving
Mean	4.45	4.28	2.90	4.64	5.12	3.17
SD	1.49	1.64	1.22	1.49	1.39	1.06
Imputed mean	4.48	4.29	2.96	4.63	5.14	3.22

In order to determine whether there was a difference between practising and student teacher scores without increasing the chance of Type 1 errors, a MANOVA was conducted. Box's test of equality of covariance matrices was non-significant for all MI data sets. Equality of variance was accepted for the retrieval practice and interleaved DVs, but for spacing was rejected because Levene's test was significant (p = 0.031). However, an advantage of MANOVA is that such violations of its assumptions tend not to affect significance level and power (Ito, 1980).

Focusing on the original data, there was a statistically significant difference between the two groups across all three concepts, F (3, 107) = 2.925, p = .037; Wilk's Λ = 0.924, with an effect size of +.076. The five imputed data sets showed p-values as follows: 0.006; 0.007; 0.016; 0.041; 0.101. It was cautiously accepted that these findings supported a difference between the two samples, albeit (on the basis of the means in table 6.5) only in some concepts and not others, and with one set of imputed data as an outlier. Comparison of between-subjects effects for the three component concepts then revealed a significant effect of experience level on accuracy of beliefs about spacing [F(1, 109) = 8.178, p = .005] with an effect size of 0.07, but no such effect for retrieval practice [F(1, 109) = 0.399, p = 0.53] or for interleaving [F(1, 109) = 1.50, p = 0.22]. This showed that while students scored better than in-service teachers on the task overall, the difference can be put down particularly to the spacing vignettes, at which they performed better, and the effect size was low.

Overall, these findings clearly did not support the predicted advantage of practising teachers over student teachers, albeit that the evidence for the reverse (an advantage of student teachers over practising teachers) was mixed.

The imputed data from the three concepts were also combined to form an overall average accuracy score for each practising teacher who had responded to at least one vignette on each concept, with a minimum possible score of 3 and a maximum of 21; this will henceforth be referred to as a participant's *total score*. A 2-tailed Pearson's product moment correlation revealed that there was no significant positive relationship between experience and accuracy by these measures, with the trend in the opposite direction (n = 38; r = -0.113; p = 0.475).[27]

Again focusing on practising teachers in particular, this group of participants was asked to report their confidence in their own responses to the scenarios. There were insufficient data to complete MI for the confidence scores, and therefore only in-service teacher participants who had completed this question were included. In effect, this amounted to listwise deletion of participants who did not finish the survey. Confidence was not normally distributed according to a Shapiro-Wilk test, and therefore bootstrapping was applied (see Efron & Tibshirani, 1993; Field, 2018), again with the bias corrected and accelerated method (see Chapter 5), and using 1000 values. The analysis revealed that these two variables (accuracy and confidence in own

[27] This compares to -0.232, p = 0.155, with the original data sets (n = 38).

answers) were not significantly correlated, with a weakly positive trend (r = 0.11; p[33] = 0.540).

Finally, after responding to the vignettes, all participants (students and in-service teachers) were asked to report their level of background knowledge with respect to the three concepts studied (retrieval practice, interleaving and spacing). Responses to this question are shown in Table 6.6, below.

Table 6.6: Declared knowledge of participants regarding the three desirable difficulties.

	Practising Teachers				Student Teachers			
Declared knowledge of concept	Retrieval	Spacing	Interleaving	%	Retrieval	Spacing	Interleaving	%
‘Not at all’	8	6	14	28.3%	1	4	16	10.4%
‘Very little’	15	11	10	36.4%	10	13	31	26.9%
‘Quite well’	9	14	9	32.3%	43	42	17	50.7%
‘Very well’	1	2	0	3.0%	13	8	3	11.9%

Students’ declared familiarity with concepts was higher than that of in-service teachers, with 62.6% overall responding with ‘quite well or ‘very well’ to the three concepts overall compared to 35.3% of in-service teachers. Taking into account the ordinal nature of the data, an independent samples Mann-Whitney U test was used to compare levels of prior knowledge with each of the options in Table 6.6 scored from 1–4, and this confirmed an overall difference between the two categories of participants (p = 0.01, df = 98; z = -3.39).

It is interesting to note from Table 6.6 that despite 51 out of 67 of the student teachers (76.1%) declaring that they were at least ‘slightly confident’ with the concept of interleaving, only 3% were classified as consistent endorsers in the analysis described earlier in this section. This may fit with an interpretation that student teachers had encountered interleaving and other evidence-based strategies in their academic reading but were inexperienced at recognising them in classroom scenarios.

6.3.4 Discussion.

6.3.4.1 Accuracy of overall responses.

In terms of prediction (a), this study showed overall poor concordance with research evidence in terms of how both student teachers and practicing teachers interpret memory-relevant scenarios that connect to desirable difficulties. The mean score for three of the vignettes were below 4 (the midpoint on the Likert scale), four were between 4–5, and just two averaged above 5.

However, there is some nuance to this. The results for spacing in particular are more accurate than those found in the study of students by McCabe (2011), in which under 10% endorsed spacing as a study strategy. Just under 50% of the current participants from the whole sample endorsed spacing consistently, that is to say, on all three scenarios that presented a spacing vs. massing choice.

Accuracy on the retrieval and interleaving vignettes was less accurate, however, with exactly 30% endorsing retrieval practice consistently, and just 4% endorsing interleaving consistently. This could suggest that the idea of using ‘tests’ is seen as antithetical to learning, and mixing examples perhaps seen as a source of confusion for learners.

Subsequent to the design and data gathering used in this study I became aware of another survey with some methodological similarities. In a survey published on the website of *The Learning Agency,* Boser (2019) reported teacher responses to vignettes about evidence-based learning strategies and neuromyths. Interleaving, retrieval practice, and spacing were all included in his study, which sampled around 200 American educators. The findings of his work and the present study are summarised in the following table (Table 6.7):

Table 6.7: Comparison of the findings of Study 4 with those of Boser (2019). The main number for retrieval practice reflects a scenario question similar to the present Study 4, while the number in brackets relates to direct questioning (similar to Study 3 of this thesis).

Strategy	Present study	Boser (2019)
Retrieval practice	30%	59% (31%)
Spacing	49%	approx 60%
Interleaving	4%	20%

Although there are some differences between the two studies – particularly the better performance on a retrieval practice scenario – the overall pattern is similar. Some of the differences may relate to the specific scenarios used, and others to the fact that the Boser study used a 3-item Likert scale in contrast to the 7-item scale used here. The broad similarity between the two studies lends further confidence to three main conclusions that I have expressed so far from Study 4: that teachers are broadly out of line with the research evidence, that they nevertheless score higher than has been found in surveys of students, and that endorsement falls broadly in this order: spacing highest, then retrieval practice, then interleaving.

6.3.4.2 Differences between student teachers and in-service teachers.

Regarding the difference between student teachers and in-service teachers – research question (b) – teachers and students in this study differed in how they responded. However, rather than providing a baseline, it was students who performed better in terms of accuracy (at least with respect to items on spacing). Students also reported more awareness of the techniques under investigation.

This is surprising for a number of reasons. One is that the students were at a very early stage of their careers. Most[28] had no formal classroom experience, and they had had little time to take on board the academic lessons of their course, either. In addition, prior research by Surma et al. (2018) has suggested that retrieval practice and spacing – the two concepts on which students answered more accurately – tend to be absent from teaching textbooks (and it can be assumed that interleaving is, too).

It is entirely possible, however, that the student teachers had read about these concepts online or heard about them from tutors. Although it was an early stage of their teaching course, they may have spent months (or even years) in the lead-up to starting the course taking time to read books and academic journal articles about teaching. And as noted in Chapters 2 and 5, a number of recent books and reports include mention of spacing and interleaving.

One aim of the present study was to provide a baseline measure; instead, novice teachers have outscored experienced ones. This suggests that another approach needs to be taken in future studies to better establish a baseline for related tasks.

[28] It is possible that some had shadowing experience prior to the PGDE course, or had perhaps worked in classrooms where a formal teaching diploma is not required, e.g. as a language assistant abroad.

A second consideration relates to professional learning. Student teachers made more accurate judgements than their experienced peers, and showed some knowledge of concepts relating to memory. In contrast, experienced teachers appear to have forgotten some of the academic understanding of these concepts, in a way that links to the points made by Ericsson (2017) about forgetting of technical knowledge post training. Clearly if classroom practice is to be optimal, more needs to be done to ensure that current teachers understand evidence-based practice and are able to make good judgements about learning situations.

This latter point leads us on to the data on teacher experience and related concepts which were investigated as moderator variables.

6.3.4.3 The role of experience, confidence and prior knowledge.

Study 3 found no evidence that accuracy of beliefs and years of experience are positively correlated and despite the different methodology for gathering the accuracy scores, this finding was replicated in Study 4. In terms of research question (c), it appears that accuracy in memory-related judgements do not improve in line with years of experience in the classroom. Again, there was no evidence of a positive relationship between years of classroom experience and the accuracy of answers, with a non-significant trend in the other direction.

The student teacher data also lends some support to prediction (c). It might be assumed that on the basis of a combination of both training and experience, practising teachers would have a great advantage in the task. However, the student teachers outperformed their more experienced counterparts as discussed above. It may be interesting to note that the two (tied) highest overall scores (16.67) came from a practising teacher with 3 years' experience and a from student teacher. The lowest (5.00) was from a teacher with 23 years' experience.

However, as with any correlational findings, the results have to be interpreted with caution. The more experienced teachers completed their education and/or training at an earlier point in time, and as with any inter-generational comparison, age effects have to be considered. For example, older teachers may be more conservative in their judgements, and therefore less likely to select an option that seems unusual or unfamiliar. That is to say, participants may have been using different heuristics to judge the scenarios in cases where they felt uncertain.

In terms of confidence in one's own choices (research question d), past metacognitive research suggests that confidence is a poor indicator of accuracy, both in terms of judgements of learning and more broadly (e.g. in legal contexts; Loftus, 2019). In line with this, the correlation between performance on the task and declared confidence was small and non-significant. While again some caution is needed in interpreting correlational data, the findings here seem to fit with the broader literature which suggests that when it comes to desirable difficulties, people don't know what they don't know. From a professional learning point of view this is important, because teachers may be making flawed classroom choices and yet feel highly confident in those choices.

However, it should be noted that the confidence ratings were not on a ratio scale. In future studies, it would be a good idea to consider using a percentage prediction of performance, as was used in Study 2.

In terms of subject specialisms, it would have been interesting to follow up on the points made in the previous section regarding the proficiency of psychology teachers. However, in the present sample, there were too few practising teachers of psychology (n = 2) among the in-service teachers to make this judgement. This is perhaps not surprising, as many secondary schools in Scotland do not have a psychology department at all. It would be interesting to follow up the finding from the pilot study with an investigation that looked at psychology teachers particularly, as discussed earlier, and which used the in-depth survey from the present study.

It was possible to analyse the mean total scores in terms of subject specialism across both teachers and students, and this revealed no particular pattern (see Figure 6.2); while Psychology specialists scored better, they were not top overall, and none of the differences reached significance. A broadly similar pattern could be seen with the imputed data.

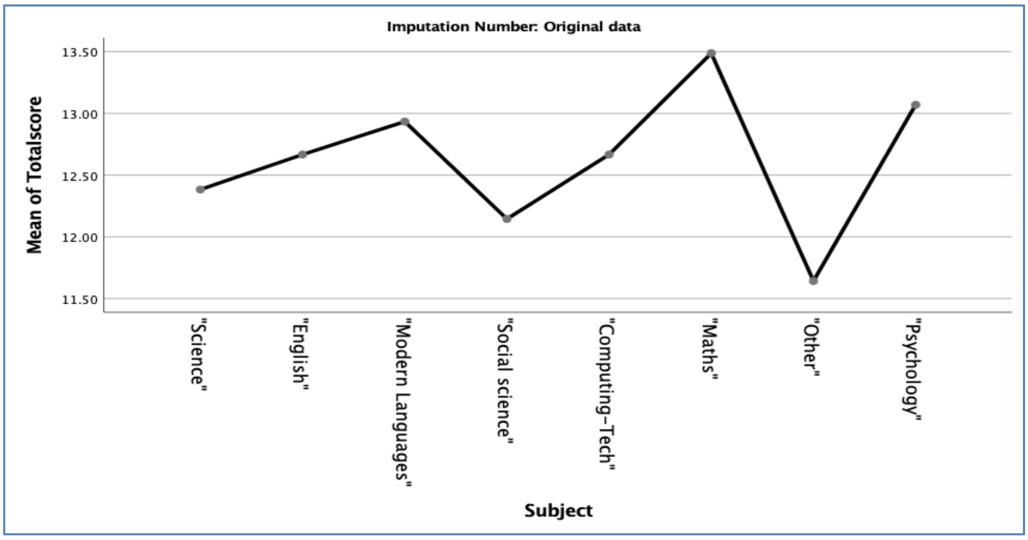


Figure 6.2: Total score analysed by subject specialism

6.3.4.4 Analysis of methodology.

Study 4 presented a series of vignettes to teachers and student teachers, and analysed their responses on the basis of what was viewed as the optimal research-based response. A reasonable question could be raised over whether the specific scenarios presented are sufficiently in line with the research evidence that the optimal answer is actually correct. What if the teachers were right to not endorse interleaving as it was presented, for example?

Although they were largely novel, the vignettes did draw heavily on research evidence. Scenario 9, for example, is based around the Birnbaum et al. (2013) study of interleaved butterfly images, and scenario 4 is based on interleaved psychological case studies from Zulkiply et al. (2012). Scenario 1 follows the same pattern with different items. The retrieval practice vignettes all feature tests. Tests are not the only means of prompting retrieval, but have repeatedly been found to be superior to re-reading (e.g. Agarwal et al., 2012; Chan et al., 2018, McDaniel et al., 2013; Roediger & Karpicke, 2006), and retrieval practice is consequently sometimes known as the ‘testing effect’. In terms of spacing, scenarios 2 and 7 clearly feature spaced practice; it could be argued that scenario 6 does not, as it refers to delays mid-way through a topic, rather like the work of Bird (2010) and as was done in Study 3. In this sense, it is less clearly evidence-based than the other scenarios (and was also the scenario which was most endorsed overall; see Table 6.4, above, ‘Spacing 2’). On balance, while it doesn’t specifically refer to spaced out practice of specific skills or knowledge, the delayed study described in scenario 6 would at least provide more of an opportunity for such practice than would the more intensive option. Ultimately, the only way to be certain whether the novel scenarios are actually superior would be to run empirical studies testing this – they are evidence-based, but an improvement for future studies would be for a team of experts in desirable difficulties to review and comment on the vignettes.

In terms of the data analysis used, endorsement for a strategy was operationalised as a score of 5 or more on all three relevant vignettes. Granted, this was a strict criterion, but it is an important one – if a technique is useful and its benefits fit with a teacher’s beliefs about learning and memory, then they should use it all of the time, not just occasionally. It should be borne in mind that these are *judgements* of what would lead to better learning, so 100% endorsement does not imply that the teacher believes their own practice to always be optimal, just that they recognise and support good practice in principle.

On the other hand, it can be instructive to consider the flip-side of this criterion. Out of practicing teacher participants who responded to at least one vignette on spacing (n = 36), 83.3% endorsed the correct option with a score of 5 or above on at least one occasion, and just 16.6% never did so. This may suggest that while less than half of the profession endorse this desirable difficulty overall, there is some sympathy for it – potentially indicating an openness to professional learning on the issue.

Multiple imputation using the MCMC method was used to replace missing data in some steps, and this is generally seen as more valid than listwise deletion of data (Dong & Peng, 2013; Manly & Wells, 2015), as the latter is open to bias. The main reason for this is that participants who dropped out mid-way through a task may be non-random. However, many common reasons for dropping out of longitudinal studies are unlikely to apply to a brief questionnaire, and it seems more likely that teachers who dropped out were either too busy to continue or had technical glitches with the software. It is possible though, in some cases, that they did not proceed because they found the early stage of the task too hard, as is indicated in the results from Study 3.

It was felt that multiple imputation was not appropriate for item-level analysis (such as whether teachers had responded to specific vignettes) and it was therefore used on the total score statistic. IBM SPSS Statistics provided an analysis that shows both pooled results and imputed data alongside the original data, and in no case did the overall conclusion of significance or direction differ between these different data sets, although the p value did.

6.4 General Discussion of Studies 3 and 4

6.4.1 Review of key findings.

The two studies reported in this chapter have found strong evidence that teachers' knowledge of memory differs from that seen among the general public and is accurate in some spheres, but is nevertheless out of line from the research consensus when it comes to desirable difficulties. This extends research previously conducted on legal professionals into a new and previously-unstudied domain. They are also the first published studies to investigate the relationship between years of classroom experience and teachers' understanding of memory concepts, and the first to look specifically at the memory beliefs of student teachers; in both cases, the results were surprising and may run contrary to most people's expectations.

Study 3 found that teachers were accurate on matters relating to the importance of meaningful understanding, but inaccurate on statements pertaining to desirable difficulties, especially spacing and retrieval practice, though the questions on interleaving in Study 3 were very limited, making this section inconclusive. The findings of Study 4 confirmed that most teachers do not endorse retrieval practice, and interleaving was even less widely endorsed when framed in scenarios that reflected authentic (initial) learning situations.

Findings relating to spacing in Study 3 were mixed, with most participants in that study rejecting the idea that practice should happen when students are "on the verge of forgetting", but also rejecting cramming as a learning strategy. The findings of Study 4 present a clearer and more consistent picture, perhaps because the vignettes provided realistic detail and related to real classroom situations rather than the concepts in the abstract. While spacing is not intuitive to teachers, it was less widely rejected than the other desirable difficulties under investigation, and the finding that most endorsed the strategy at least once (see 6.3.4.4) suggest that many may well be sympathetic to this idea if they find out more about how to use it, with professional learning implications.

Perhaps surprisingly, both studies found that experience made no difference to accuracy on the questions about memory, and the trend was in the opposite direction when it came to desirable difficulties. This pattern seems to confirm the speculation made at the outset – that such features are counterintuitive, and teachers will not learn them simply through time in the classroom. The finding that students scored better than in-service teachers, while surprising, confirms the idea that a technical, theory-based understanding of such phenomena may be more useful than classroom experience alone.

There is good reason to suppose that flawed thinking about memory will lead to flawed practice. It appears that some teachers mistakenly think that the "*best way to learn something is to go over it repeatedly within the same hour*" (Study 3, item 10), for example, or that consolidation should occur as soon as possible. These are likely to lead directly to flawed classroom choices over the timing of practice which can undermine the development of flexible, lasting long-term memories, as discussed in Chapter 2. Study 4 helped to confirm that rather than just occurring in response to general questions, such flawed thinking is still present when survey items are tied to specific, fleshed-out classroom scenarios. Although there is as yet a lack of research investigating the specific effects (if any) of such beliefs on attainment (and the effect of such beliefs is therefore just theoretical at present), there is evidence connecting flawed beliefs about memory with poor outcomes in independent study (e.g. Hartwig & Dunlosky, 2012). The present research will help to pave the way for such research to be conducted in classroom teaching contexts as well.

6.4.2 Analysis of methodology.

If the research reported above is to be replicated and extended, certain methodological limitations should be addressed. One is the method of sampling. The limits of social media have been discussed; the findings from Study 3 suggested a bias, with Twitter users scoring more highly than participants sampled from among the staff of a secondary school.

In study 4, there was a lack of independence between the samples taken of both the student teachers and the practicing teachers in Study 4. The students were all drawn from the same ITE institution; the teachers came from a limited set of secondary schools. It is entirely possible that teachers in the same school will show performance that correlates, given that in-service professional learning may be done in common across a school or even across a local authority. This could have affected the validity of my calculations of the standard error. In a follow-up study, the methodology could be adapted (at least for the practising teachers) by treating each participating school as a cluster.

When it came to the student teachers sampled for Study 4, these were recruited very early in their course, but had still been exposed to the reading, meaning that this group did not serve as a baseline. Indeed, it seems likely that their level of endorsement of spacing (which tends not to be endorsed by students more broadly) reflects some professional reading of the research literature.

In order to establish this more definitely, a follow-up study could recruit student teachers before they begin their course, for example by distributing the questionnaire at the course interview stage, or draw a broader comparison with members of the public on the same tasks, for example adults in other professions, matched according to past educational level. This would allow a true baseline, unbiased by possible interest in matters relating to teaching practice. It would also be helpful to establish an upper boundary by replicating with a purposive sample who have particular experience and/or knowledge relating to desirable difficulties, for example a group of cognitive psychology researchers (as done by Simons & Chabris, 2011) or teachers with a particular interest in this area. This would give an idea of what good performance on the task would look like, which would be useful for professional learning purposes as a target, and would also provide a further check of the accuracy of vignettes.

In comparison to the pilot study, the vignettes left less room for subjective interpretation by participants, making them more accurate reflections of participants thought processes.

Nevertheless, the delays referred to in the spacing vignettes are not actually specified[29]. This could be important given that Study 3 found that teachers tended to reject lengthy spacing but also to reject cramming. Taken together with the present data showing that the spacing vignettes were more frequently endorsed than the others, there remains a question of exactly what time delays teachers would see as ideal, and future studies may need to add more subtlety in the sense of not presenting spacing as a binary 'do or do not' choice.

The wording of the retrieval and interleaving vignettes largely followed previous research studies, but it could also be reviewed. If a term such as 'practice quiz' was used instead of 'take a test' for the retrieval items, teachers may be more amenable to the possibility that this would promote learning (especially as practice quizzes or self-tests are often used as formative assessment strategies; see Black et al., 2004). Interleaved alternatives may be more appealing to respondents if they are phrased in terms of providing 'contrast', rather than being 'mixed' (which, as a partial synonym of 'confused', is potentially off-putting to educators).

These choices have obvious implications in terms of the internal validity of the findings, but wording which makes evidence-based techniques sound more appealing could also be useful for professional learning applications.

Another issue for future studies was the length of the task and the drop-out rate. Even though the task was quite short (median completion time was 8 minutes, and the advert suggested that it would take "approximately 10 minutes"), a proportion of participants did drop out before the end. A shorter task would be preferable – perhaps one that could be done at the beginning of a lecture or in-service training session, and took just 2–3 minutes. One thing that is of interest here is the performance of individual vignettes. If the present findings can be replicated, it may be possible to develop briefer versions of the task that work equally well.

[29] For example, the third spacing vignette states "the same coding processes are practiced across several lessons, only once per lesson (that is, the learning is SPREAD OUT)".

6.4.3 Implications for future research.

One of the most useful aspects of the two studies reported above is that they helped to establish a novel research agenda, and to lay the groundwork for future investigations. They shine a spotlight on teachers' beliefs about memory in a way that has not been done before; despite multiple studies of teachers' beliefs in neuromyths, only a couple of recent studies have surveyed beliefs about memory. The present studies used both short items and vignettes, and explored the role of classroom experience in a way that is entirely novel in the literature.

It is therefore worth considering the existing findings as the beginning of a research programme. What, then, would be the next steps?

In terms of the sample used, it has already been suggested that future research could establish a better baseline – what exactly would the findings be if someone with no technical or practical experience of teaching were to respond to the vignettes? Establishing this would help any particular sector (for example, a school or national education system) to use similar techniques to gauge how well established an understanding of desirable difficulties was among their staff, as would repeating the task with a purposive sample who were selected for their interest in desirable difficulties to provide an upper boundary score or target.

Some recent studies have also compared teacher beliefs with other relevant sectors. For example, a survey by Macdonald et al. (2017) on neuromyths looked at the general public, educators, and those with high neuroscience exposure. McCabe (2018) looked at professionals in academic support centres. Future replications of the present study could look at non-teaching staff in a similar way, perhaps including school leadership teams, or drawing a comparison between department heads and classroom teachers.

In terms of materials, both surveys were quite short, and it would be worthwhile to take some of the stronger elements of both; the best-performing vignettes, for example, alongside the most informative items from Study 3. To address the restrictive and binary presentation of spacing in the vignettes, understand this further, (as in Study 4) but instead asking participants to estimate the optimal amount of a spacing for one of these scenarios. Alternatively, this could be tested using an authentic planning task, for example via filling in a weekly or monthly schedule of lessons, and distributing study sessions over a finite number of days or weeks. Such a test could be timed before and after an intervention such as a professional learning course to gauge the effectiveness and time efficiency of various professional learning options.

Further research should also determine the practical effects (if any) of misconceptions about memory, perhaps via classroom observations. This would clarify whether flawed teacher beliefs have a detrimental impact in line with previous findings relating to learner beliefs. It would also be helpful to develop a taxonomy of memory-relevant educational tasks to inform the selection of future survey questions, and to survey memory experts (as in Simons & Chabris, 2011) to better establish the scientific consensus on the issues covered.

This more extended survey could further set the groundwork for future investigations into teacher professionalism. For example, longitudinal work could be carried out, following teachers over the first several years of their time in the profession. Retired teachers could also be investigated, to get an idea of how professional understanding of memory looks when the number of classroom experience is at its maximum.

6.4.4 Implications for practice and professional learning

The findings of both studies in this chapter suggest that teachers' judgements – and presumably therefore their choices and behaviour – will be flawed with respect to certain memory processes. While it is likely that they will avoid cramming approaches and will value the role of meaningful information in the classroom, most appear to eschew strategies that would lead to desirable difficulties in learning, and this is to the detriment of students. While spacing was more widely endorsed than retrieval practice (specifically, testing) and interleaving in Study 4, fewer than half of the in-service teachers consistently endorsed it.

Another major implication is the ability of teachers to guide learners' independent study, as discussed at the beginning of this chapter. When using desirable difficulties such as interleaving, students are likely to notice immediate negative consequences of these difficulties – mistakes, psychological effort, and a slower rate of perceived progress – and interpret them as reasons to avoid strategies such as spacing and interleaving in favour of 'easier' alternatives. It will often fall to the teacher to help build a better metacognitive understanding of how memory and learning work.

If teachers are avoiding either using such strategies directly or recommending them, then it could be suggested that professional learning experiences to address this would be beneficial, an idea which is supported by the superior performance of student teachers in Study 4 (together with their self-reported greater awareness of the three desirable difficulties studied). One possibility is that teachers could undergo a programme which would systematically teach theories of memory and metacognition, helping them to understand some of the counterintuitive ways that memory plays out in the classroom, such as the performance vs. learning distinction.

Similarly, the difficulty that teachers had with the vignettes in Study 4 compared to the relatively good performance in Study 3 supports that idea that transfer of conceptual knowledge to real situations is a challenge, and that simply providing information may not be enough. The idea that knowledge does not always transfer easily has been raised throughout this thesis in the context of transfer (e.g. see Chapter 2, section 2.3.4). In the legal profession, too, there is evidence that simply providing knowledge about memory does not always lead to good performance (Cutler & Penrod, 1995; Magnussen & Melinder, 2015). However, there is also evidence of a possible solution from that field. Wise and Kehn (2020) provide evidence that the interview–identification–eyewitness factor (I-I-Eye) method can help with more accurate analysis of eyewitness testimony among jurors and legal professionals. The I-I-Eye method provides a framework which professionals can use to work through the validity of evidence and apply their own technical knowledge at each step of the process.

A similar approach in education could be possible, but to do so, technical theory-based knowledge of memory would be required rather than simplistic generalisations or approaches that lean too heavily on professional reflection, given that the present studies have confirmed that experience and intuition cannot correct errors related to desirable difficulties. Professional learning with this purpose should also directly tackle misconceptions rather than just providing information (Will et al., 2019), helping participants to develop a metacognitive understanding of common misconceptions. Appendix 14 (parts 1 and 2) describes such a course that I wrote myself and have taught to primary education student teachers alongside one of my university colleagues. The aim of the course was not just to provide information but to guide participants to recognise and analyse misconceptions about memory and learning which are commonly shared online and in teaching-related documents. However, the course did not provide a step-by-step framework like the I-I-Eye method, and this is an area for future development.

Finally, and as noted in the interim discussion of Study 3 (section 6.2.4), a teacher's professional identity plays a role in terms of their willingness to engage with the research-based concepts, and this is supported by some of the less prominent findings here, such as the superior performance of student teachers (Study 4) and of Twitter users (Study 3). I will return to the implications of this point – taken in the context of all of the foregoing chapters – in the general discussion (Chapter 7).

7

General Discussion

7.1 Review of Findings

In this chapter I address the implications of the findings for both teaching and professional learning. What is novel in my findings and analysis, what can be used immediately, and what areas require further exploration and confirmation before the findings can start to form the basis of practical applications? In order to answer these questions, I will first briefly review the findings themselves – both theoretical and empirical – highlighting the main points raised in the previous chapters:

7.1.1 Key points from Chapter 1.

Chapter 1 outlined what we mean by 'learning' in classroom contexts. It highlighted and discussed the theoretical idea that learning should be distinguished from performance, and outlined a series of techniques – desirable difficulties – which harm performance but boost learning over the long term. These techniques include spacing and interleaving, as well as some other education-relevant techniques such as retrieval practice and self-explanation. It was also noted that delays and variation fit well with the theoretical explanation of desirable difficulties centred on transfer-appropriate processing, on the basis that real-world demands to use what has been learned tend to occur after a delay and in an uncategorised manner (Bjork & Bjork, 2019; Bransford et al., 1979). It was also noted that there is an important difference between taking in information and learning new concepts, and that prior research in the field has often neglected this in favour of simple experimental models of spacing.

7.1.2 Key points from Chapter 2.

Chapter 2 reviewed the background research, focusing in particularly on two desirable difficulties that pertain to item order – spacing and interleaving. These techniques can be considered to be evidence-based, and are relatively straightforward to apply in the classroom compared to other interventions. However, limitations in the research evidence were further explored, in particular the idea that experimental evidence often lacks context and does not connect to ongoing programmes of learning. The evidence was also put into the context of its cognitive and metacognitive underpinnings, commonalities and interactions between the techniques were identified, and implications for professional learning were discussed. This chapter helped to establish some tentative research priorities, including the need for a systematic review of interleaving.

7.1.3 Key points from Chapter 3.

Having identified this as a research target, **Chapter 3** described a systematic review of interleaving that has been carried out – a piece of scholarly work which was pre-registered via PROSPERO, and which includes both a narrative review and a meta-analysis. 26 studies met the inclusion criteria for the review, with a subset of constituent experiments forming the basis of the meta-analysis. Memory (as tested by presenting studied items from a learned category) showed an interleaving benefit with effect sizes of up to +0.65, 95% CI [0.50, 0.80], and transfer (as tested by presenting novel items from a learned category) a benefit with effect sizes of up to +0.66, 95% CI [0.49, 0.80]. Interleaving was found to be of greatest use when differences between items are subtle, and the benefit extended to both art- and science-based items and also to delayed tests, with implications for education. The review revealed that the literature is dominated by laboratory studies of university undergraduates and the need for future school-based research using authentic classroom tasks is discussed.

7.1.4 Key points from Chapter 4.

Having identified the need for further empirical studies in Chapters 2 and 3, **Chapter 4** reviewed and evaluated the methodology choices available. It addressed three ‘problematic assumptions’ that often pertain to practice and research when it comes to memory – that education-relevant memory processes are absolute and universal in that they pertain to every learner, that memory processes are relatively stable and amenable to investigation and manipulation, and that memory processes are open to scrutiny by their own users and can be guided on the basis of intuition. All three of these assumptions were analysed and criticised on the basis of the research literature, drawing particularly on the literature on metacognition. On the basis of this analysis, key methodology options were set out and evaluated, and a plan was devised which included field experimentation and a practitioner survey. Among the major methodology points highlighted, I identified the potential benefits of interleaving manageably small category sizes, using syllabus-related materials including skills-based tasks, and using between-participants designs in order to gauge metacognitive beliefs. Computer-based methodology on the classroom was identified as a way of ensuring both randomisation and authenticity. The analysis of the use of surveys reflected on techniques for gaining an unbiased sample, and focused on the benefits of vignettes.

7.1.5 Key points from Chapter 5.

Chapter 5 then outlined two experimental studies conducted in schools. *Study 1*, a small-scale pilot study, used material derived from psychology topics in the Scottish school curriculum to present information about phobias in an interleaved format and supporting evidence about phobias in a spaced format to beginner pupils. A number of limitations in the way that interleaving and spacing were conceptualised limited the conclusions from this study, but it was successful and novel in that it trialled a computer-based methodology and featured school pupils and authentic, curriculum-based materials – features that are largely absent in the broader literature on interleaving. Statistically, there was no advantage of interleaving and spacing led to a short-term deterioration in performance. The findings also provided some evidence that in the context of real sequences of classroom activities, interleaving and spacing may interact. *Study 2* followed this up with a novel experiment into the desirable difficulties interleaving and self-explanation applied to the higher-order skills required for exam-based courses at secondary school level. In this study with its larger sample, a significant advantage of interleaving over blocking was found. Self-explanation also had a significant effect on performance on the final test, but had the drawback of taking more time than the control condition. As well as being one of the first to use interleaving in a school context, this study is the first to demonstrate that skills-based learning can benefit from interleaving.

7.1.6 Key points from Chapter 6.

Finally, **Chapter 6** outlined two survey-based studies of teachers, each of which attempted to gain an insight into practitioner beliefs about learning and memory. *Study 3* featured a wide-ranging survey about the features of human memory, and found that while the teacher sample were generally more accurate than past surveys of members of the public, errors were shown on the questions on desirable difficulties, and overall accuracy did not increase in line with experience level. *Study 4* used a set of vignettes that portrayed the classroom use of three desirable difficulties in particular: interleaving, spacing, and retrieval practice, presenting participants with a relatively straightforward choice between using these desirable difficulties or avoiding them. In a more nuanced analysis, this study again found no correlation between years of experience in the classroom and a teacher's 'accuracy' on the scenarios as defined by how consistent they were with the research literature. All of the vignettes divided opinion, and it was clear that a large proportion of the teaching profession do not know about interleaving, never mind knowing how to apply it. Awareness of the spacing effect was at a higher level. Perhaps most surprisingly, student teachers in their first few weeks of PGDE study scored higher than practising teachers.

In the following section I attempt to synthesise these various findings and to draw out the main implications that arise. I then evaluate the methodology used (section 7.3), and address some of the major flaws and limitations in the evidence that I have presented (sections 7.4 - 7.5). Finally, I will make recommendations both for future research and for applications to practice (section 7.6).

7.2 Analysis of Findings

Interleaving sits together with other related techniques as both a 'desirable difficulty', and, more specifically, a technique that allows educators to modify a student's experience simply through changing the order in which examples or tasks are studied. It links to and can at times be combined with other desirable difficulties such as retrieval practice, self-explanation, and particularly the spacing effect.

It is apparent from the findings presented in this thesis that interleaving is a valuable technique which has considerable potential for application to classroom learning and to professional learning as well. A new but relatively consistent evidence base exists, and although the prior studies are largely lab-based, they feature an impressive range of classroom-relevant materials, ranging from art (e.g. the artwork used by Kornell and Bjork, 2008) to maths and science (e.g. the statistics example of Noh et al., 2016). Spacing, too, applies very broadly. The reason for this breadth of application could link to the underpinning processes of interleaving and spacing (contrast and forgetting) as well as the biological mechanisms also being very general.

Despite the breadth and potential utility of interleaving, it is also apparent that the technique is little known and poorly-understood. The evidence from Study 4 shows that most teachers report little or no prior knowledge about interleaving, and the majority do not appear to make a good intuitive judgement about the benefits of mixing examples together. Although the research literature is quite consistent in that it defines interleaving as the mixing of short items or pictures, the educational discourse around the technique often mistakenly interprets it as involving the interleaving of long sections of work, lessons or entire topics (see Chapter 3; Chapter 4). There is also confusion over the difference between interleaving new learning (as exemplified by most of the literature reviewed in Chapter 3) and studies where interleaving is applied only to practice or exam revision (such as in the studies of mathematics by Rohrer and colleagues).

In addition, there is a growing body of evidence which suggests that learners misjudge the value of the technique even when they are given a chance to try it out, and in this respect it concords with other techniques such as spacing which also lead to flawed metacognitive judgements (e.g. Study 4, as well as the work of Yan et al., 2017; see also Chapter 6). Part of the reason for this is that as desirable difficulties, these techniques feel harder over the short term, and because learners and teachers are typically unaware that memory is malleable and slow to develop (study 3; Boser, 2019; Loftus, 2019; Simons & Chabris, 2011, 2012; Soderstrom & Bjork, 2015), they assume that techniques that increase effort and errors are best avoided.

It is therefore reasonable to suppose that teachers will often eschew interleaved learning on the basis that their students find it more difficult to study mixed examples compared to a neatly categorised set, although more research is needed into the specific question of how teachers would present examples spontaneously in classroom situations. It is likely in any case that most teachers will tend to cover a specific sub-topic within a lesson rather than mixing several topics together. As such, interleaved contrasts are unlikely to arise unless a teacher specifically chooses to refer back to prior topics (and such a review is likely to be passive if teacher-led, and will also be challenging for learners due to the rapid progression of forgetting).

However, despite the strength of the evidence base, it is not enough to simply advise teachers to interleave more, for there are important considerations about how the concept is operationalised and implemented as discussed in Chapters 3 and 5.

Perhaps most obviously, the degree to which interleaved examples differ from one another makes a difference to how effective the strategy is (Brunmair & Richter, 2019; Carvalho & Goldstone, 2015b; Rohrer, 2012). When these differences are subtle, the side-by-side nature of interleaving helps to make differences more salient than they would otherwise be; classroom examples of this phenomenon might include small but meaningful differences between two mathematical procedures or between two chemical molecules. However, this interleaving advantage disappears and can even reverse when the between-category differences are major (such that no learner would mix up members of the different categories). For example, there is little point in a biology teacher interleaving examples of plants and animals, as no beginner student, even a very young child, is likely to mistake a plant for an animal, even if examples of these life forms are given in separate lessons. The similarity of the within-category difference also plays a role; if the category is diverse (e.g. mammals), then blocking (i.e. showing example mammals together rather than interleaving them with amphibians, birds and reptiles) may help learners to perceive and remember similarities (Carvalho & Goldstone, 2015b; MacKendrick, 2015), and thus to correctly allocate new items to the correct category – i.e. to transfer their learning.

This issue is further complicated by the nature of the learners themselves and their prior knowledge. For, as raised in Chapter 1, a student on a course does not approach new learning like the experimental subject who is trying to memorise a set of trivia facts. Instead, they use their existing knowledge to connect what they see to what they already know, building on their existing schemas and modifying those schemas in such cases as the new information seems to warrant such a re-organisation.

We therefore cannot see interleaving as inherently beneficial to all (a flawed, universalist approach as discussed in Chapter 4). To continue the mammals example used above, a learner's current knowledge of classes of species would make a fundamental difference to the benefits (or lack thereof) of interleaving (see also Chapter 3). In a similar way – as noted in Chapter 2 – the timing of spacing is a highly context-specific judgement which depends on prior learning and the level of understanding of the individual student. This is a judgement that teachers are in the best position to make.

Overall, then, the order in which examples or tasks are presented to students is a professional choice on the part of the teacher. This professional choice will be guided by three major factors: a general understanding of the theoretical concepts underlying learning (in this case, interleaving), a technical proficiency for how to apply those concepts to the classroom situation (that is, applying theory to practice), and a knowledge of the learners in their own classrooms (or, failing that, of typical learners who they are assumed to resemble, e.g. in the case of a brand new class).

The practitioner themselves is most likely to be the locus of decision making in the classroom (Hunter, 1979), and are well placed to have the kind of deep knowledge of the curriculum that facilitates an application of theory to practice (e.g. knowing what kinds of examples learners will respond well to). It is also highly likely that teachers have a good knowledge of the learners in their classrooms, and that this knowledge develops in line with their years of experience (a Psychology teacher with 10 years of teaching experience, for example, should be able to predict the areas of difficulty that a class will face when studying a popular topic, even if he/she is new to the specific class). If they have spent a lot of time working with a particular group of learners, then this situational experience will be better still.

However, Studies 3 and 4 in the present thesis show that many of the benefits of experience fall down when it comes to desirable difficulties; professional understanding of these phenomena do not appear to develop spontaneously. This has been analysed in terms of the new theory of disuse, and the related contrast between performance and learning. Students and teachers alike, it appears, tend to mistake relatively temporary gains for long-term learning, and to underestimate forgetting. They may see the outcome of brief, formative tests, quizzes and even self-assessments as indicative of learning, (thus mistakenly seeing memory as amenable to scrutiny by its users, as discussed in Chapter 4). While teachers are apparently well aware of the importance of developing a meaningful understanding among their students (see study 3), they appear to avoid techniques that introduce short-term difficulties, and are in most cases unaware of these techniques.

A possible solution evident in the policies of some countries (for example England, where the Department for Education's 'Early Career Framework' is infused with concepts from cognitive science; see Scutt, 2020) would be for evidence-based practice to be mandated for all. Such policies have the advantage that they are universal, and circumvent potentially flawed decision making in the classroom. Rather like evidence-based medicine, an argument can be made that if something is effective, then every practitioner should be doing it. Allowing teachers free reign to guide their practice on the basis of their own beliefs and hunches is akin to allowing doctors to recommend homeopathy over conventional medicine.

On the other hand, teachers know their content, and they know their own learning situations. Unlike external policy-makers, they are in a position to make on the spot judgements of what needs to be done, and how, and when. They can use evidence-informed techniques not just to guide planning and material choice, but also to remediate and re-teach where necessary. Policy-makers, on the other hand, can only ever make generalisations. In addition, the 'what works' approach to evidence lacks nuance, fails to account for the negative side effects of applying evidence, and does not fully take student individual differences into account (see Chapter 2).

If teachers' understanding of learning and memory was improved, therefore, they may be in a better position to both apply techniques such as desirable difficulties and to do so in a way that is responsive to their specific learners (Firth, 2017) – at least if evidence from other populations such as students and legal professionals can provide a guide to the possible effects of knowledge of memory on later action. They would also be able to enact techniques immediately and at times responsively when faced with learner misunderstandings. However, an argument against empowering teachers in this way comes from the difficulty of asking teachers to take on board complex research evidence on top of their other duties. Such concerns may have influenced academics such as John Hattie to state "the whole research side, leave that to the academics" (Stewart, 2015) and Dylan Wiliam (2019) to argue that classrooms are "too complicated" for teaching to ever be a fully evidence-based profession.[30]

Other researchers have been similarly negative about the capacity of the teaching profession to successfully engage with and apply evidence from research. Willingham (2017), for example, argues that while teachers do need a mental model of the learner, this theory should be greatly a simplified 'modal model' (his suggested model is very similar to the theories of Murdock, 1967, or Atkinson & Shiffrin, 1968), and should stick to basic empirical findings presented in terms of "folk constructs" (p. 171) rather than scientific terminology. In short, he feels that teachers would not understand evidence if they were to encounter it.

However, a counter-argument can be made both logically and empirically. Chapter 1 presents some of the flaws of the modal model – a simple theory that does not account for issues such as timescale or variation in learning[31]. A focus just on findings as Willingham recommends would most likely lead to the misuse of techniques such as spacing and interleaving as discussed above, lacking the nuance and about when and how to use them. What's more, failure to engage with the research evidence and terminology would leave teachers ill-equipped to self-correct their practice if and when generally accepted concepts become discredited (as happened with 'learning styles' in the 1990–2000s – see Chapter 6 – and the perceptual-interference effect in the 2010s – see Chapter 1).

While an argument could be made that Willingham is only talking about beginner teachers in his 2017 article, the evidence presented in Chapter 6 shows that student teachers can be

[30] A full analysis of this claim is beyond the scope of this thesis, but the essential argument appears to be that classrooms are variable and unpredictable, and that therefore evidence can only guide us to what has happened in the past, not what will happen in the future. This argument appears to be founded on a misunderstanding of how scientific theories are developed and applied.

[31] There are many further flaws of the modal model, not least the fact that both working memory and long-term memory are presented as single systems; there is now a consensus that both systems are functionally subdivided. This has important implications if, for example, WM is conceptualised by teachers as a box with limited storage (the metaphor used by the modal model and also by cognitive load theory). In fact, it is better seen as a system which can simultaneously process and combine verbal and visual information (e.g. Baddeley, 2000), meaning that the limits of verbal WM can be extended if teachers support verbal activities with visual scaffolding.

superior not just in their technical recall of research concepts but in their ability to transfer this understanding accurately to real classroom scenarios. In addition, both Study 3 and Study 4 demonstrate that this kind of awareness of the workings of memory does not develop spontaneously through experience, and there is therefore no reason to suppose that an inadequate explanation given to beginner or early-career teachers would spontaneously improve with practice. This leads to the conclusion that gaps in professional knowledge and understanding are likely to persist if they are there at the outset. If teacher decisions are not informed by evidence about learning and memory then they are likely to be informed by hunches and intuition instead (Firth, 2019a), and these are likely to be flawed on the basis of the evidence already presented.

There is, however, a growing movement in favour of teachers engaging with evidence and making evidence-informed decisions part of their everyday practice, as demonstrated by the appearance of numerous guides over the last decade such as the Education Endowment Foundation's (2018) 'Teaching and Learning Toolkit' and the Society for the Teaching of Psychology's report 'Applying science of learning in education' (Benassi et al., 2014), as well as many popular books. This principle is clearly and directly stated in the highly-regarded report of the Deans for Impact (2015) of the University of Austin, Texas: "*The Science of Learning does not encompass everything that new teachers should know or be able to do, but we believe it is part of an important — and evidence-based — core of what educators should know about learning*" (paragraph 4). At a more grassroots level, the organisation researchED (researched.org.uk) has run dozens of teacher-led conferences exploring the application of research to practice. Together with the policy changes mentioned above, these developments reflect a growing consensus that teaching can and should be an evidence-informed practice, with teachers as evidence-informed practitioners (see Chapter 2).

As noted already, teachers may be in a better position than policy-makers to identify ways in which this evidence applies to their particular learning situation, their particular course, and their particular learners. They are therefore in the best position to take a nuanced, learner-specific approach when applying evidence (Firth, 2017; Scutt, 2020), recognising that techniques like interleaving and spacing are helpful in many situations but should be avoided in others (see above, and also Chapter 3).

Teachers can also go beyond interpreting and applying evidence; they may also, at times, become co-producers of evidence. As recommended in Chapter 3, one of the best ways of extending the evidence on interleaving to more classroom-relevant materials and tasks would be for teachers to participate in the research process. Likewise, education and educational psychology researchers could benefit from engaging more with schools when planning their research programmes. By forming productive partnerships, schools and universities can advance knowledge through the investigation of specific classroom situations and problems. A good example of such a partnership in practice is the work of psychologist Pooja Agarwal and teacher Patrice Bain (see Agarwal & Bain, 2019[32]).

A full analysis of how teachers can engage with research evidence and collaborate with university-based staff is beyond the scope of this thesis, though it is worth noting that that Beck et al. (2020) have set out multiple models for such partnerships (see Appendix 15). Doing so may help to conceptualise the issue in a way that is more helpful to the classroom teacher and which for the researcher problematises a research question in a way that is more likely to facilitate broad application and, thereby, greater impact. Three examples of research goals

[32] However, as a school teacher myself, I would caution against a structuring such partnerships into a simplistic division between ‘expert researchers’ on the one hand and ‘teachers’ on the other. This is partly because many researchers teach, and teachers – as already discussed – can engage with and contribute to research. But there are also psychological advantages from a shared collective professional identity rather than one which emphasises group differences, and more shared ownership and agency is likely to arise if teachers are equal partners rather than subjects of the research.

which have been well linked to practice as demonstrated by studies 1 and 2 (albeit imperfectly) include:

- Applying desirable difficulties to authentic school materials, thus easing the translation of research into practice (both Study 1 and (Study 2).

- Linking spacing and interleaving together in a series of classroom tasks that better reflect the normal progression of activities within a school teaching period (Study 1).

- Applying desirable difficulties to the higher-order skills that are valued by examiners and rewarded in the scoring of marking schemes, and which constitute a bridge between the conceptual categories studied in psychology and the transfer-related skills that are of interest to educational researchers (Study 2).

Further general ways that would be valuable to link theory and practice in school-based research investigations of desirable difficulties might include:

- Investigating the link between desirable difficulties and students' workrate or motivation to learn.

Overall, then, interleaving and other techniques that involve the manipulation of item order are potentially powerful and can apply broadly in the classroom, but their use is situation dependent. This is problematic, given that the techniques are desirable difficulties, and tend to be accompanied by misconceptions on the part of students and teachers alike. The best way to address this would not be to mandate them on a policy level or require them as part of a 'what works' strategy, but to help to ensure that teachers have the skills and knowledge to apply the techniques when it would be helpful to do so, and to avoid them when it would not.

7.3 Evaluation of Methodology

In addition to the broader reviews of research literature and the theoretical analysis arising within from it, the primary research methodology used in the current thesis comprised of a systematic review with meta-analysis (Chapter 3) and a series of computer-based field experiments and practitioner surveys (Chapters 5 & 6). It is now appropriate to reflect on the strengths and weaknesses of these methods ahead of making recommendations for future research in this field.

7.3.1 Systematic review.

I decided to carry out a systematic review into interleaving, a widely-cited evidence-based learning strategy. There was not an extant review at the time this was initiated (although one with overlapping aims was later published by Brunmair & Richter, 2019), and the technique had therefore been less fully explored and reviewed than spacing and some other desirable difficulties.

Since the work of Mulrow, Thacker & Pugh (1988), systematic reviews have become increasingly influential first in medical research and more recently in the social sciences and related fields, including education. Standards for methodology and reporting in such reviews have been developed, with the Cochrane Collaboration[33] being a leader in this field. Due to their depth and comprehensiveness, systematic reviews are considered to provide a more objective and thorough overview than other forms of review, given that evidence cannot easily

[33] https://www.cochrane.org/

be missed through bias or carelessness. Consequently, these reviews are often highly influential with policymakers.

On a technical level, systematic reviewing has certain advantages over traditional literature reviews, but it is not without its critics. Because a systematic review has a clear methodology it lends itself to pre-registration and also to replication, helping to confirm the findings and ensure credibility among the field more broadly. However, some critics believe that it is wrong to elevate systematic reviews above traditional narrative reviews. For example, Greenhalgh, Thorne and Malterud (2018) argue that the two types of review serve different purposes. A systematic review can answer a focused question with a lot of data, while a narrative review allows for interpretation and exploration of phenomena, according to their view.

This dichotomy is partly a matter of definition, however, and it is not one that all researchers adhere to. Fyfe et al. (2014), for example, conducted a study they refer to as a 'systematic review' of concreteness fading, but it does not feature a systematic literature search or a meta-analysis.

In addition, the recent interleaving review of Brunmair and Richter (2019) demonstrates some of the difficulties of engaging in statistical analysis without first carrying out a careful narrative analysis of the concepts under review. Their paper did not fully distinguish between different types of stimuli, leading to somewhat mixed results (e.g. a positive effect for images, a weaker effect for maths, and a benefit of 'blocking' in the case of words). It may have been preferable to explore the mechanism of interleaving before carrying out a statistical analysis, as there is otherwise a risk that the studies grouped together do not all reflect the same phenomena.

Therefore, while I broadly agree with the point made by Greenhalgh and colleagues, a systematic review can very well serve both purposes. The example in Chapter 3 both investigates and explores the interleaving phenomenon and its boundary conditions, and –

having established these things via a narrative section – carried out a meta-analysis on the specific subset of studies where doing so made theoretical sense. Carrying out meta-analyses without a broader contextualisation of the issue at hand is problematic; as with any statistics, useful answers only arise if the question is framed appropriately.

A final point that is worth highlighting is that the critics of systematic reviewing have themselves been criticised. In a study of medical articles, Forsyth et al. (2014) found that articles which were critical of systematic reviewing had industry ties more than twice as often as those which supported the use of the method. While these conclusions cannot be securely generalised from medicine to other fields, it is worth considering, in an era of increasing marketisation of education (e.g. see Dovemark et al., 2018), that the systematic review could be seen as a prophylactic against polarised arguments over pedagogy.

7.3.2 Classroom studies of interleaving.

The two classroom-based studies of school pupils reported in Chapter 5 both made use of online protocols. The studies thus resemble laboratory experiments in some respects, but were carried out as field experiments in terms of the setting. This fusion of methodology allows some of the control inherent in a laboratory experiment (for example, randomisation to conditions, measurement of task time), while maintaining the authenticity of the setting. Of course, in doing so it also maintained many of the drawbacks of classroom experimentation, too, for example background noise.

Granted, there are down-sides to using a computer-based presentation; the tasks required access to internet-enabled devices and this is not a norm in all classrooms (though it is arguably becoming one). The experience of reading textual examples on a computer screen is not alien to pupils but it does not represent the bulk of classroom experiences, though again it is becoming more typical for school pupils to spend some or all of a school period on internet-based activities such as online reading and research.

Past literature on interleaving is dominated by replications of the Kornell and Bjork (2008) ‘artworks’ study (see Chapter 4). It is clear that this is a useful paradigm for research purposes, but it is also one with a relatively weak connection to the classroom, given its lack of connection to STEM or social science, the non-core nature of the material covered (obscure modern artists), and the nature of the task itself (school and university students do not tend to be asked to name the artists responsible for previously-unseen paintings). Indeed, it is a task that relates better to informal learning of general knowledge among adults.

This situation leads to an unfortunate bias in the research literature on interleaving. However, a number of the studies reviewed in Chapter 3 of this thesis are much more directly applicable to classroom situations in science or social science. The use of psychiatric case studies by Zulkiply and colleagues (e.g. Zulkiply et al., 2012), images of chemical molecules by Eglington and Kang (2017), statistics examples by Noh et al. (2016) and psychological definitions by Rawson et al. (2015) are notable examples. Replications of those pieces of work, perhaps extending them to related subjects (for example, replicating the Eglington and Kang methodology with biology/anatomy images), would be highly worthwhile.

The present study provides the basis of another paradigm – the use of short textual examples to study learning of the skills of analysis and evaluation. While these skills will inevitably differ somewhat across academic subjects (for example analysis of a text in English may be quite different from analysis of a research study in Psychology), in a broader sense the presence of such higher-order skills are educationally universal, particularly at more advanced stages of academic study.

Overall, then, one of the main strengths of the current work is that it piloted a novel classroom procedure for implementing and testing interleaving and spacing in the context of real school materials, and applied it to skills learning.

7.3.3 Studies of teachers' beliefs about memory.

Studies 3 and 4 also made use of a computer presentation of materials, though the limitations were rather different; some teachers may have carried out the survey in their classrooms, but given the demands of teaching a class, it is likely that many did so during their time off timetable, either at home or (possibly) in staff preparation rooms. The student teacher participants could have carried out the survey in either university buildings or their homes. An advantage of the chosen methodology is its flexibility and the relative ease of accessing they survey. Participants could even fill it out via a tablet or phone if they so chose.

A limitation of Study 3 was the use of rather general statements about memory, Nevertheless, it was novel, and has helped to establish a new area of educational research, and one that complements the existing literature on teacher effectiveness (and studies of memory in other populations). In short, how accurate are teachers' professional beliefs about memory?

Study 4 improved on Study 3 in that it used a series of largely novel vignettes, each of which portrayed a teaching situation and asked participants to make a judgement. Because each vignette portrayed a different teaching subject, it would be reasonable to suggest that participants may have felt distant from the situations, and not have fully identified with them. A solution for future studies would be to either dispense with the references to teaching subjects entirely, or to use the participant's own declared teaching subject and level to select the most relevant vignettes from among a selection.

One broader consideration is whether the survey itself could be briefer; each vignette performed similarly, and there is an argument for attempting a replication with just three vignettes, one per concept (interleaving, spacing and retrieval practice). Doing so would reduce the demand on participants, and open up the methodology to potentially adding additional questions in the future without making the survey overly-long – questions about participants' background training, for example, or about their professional reading, or a measure of their endorsement of neuromyths.

Overall, studies 3 and 4 provided a novel insight into an under-researched area. Study 3 was novel in that it provided a wide-ranging survey into beliefs about memory. To my knowledge, only one other recent survey – a study of American teachers by Boser (2019) which did not appear in a peer-reviewed academic publication – has used vignettes to investigate beliefs about desirable difficulties in learning, and none have investigated student teachers or compared levels of knowledge across the career span, or surveyed a UK-based sample.

On the latter point, both Study 3 and Study 4 are limited in that they recruited samples specific to the Scottish education sector, but nevertheless they provide evidence that prior investigation into beliefs about memory among the general public and among the legal profession may not fully apply to teachers. And the relatively similar findings from Boser (2019) in terms of preference for the three desirable difficulties under investigation (see Chapter 6) suggests that the finding were not strongly biased by the nationality of my participants.

7.3.4 General limitations of the findings.

As many of the points so far in this chapter rest on the findings of empirical sections of the thesis, it is worth taking some time to consider any issues that might affect data validity. Given that I have critiqued the methodology and analysis of the specific studies in the foregoing chapters and sections, I will focus here particularly on external validity; that is to say, I will address the question of whether the current findings can be confidently generalised to other learning settings, and used as a basis for recommendations and for further research – a notoriously problematic process (Boyle, 2012; Coe, 2020; McDaniel et al, 2007).

A first consideration, touched upon in Chapter 1, is whether the methodology used is sufficiently meaningful to be applicable to classroom practice. An issue that has been raised already in this thesis is that many studies of desirable difficulties (e.g. Cepeda et al., 2008; Hausman & Kornell, 2014; Kornell, 2015; Landauer & Bjork, 1978) use stimuli which consist of brief items (lists of words or facts) which are presented out of context. As such, they cannot be considered analogs of classroom learning, during which learners build information actively and cumulatively into meaningful schemas (a possible exception is the practice of revising with flashcards, but this tends to happen after meaningful understanding has already been established).

In the present studies, however, realistic classroom materials were used in their authentic contexts. Study 1 used a full-text psychology research study, widely taught on school courses, and sets of information about phobias which closely resembled what might be seen on classroom slides. Even the materials in study 2, where the examples were presented as individual sentences, connected with a broader meaningful context, in that each example related to a previously-studied mandatory piece of psychology research (e.g. the Milgram obedience study; Milgram, 1963). As such, the learning task allowed learners to develop new conceptual knowledge (of analytical skills) by linking what they saw to prior learning.

All the same, it would be worth extending the findings by using materials in their original context. Rather than extracting example evaluation points from a textbook, for example, the original textbook material could be used in future studies, perhaps with two subtly different versions created which vary the order of examples but look otherwise identical to students.

Related to the point above, I have explored the idea that much of the previous literature aims to provide learners with novel, previously-unseen items to avoid bias from prior knowledge. In studies of desirable difficulties and language learning, for example, Swahili words are often used (e.g. Karpicke & Bauernschmidt, 2011) in preference to more widely-studied languages. Even richer materials such as the artwork used in interleaving studies is selected to be obscure and therefore (it is assumed) novel to most participants. In the present research, I would argue that while the learners did not have much background knowledge in the case of Study 1, and in the case of Study 2 performed at a relatively low level in the pre-test phase, this does not mean that the task content was disconnected from their prior learning. Learners are always novices when they experience new material, but an important issue is whether they have relevant background knowledge and interests. As learners in a psychology lesson, the material will have fit in some way with their prior expectations and understanding, and engaged with their motivation to learn the subject. Some factual details may have differed from their prior assumptions, but it will have done so in ways that were meaningful for them. In contrast to a laboratory task which the learner typically cannot connect to any aspect of their interests or broader educational context, these tasks represented a natural next step to their learning.

It could be argued that the samples used in the metacognitive studies of teacher beliefs were too small and insufficiently diverse; I would acknowledge that this is a limitation, as was the drop-out rate (for example, 84% of those who completed vignette 1 in Study 4 went on to complete all eight of the other vignettes). Besides the implications of this for the statistical analysis as discussed in the previous chapters, it is worth questioning how representative the sample was. In the case of those teachers who dropped out during study 4, did they have certain characteristics that differed from those who completed? It seems likely that if there was a consistent difference, that those with weaker knowledge of the strategies at hand would be more likely to drop out (due to finding the task too hard), and that overall performance would be, if anything, be worse than what was found if they had all completed the survey. The first vignette was the most answered (because most drop outs occurred after this question) and was also the worst answered (with a mean of 2.72), which is consistent with this speculation.

A broader issue is that the entire sample will have been subject to bias by the voluntary, interest-based nature of recruitment, a problem that was particularly the case with study 3 due to its use of social media as part of the participant recruitment strategy. Future studies might look at incentives for participation in order to reduce this problem, and aim to sample as large a proportion of staff as possible from a single school. Again, though, if anything a wider sample would most likely show worse performance and awareness of the techniques under investigation, so it is unlikely that any of the conclusions of study 4 would be invalidated.

A related issue is that the staff members from the same institution do not have statistical independence. This means that findings which draw heavily on a particular school could be biased; it may be the case, for example, that the school in question has an especially effective in-service training programme in place. Future studies may seek to navigate this problem by recruiting comparable departments (e.g. the science department) from multiple schools in different local authorities.

Studies 1 and 2 also have limitations in terms of their samples; in the former case I recruited participants from within a single independent school, and in the latter, just two comprehensive secondary schools. All of the participants in both of these studies had voluntarily chosen to take psychology as a non-mandatory subject in their final year of school, and may have been slightly above average in terms of their academic ability and/or motivation (although perhaps not by much, given that a very large proportion of pupils from each year group take the subject).

However, any field experiment which looks at learning in a classroom is subject to some of the same constraints and biases in the sample.[34] The benefits of classroom-based experimentation (Taber, 2013) tend to be tied to flaws in both sampling (which tends to be opportunistic) and allocation to conditions (which tends to be non-random), and some attempt was made to overcome this where possible. The tasks related to the study of psychology and therefore couldn't be used with a random sample of school pupils, but randomisation to conditions was achieved via the software. In study 2 in particular, some diversity was achieved by recruiting schools from two different local authorities. The surrounding areas were also varied in their socio-economic make-up.[35]

The task itself was rather short-term, and again I acknowledge this limitation. Future studies could build in a longer-term follow up, as Soderstrom and Bjork (2015) have recommended for studies of desirable difficulties. This is important because the dynamics of forgetting are subtle, and at times a condition associated with equivalent or superior performances at the time of study can lead to poorer performance after a delay (for example Bird, 2010; Roediger & Karpicke, 2006). However, it is important to note that previous studies of interleaving have not revealed these same dynamics as have the studies of retrieval practice and spacing that appear in the literature. Three studies described in Chapter 3 which included a delay found no interaction between the effect of interleaving and a delay (it is worth reiterating that a delay or lag between practice sessions would be expected to differ from spacing within the task itself; the latter can interfere with interleaving; Birnbaum et al., 2013).

Overall, the four empirical studies here were all exploratory, aiming to test out a relatively novel set of methodological procedures and to explore known issues in new contexts. While each one had its flaws, these did not exceed what are typically found in studies of these type,

[34] And indeed, much of the laboratory research in this area also features highly-academic students who have chosen to study psychology.
[35] For example, one of the host local authorities has a rate of unemployment of 4.1% and the other 2.7%. These compared to the national average of 3.5% (source: https://www.nomisweb.co.uk/)

and a number of factors support the idea that the overall conclusions – and thus the external validity of the work – remain sound.

While this thesis has described a programme of academic research, I have – as highlighted in Chapters 1 and 2 – been influenced throughout by my perspective as a teacher as well as a researcher. Understanding of learning and memory processes is always the key goal, but there have been times when compromises have been made in the interests of ecological validity or to align with what is practical within a classroom.

The teacher-researcher perspective also requires a close focus on classroom applicability, sometimes at a cost to reliability and standardisation of procedures. For example, there were aspects of the procedure in Study 1 which do not align with common practices in laboratory-based studies of spacing and interleaving, while the nesting of variables in Study 2 was also a limitation. While I acknowledge and take responsibility for these weaknesses, there is also an important place for more teachers to engage in and with research (Firth, 2019a), as well as to form research partnership with HE institutions (see Appendix 15). Any such partnerships will need to navigate conflicting priorities. Lowering ecological validity, increasing stress on classroom students or switching to more artificial stimuli in the interests of control are unlikely to be acceptable compromises for most teacher-researchers.

We may instead need a triangulation of evidence – drawing on theory (such as that reviewed in Chapter 2), on laboratory studies (such as those reviewed in Chapter 3), and on field research (such as the studies reported in Chapter 5). Each of these sources of information on its own is likely to be imperfect when used to generate recommendations for the classroom. Combined, however – as I have attempted to do in this thesis – these sources can begin to provide genuine insights into psychological processes and the way that they manifest in the classroom, and can point the way towards implications for effective instruction.

7.4 Priorities for Further Research

Having considered some of the methodological flaws in the present research, it is now appropriate to consider the areas where my findings can be extended and applied. I now therefore outline what I see as the priorities for further investigation – the logical and practical next steps for exploring the interleaving effect, and for in linking it with the broader evidence on desirable difficulties – as well as priorities for further research into the professional learning implications of these phenomena.

Research into interleaving appears to be on the increase; consider that out of the 26 studies reviewed in Chapter 3, six were published in the five years from 2008–2012, and twenty were published in the five years from 2013–17. The awareness of the technique among teachers more broadly is low (see Study 4) but it may also be growing, given that it is now referred to in some popular books about teaching. It would be useful for future studies to follow up on the current findings by tracking this professional awareness – how many teachers have heard of interleaving and other desirable techniques, and what do they think these techniques involve? It could well be the case that awareness is increasing year on year, but it is also possible that sharing of the evidence is restricted to a particular subset of teachers, and the information is not reaching others outside of this circle.

With some of the key practical effects and interactions associated with interleaving now well established via laboratory investigation, attention is increasingly turning to classroom application of the technique. This thesis summarises a contribution to this effort, but there are still major gaps in our knowledge. With regard to the way interleaving can and should be used in the classroom, I recognise that the studies included here provide only a beginning. One aspect that should continue and be extended is the use of more curriculum-relevant study materials.

The use of computer-based presentations had certain advantages as discussed above, and could be extended, moving from software that is designed for psychology experiments to more authentic screen-based contexts. For example, short learning videos could be designed in such a way that one version presented interleaved examples and another version presented the same examples in a blocked fashion. Online learning environments are becoming ubiquitous, and could be used to host quizzes and tasks that were later analysed as a source of data, improving the validity of the sample in a study (though also presenting some ethical issues; this idea may work best for practitioner enquiry projects). A more innovative option would be to make use of computer games such as Minecraft (as discussed in Chapter 4); the images and environments in such games could be interleaved or spaced to model such experiences as learning on field trips. Such techniques could also help when it comes to experimentation with younger learners who are not taking examined courses with prescribed factual material and skills.

The present research was innovative in that investigated the potential of interleaving for skill-based learning, and future studies could follow this up by investigating the same skills in different topics. Although these are often seen as the same 'transferable' skill (e.g. 'analysis') as far as exam boards are concerned, the reality of what students are doing in the classroom may be quite different across disciplines, as mentioned above. It is unlikely that this skill transfers especially well from one subject to another, but the potential of interleaving to help with this transfer could be investigated, and its potential use in *meta-transfer* (that is, the ability to learn how to transfer a strategy) could be revisited, following the exploratory study by Birnbaum (2013).

The potential of interleaving to assist with the literacy skills could also be assessed; on a theoretical level it would appear to lend itself well to areas such as initial learning of reading (interleaved rather than blocked examples of words or letter sounds), the learning of such features of written language as genre or figures of speech (via interleaved rather than blocked example texts) or speaking skills. Such research would considerably extend the current body of research with its predominance of experiments testing older students.

It is noted in Chapter 5 that the interventions used in my classroom work on interleaving (studies 1 and 2) was very brief – typically lasting less than 30 minutes – and the criterial test occurred within the same school hour. A useful follow up, and one which would further increase the external validity with regard to classroom recommendations, would be to use at least one delayed follow up test, or to use naturally-generated evidence such as progress tests or prelims to determine whether interleaving has a sustained benefit. Both practically and ethically it would be difficult to use an experimental methodology where a class served as a control group, but this problem could be circumvented by using historical school data for comparison (Firth, 2019a). For example, a teacher could compare the performance of their class in the present year with previous cohort(s) used as a baseline, perhaps with some allowance made for potential differences between the different groups by scrutinising other relevant data such as pupils' prior attainment. Alternatively, teachers/researchers could interleave examples in one topic and block them in another topic on the same course, counterbalancing allocation to conditions across two comparable classes.

Were such studies to be done with an exam class, it would be possible for this to work around existing schedules; most teachers run occasional short review tests, and most schools run prelim exams at around the turn of the calendar year. Such scheduling tends not to be flexible, so any initial interleaving intervention should be scheduled to begin a number of weeks ahead of such assessments. Planning must be meticulous when timing is so inflexible; as was the case in the current study, the process of gaining ethical permission via the university and then local authorities can lead to some delay, with the potential to push the entire data gathering process into the following year if permissions are not in place soon enough.

On a related point, future work could follow learners over a longer period to see how desirable difficulties play out across the learning of different parts of a topic. One way to do this would be for a researcher to spend some time in a single school, guiding the choice and timing of materials used. However, such options are likely to face some practical difficulties given the need to 'borrow' time from a school class. This is especially the case when the participants are senior pupils, working towards exams in a scenario where the amount of contact time with their teacher is constrained by the school year (the bulk of the course content for Higher and N5 Psychology, for example, is usually covered between September and January).

A potential solution to this methodological problem would be the use of 'practitioner enquiry' approaches, whereby a teacher investigates their own practice. Practitioner enquiry suffers from some limitations, not least the lack of ethical oversight (and groups should not be disadvantaged by being a control group, for example). The practitioner approach tends to still suffer from a small number of participants in a unique context, making any findings hard to generalise and from a statistical point of view, causing the conclusions to be underpowered. But this issue could be overcome via collaborative practitioner enquiry, with multiple teachers and schools and perhaps HE partners as well. In light of some of these issues, I have written a book which guides teachers about how to find, engage with, and apply research in the classroom in a way that is ethical and valid (Firth, 2019a; see Appendix 16 for an outline and introduction to the book).

It is also important to consider how interleaved tasks fit within a broader set of realistic classroom activities, and the present research included some initial attempts to combine desirable difficulties such as spacing and self-explanation with interleaving. Study 1 found evidence that the placement of an interleaved task in a lesson could interact with other, ostensibly separate tasks that come before and after it; more broadly, attentional processes vary across the course of a one-hour period (Risko et al., 2012), as well as across the school day (e.g. see Valdez et al., 2014).

Future studies could go further in to combine other desirable difficulties such as retrieval practice with interleaving. It would also be helpful if such techniques were used in tasks that fall within the same lesson, to further investigate whether the benefits of a technique on one task could have knock-on benefits or costs elsewhere in a lesson.

Chapter 3 raised a number of methodological issues that apply to the study of human memory and learning, and these should be taken into account when planning future investigations. In particular, learners often lack insight into their own learning process, and underestimate forgetting. As such, it is useful to develop a metacognitive awareness of how learning works among students (Pintrich, 2002). One way of doing this would be to design a course specifically on ‘how to learn’ and make this available to school students. In addition to its utility as a study guide, such a course could build in research into certain aspects of desirable difficulties, for example by surveying student beliefs as they work through the course, and testing their recall and forgetting of material from practice tasks (to which spacing or interleaving could be applied). Advantages of doing this as a computer-based course would include the potential to monitor and analyse task timings and login dates.

Besides the use of interleaving and other techniques to improve concept learning, the current research has tackled some of the psychological and belief-based barriers to the use of interleaving and other desirable difficulties among teachers, and these findings, too, could be built upon.

One of the strengths of the present series of studies was to develop and trial two survey-based instruments. The use of these could be extended; in particular, the vignettes used in Study 4 could be tried with on different populations to better establish what good or poor performance on the task would constitute. The same tasks could also be tested on student teachers in other institutions and in other countries. As an online survey it is relatively easy to distribute, and international comparisons would provide some insight into the pros and cons of different teacher preparation systems (as has been attempted by Betts, et al., 2019, for higher education; see also Howard-Jones, 2014, with respect to neuromyths among school teachers). Some of the questions from study 3 could be combined, potentially allowing factual knowledge about memory to be compared with skill at judging specific classroom situations.

Of course, a perennial problem with the validity of surveys is the question of whether real behaviour will match survey responses. For this reason, it would be worthwhile to test the validity of survey responses via a case study which followed up certain respondents and looked at their classroom practice and planning process to see whether these reflected their survey responses. In-depth interviews or even focus groups could also be used, allowing the researcher to find out more about the thinking behind the choices that participants make on the Likert scale, and to investigate how they conceptualise successfully learning.

A broader metacognitive point, briefly noted earlier, is the barrier caused by specific misconceptions about the nature of interleaving itself. Even sources which recommend interleaving as an evidence-based strategy either misunderstand it completely or recommend it in ways which are far from clear. For example, sources often refer to interleaving of 'topics' or of interleaving entire lessons rather than of specific textual examples, images, or questions as typically used in the literature (see Chapter 3; Chapter 5). This is also worthy of further investigation. One way to do this would be to conduct a literature review which focused not on the academic papers but on professional texts that refer to interleaving and are aimed at educators – government policy documents, articles in the professional press, and books about teaching, for example.

More broadly, it emerged that interleaving itself is not always clearly distinguished from spacing, even in the research literature. This is in part due to the fact that early work by Kornell and colleagues attributed the benefit of interleaved schedules to the increased delays between items from the same category (e.g. Kornell & Bjork, 2008; Hausman & Kornell, 2014). A further area of confusion is the distinction between interleaved learning and interleaved practice. The latter is primarily a revision technique, and has been successfully applied to maths, as discussed in Chapter 3. Even in Study 2 in the present thesis, there was an element of revision, though the methodology generally reflected studies of initial learning. Future research could explore the difference between these concepts theoretically, and also make direct empirical comparisons between interleaving at different stages of the learning process. The small interleaving effect for maths revision found in the review by Brunmair and Richter supports the idea that interleaving may be best utilised early; similarly, experimental evidence from Yan et al. (2017) found no support for the idea that moving gradually from a blocked to an interleaved schedule was beneficial.

A more general but related issue is the conceptualisation of desirable difficulties as a whole. As noted in Chapter 1, this can be misconstrued as a generalisation that harder tasks are always better. Actually – in line with the transfer-appropriate processing principle – difficulty is only better if it prompts processing which reflects the situation in which learned information will later be used (Bjork & Bjork, 2019), and can otherwise be worse for learning (e.g. Taylor et al., 2020). If this misinterpretation were to take hold among the teaching profession, there is a risk that they would aim to make interleaving as difficult as possible rather than focusing their efforts on interleaving easily-confused examples that would boost discriminative contrast. This presents a barrier to effective practice that is worthy of further investigation; interviews and focus groups could be used to explore teachers' conceptualisation of 'difficulty' and when it might be beneficial.

7.5 Recommendations

This thesis has discussed the practical implementation of interleaving and other evidence-based strategies to learning. The most obvious applications of the work, then, lies in informing educational practice and/or policy. But what exactly are these applications, and how could they be implemented in a way that takes account of the inherent complexity and counterintuitive nature of the phenomena investigated? I will address this question now, and in doing so attempt to indicate how best to overcome the major barriers to implementation.

So far in the thesis I have proceeded in the following order – exploring the nature of interleaving and related techniques first, and then considering teacher beliefs about these techniques afterwards, and I will continue with the same order here (rather than interleaving it). I will therefore begin with recommendations for the classroom (7.5.1) followed by recommendations that aim to apply the evidence to informal learning situations (7.5.2). I will then provide a series of recommendations and considerations for professional learning (section 7.5.3),

7.5.1 Recommendations for formal learning settings.

As I discussed at the outset of this thesis, desirable difficulties have considerable potential for application to the classroom. They are evidence-based strategies that appear to boost long-term learning and transfer (see Chapters 1–2). Interleaving, spacing and other techniques in this category have been described as "low hanging fruit" (Roediger & Pyc, 2012, p. 242) on the basis that they are easy and low-cost to apply. I also identified evidence at the outset which suggested that they are nevertheless little known and used among the profession (Dempster, 1988; Roediger & Pyc, 2012), amid a broader literature on beliefs about memory and learning which suggests that people often misjudge the workings of memory and that teachers commonly hold flawed beliefs and subscribe to neuromyths (see Chapter 6).

The findings of this study largely support those initial speculations. I identified that interleaving was worthy of more detailed study, given that compared to the spacing effect it has a new and relatively unclear evidence based, but yet has a considerable potential for concept learning and transfer. Much of the past research on these techniques has been conducted on samples of university students (see Chapter 3; Cepeda et al., 2006) but interleaving has been demonstrated with very young children, too (Vlach et al., 2014) as well as with older adults (Kornell et al., 2010) suggesting that the techniques have potential to be used at any level. The effect sizes found in laboratory work are large enough to suggest considerable potential for the classroom and beyond. And while learners' level of prior knowledge is likely to be important, the techniques are not inherently dependent on adult-level cognitive functioning.

Some specific recommendations for practice are as follows.

7.5.1.1 Classroom presentations.

Teachers verbally provide explanations and examples on a regular basis, either as part of a planned presentation of new material (for brevity, a ‘lecture’) or more ad-hoc remedial explanations when learners appear to be struggling (for brevity, an ‘explanation’). In either case, it is possible – indeed likely – that teachers are giving more than one verbal example, and the order of these examples could be considered in light of the evidence on interleaving. Lectures could be planned to contrast concepts rather than presenting them one at a time in chunks (as may be typical). While teachers may be concerned that this would lead to lengthier explanations and that student attention would wander, evidence from Study 1 and related work (see 5.2.4) suggests that students may see such situations as problems to be solved, and process the information more actively and more deeply as they try to work out the connections between items. It may not be necessary or even desirable for teachers to point out the links; work by Eglington and Kang (2017) suggests that drawing attention to differences has no effect, and learners may find it motivating and memorable to make these links for themselves. A real-world analogue to this could involve reading a mystery story, where plot lines, clues and characters are introduced without full explanation, leaving readers to speculate on how they are connected and actively attempt to work this out.

When preparing materials for lessons, then, teachers should be mindful that the order and timing of examples can play a role in learning and memory, as well as in students’ later ability to transfer their learning. Teachers should aim to juxtapose subtly-different items or skills in order to aid discrimination and promote learner attention to the differences. Such interleaving of examples may occur more naturally in the absence of clear rules to follow (see Noh et al., 2016), and as Schalk et al. (2018) recommend, there can be a place for problem-solving-followed-by-instruction rather than a more traditional tell-and-practice lesson sequence.

7.5.1.2 Materials and course design.

Materials can also be designed in a way that takes account of interleaving. Rather than separating sub-topics onto different sections and pages, the content can be mixed together such that easily-confused concepts appear side by side. Again, though, there is no advantage in mixing items that are clearly distinct. Practice activities should have their order mixed or shuffled as happens in exam practice, and the ability to discriminate between easily-confused concepts and examples should be a directly practised. Delays can also be built into the practice process such that consolidation is delayed by a few days or even weeks after an initial (effective) teaching session. There is no need, however, to interleave longer sections of lessons or entire topics. Interleaving works at an item level, by modifying attentional processes and promoting contrast. These processes can't take place if the interleaved items occur in different study sessions.

7.5.1.3 Independent study and revision.

Learners' assumptions about what helps to make information stick in memory tend to be inaccurate (Schwartz & Efklides, 2012), and they are more likely to make use of ineffective strategies such as re-reading and highlighting than effective ones such as self-testing and interleaving (Hartwig & Dunlosky, 2012). Teachers are in the best position to advise on more effective independent study and revision; indeed, given the large-scale misconceptions about learning which most learners retain through to university level (see Chapter 6), it could be seen as a moral imperative for better guidance to be given. But for teachers to be the experts on learning, guidance must be evidence-based rather than derived from experience or intuition

(Firth, 2017; see Chapter 6). In terms of interleaving specifically, learners may choose to strictly categorise their learning by topic rather than seeking out comparisons and contrasts, and thus fail to recognise subtle areas of potential confusion or meaningful cross-topic links. One way that teachers could assist with this would be to spend time practising revision and study habits in the class, easing the transfer process to private study (Riazat & Firth, 2020). The shuffling of items within a topic could be encouraged during such sessions, as could active retrieval practice of terms and concepts. In terms of spacing, shorter learning sessions involving active tasks are also likely to be more effective than more lengthy sessions during which attention levels are likely to dip. However, short sessions may detract from the benefit of interleaving. It is therefore important that individual items which benefit from discriminative contrast are covered within the same session, and learners could be educated to appreciate this principle.

7.5.2 Recommendations for informal learning

Outside of formal education, workplace training typically faces the same problems as those discussed throughout this thesis (e.g. see Chapter 1); learning needs to last, and it needs to transfer. Safety briefings for those who operate machinery, for example, are going to have more impact if the workers remember the information (so the learning is lasting) and are able to put what they have learned into practice in the workplace (so the learning is transferable). Indeed, this kind of scenario would be an ideal locus for applying interleaving, given its role in discriminability and contrast of subtly different stimuli, its mnemonic benefit, and evidence of its efficacy with image-based material. If a workplace were to interleave images of safe vs. unsafe ways of setting up or using machinery, for example, it would be reasonable to predict that this would lead to more effective learning than would blocking of the same images.

More broadly, for workers to better remember new information and concepts, desirable difficulties can be brought to bear. Information should be spaced out over time (especially if it is important that learners retain it for a long duration; Cepeda et al., 2008), and retrieval practice can help to ensure factual retention. Variation of the learning context (for example, studying in several places or with different groups of colleagues) can boost transfer, and challenging aspects of learning such as note-taking and self-explanation can help to cement understanding by increasing reflection and focussing attention on the key content for longer and in a more deeply meaningful way. In an informal, professional learning context, the time required for self-explanation may be less of a barrier than is the case in the school classroom.

Many of these points apply to the professional learning of *teachers*; this is considered next.

7.5.3 Recommendations for professional learning.

The point has already been made that a deeper understanding of memory and learning would be beneficial for the teaching profession (see Chapter 2). A barrier to the implementation of interleaving and other evidence-based teaching strategies is the flawed beliefs which characterise teacher and student metacognition, and which formed the focus of Studies 3 and 4. A straightforward recommendation, then, is that teachers should have more knowledge about desirable difficulties. They should know what interleaving and spacing are, and have a clear understanding of how to apply them, in line with the recommendations in the previous section.

However, embedding an understanding of memory in the teaching profession is challenging. If they do not already know about concepts such as desirable difficulty then they may suffer from the Dunning-Kruger effect (Dunning, 2011) – the techniques under discussion are ‘unknown unknowns’; teachers cannot and will not strive to improve their professional practice if they don’t realise that they are ignorant of these issues.

One starting recommendation, then, is to make information about these techniques more widely available to practitioners. An example of this recommendation in action is a recent short article that I wrote for teachers of psychology outlining how the techniques can be applied to popular topics in the subject, as well as a chapter on how to apply desirable difficulties to mentoring new science teachers (see Appendices 17–18). More generically, books such as ‘Psychology in the classroom’ (Smith & Firth, 2018; see Appendix 19 for the introduction to this book setting out its aims) help to make the concepts more accessible to teachers while maintaining technical accuracy and not losing the theoretical underpinnings. Other accessible books on evidence-based practice include ‘Powerful teaching’ by Agarwal and Bain (2019); ‘Make it stick: The science of successful learning’ by Brown et al. (2014); and ‘Understanding how we learn: A visual guide’ by Weinstein et al. (2018). As can be seen from the dates of these publications, the past few years has seen numerous useful titles released; relevant information is clearly available.

However, information alone is not always enough to have an impact on classroom practice. As discussed above, it is possible for teachers to increase in knowledge with little impact on their practice. A possible way forward is indicated by an experiment by Yan, Bjork and Bjork (2016). In this study, participants were given information about interleaving, or given a chance to try out interleaved learning experiences, or both. Only in the condition that featured a combination of practical experience and theoretical understanding did practice change, and even then, only when the two schedules (interleaved and blocked) were made clearly distinct. This study lends empirical support to some of the ideas discussed in Chapter 2 (and see also below) – teachers need an extended opportunity to put theoretical ideas into practice, and one that includes practical experience.[36]

[36] Indeed, this is the rationale behind the extended training and school ‘practicum’ used by most countries within their teacher preparation process.

There are also concerns over the sources of information available to teachers. Given that teachers need to make minute-by-minute decisions which are tailored to their specific classroom situations, I earlier (see section 7.2) rejected the idea that evidence-informed practice should be a matter just for researchers and policy makers, arguing instead that teachers are in the best position to make informed decisions. However, teachers (beyond their student year) typically lack access to up-to-date academic journals, and this contrasts with practitioners in other professions which are expected to be evidence-based in their practice – doctors, for example, who have full access to academic journals via the NHS (Firth, 2017). A simple recommendation to address this and minimise the power of gatekeepers to research access would be to make online mainstream scientific journals free to access for teachers, at least those in the public sector.

While it can easily be argued that teachers should engage with (or would benefit from engagement with) the evidence on memory, it is also important to think about *how* to achieve this in terms of efficacy and engagement. In terms of efficacy, it has been recognised that methods of in-service teacher professional learning are often flawed due to relying on brief sessions without the opportunity for consolidation or practice (Donaldson, 2010; see Chapter 2). And while introducing evidence at the initial teacher education stage has its advantages, it is a limited strategy both due to the propensity for forgetting by the individual practitioners, and because the evidence itself is dynamic and moves on over time[37]. A complete strategy for embedding evidence into the profession must include ongoing learning (or, as the General Teaching Council for Scotland call it, 'Professional Update'; GTCS, n.d.).

Many of desirable difficulties discussed earlier could help to tackle this problem, and as such, they are just as appropriate for professional learning as they are for delivery of the school

[37] A comparison can be made here with the work of Hans Rosling, who found that the world knowledge of many highly-educated people was at little better than guesswork, simply due to their ideas being out of date (BBC, 2013).

curriculum (in some cases, perhaps more so). A reflexive use of the ideas, then, can be achieved by considering desirable difficulties as means of guiding effective professional learning, in addition to their utility as classroom techniques.

Several desirable difficulties have the potential to be applied to teacher professional learning in order to make it more effective. These and other promising strategies are discussed next.

7.5.3.1 Interleaving professional learning.

Interleaved presentations provide opportunities for learners to compare and contrast concepts, thus allowing more accurate categorisation and the development of new categories. An implication for professional learning is that examples may need to be less structured. Just as information does not come to us neatly categorised in the real world or the classroom, so the professional learning process may benefit from the desirable difficulty of increased variability.

In this context, interleaving would be best used to illustrate the differences between professional concepts, potentially helping practitioners to both remember these concepts and apply them to novel contexts. The evidence explored in Chapter 3 suggests that this strategy will be best applied to concepts which are easily confused – two subtly-different methods of classroom management, for example, or several methods of questioning. The research into interleaving tends to feature relatively short timescales – brief study sessions where examples are directly contrasted. One practical example of how this could occur in practice would be through brief, Pecha Kucha-style presentations of classroom learning examples by colleagues. Short videos of teaching techniques in action could also be interleaved during staff professional learning sessions in order to illustrate confusable points. This would be a better way of promoting teachers' ability to compare and contrast than can be found in more occasional and lengthy lesson observations.

Interleaving could also be applied to the learning of skills needed for practitioner research. Study 2 established that interleaving is in principle an appropriate desirable difficulty for skills learning; the skills associated with practitioner research are often new to school teachers (or at least, very much in need of practice) when they find themselves undertaking projects. Interleaved examples of such things as the use of statistical tests (as done by Noh et al., 2016), types of sampling, and ethical procedures are potential area where the manipulation of item order could be applied. Again, though, it should be remembered that if interleaving is used alongside spacing, delays should be scheduled to avoid interfering with discriminative contrast (Birnbaum et al., 2013).

7.5.3.2 Spacing professional learning.

Professional learning is more effective when it is sustained, while lessons from one-off CPD events tend to be rapidly forgotten upon the return to the classroom (Donaldson, 2011). This concords with the spacing effect; massed practice is less durable, leading to forgetting which is both rapid and severe. In addition, extended professional learning provides more opportunities to integrate new ideas into classroom practice (Darling-Hammond, Hyler, & Gardner, 2017), thus supporting the process of transfer and providing opportunities for varied retrieval.

One model of spaced-out professional learning is to engage in practitioner research projects, as such a project is inherently spaced out over time. If well-conceived, a teacher research project may also engage curiosity, and benefit from the motivational factors described by self-determination theory (Ryan & Deci, 2017) – autonomy, socially connectedness, and a sense of competence. The project may also result in some incidental interleaving of concepts.

7.5.3.3 Other desirable difficulties in professional learning.

Other desirable difficulties are relevant here too. Opportunities for retrieval practice occur during professional discussions. This could be promoted by implementing a staff journal reading club (as recommended by Bennett, 2017), and/or organising professional learning sessions to promote discussion, especially discussion which involves retrieval (and thereby consolidation) of past learning.

Professional blogging could also be used; it would allow for both retrieval practice and self-explanation. Depending on the task, it could also promote summarising – another desirable difficulty (see Chapter 1) – as could giving teachers time to complete independent written summaries at the end of a session during the period typically allocated to evaluations.

Finally, the desirable difficulty of variation can be used. Varied tasks lead to poorer performance but better transfer, and this could be utilised by instigating professional learning sessions where participants try out skills in multiple contexts – with different age groups, perhaps, or even in different academic subjects. To be beneficial, however, the variation would have to be relevant to the processing involved in learning, and should relate to the kind of variety that may be found in future classroom situations.

7.5.3.4 Leveraging identity to motivate professional change.

Another barrier in applying the broader findings from this thesis lies in the time available for teachers to both engage with the evidence and to apply it to their practice. On the assumption that most practitioners want to teach effectively, it is important to ask what barriers would interfere with their engaging with (i.e. finding, reading, interpreting, and using) evidence that has the potential to improve their practice. There is a growing body of evidence developing which suggests that available time is a major barrier to such engagement (e.g. Lowden et al., 2019).

Professional learning sessions as discussed above are often school-led, but teachers will benefit more fully if they prioritise reading and applying evidence to their practice more independently as well. However, independent engagement in evidence-based practice is in part a matter of identity. Teachers are unlikely to prioritise the time required to read and analyse research papers or to attend evidence-focused practitioner conferences and talks if they don't see this as part of their role.

Even the most motivated and committed of teachers may feel that research is just 'not for them'; not only it is not seen as part of their specific remit, it does not form part of their professional identity (Beck et al, in preparation, includes a more detailed version of this analysis based around interview evidence; see Appendix 20). And this professional identity will affect how they allocate preparation time. Freeing up more time for teachers is important, but they are unlikely to use any such time to engage with evidence on desirable difficulties if they don't see themselves as research-engaged practitioners.

Identity can thus be seen as a barrier when it comes to teachers' choosing to engage with evidence (or not). Without buy-in, then there is little point in exhorting teachers to improve their understanding. Identity can be difficult to modify, and can serve to support flawed beliefs about learning (see Chapter 2). However, it is malleable, as studies that aim to modify school culture among students have demonstrated (e.g. Reynolds et al., 2015).

One way to address these issues is to use the tools of choice architecture – 'nudges' – to prompt research engagement. In essence, a nudge is a small change in the structure of an environment which facilitates a behaviour, for example by making it easier, more attractive, more desirable, or more memorable. Nudges need not be covert. They aim to ensure that people are not put off from working towards their goals, and are instead motivated to persist (Sunstein, 2014).

On the basis of self-determination theory, teachers are more likely to be motivated to engage with research if they have the necessary skills, and the sense of competence that accompanies this skill (Ryan & Deci, 2017). The process of training staff in research-related skills can serve to nudge them towards more engagement, therefore, and in doing so, modify their professional identity. Research in social psychology has also established that a person's sense of identity can vary across situations depending on how a situation is framed (e.g. Ward & Wilson, 2015), and so presenting tasks as both relevant to student learning and within their professional capabilities will be important.

As discussed earlier, professionals need to retain their past learning about research and to be able transfer it to future tasks, and memory techniques are therefore a type of nudge. On a mundane level, teachers may benefit from reminders in the workplace situations (indeed, reminders and routines can form one of the most effective forms of nudges, and are frequently applied to everything from remembering medical appointments, to investing, to sticking to exercise regimes; Thaler & Sunstein, 2008). Strategies such as circular workplace emails with recommendations of research sources, posters showing evidence-based techniques, or reminders about professional projects at school meetings could all help to promote research engagement.

More broadly, and over time, such strategies can help to build a social norm which implies that research-engagement is part of the job.

7.6 General Conclusion

In terms of broader citizenship, the abilities for students and their teachers to learn successfully – to take in new concepts and retain information in a way that can transfer to new situations – is fundamentally important.

In the year of writing, 2020, the world faces two great crises. One is the Covid-19 outbreak, a health disaster of generational significance which has caused hundreds of thousands of deaths, and the other is the climate emergency, a slower-moving but possibly (over the long-term) even more deadly threat. It would be wrong to suggest that educational psychology can provide the solution to these problems, especially given the critical importance of fields such as bioscience (vaccines) and engineering (low-carbon technologies). However, any behavioural change involves an element of learning and memory. If a government is to give out heath advice, it is more likely to be followed if people remember it over the long term, and are capable of transferring a theoretical idea (such as keeping your distance to avoid spreading a virus, or reducing your carbon footprint) to their practical day-to-day choices. That is to say, they must remember what they are supposed to do and how to do it, and then use this knowledge and skill in novel situations.

Such concept learning and transfer may be especially critical in a changing world, when past learning situations and settings may not be repeated exactly. Young people who are students today will experience rapid change as they grow older, and the information that they have taken in during the school years (based on the prevailing situation at the time) may not be sufficient. Educators therefore widely recognise that critical thinking is important, but the application of critical thinking is often theoretically shallow, and lacks connection to evidence-based learning techniques and the science of memory and transfer. The current research provides a bridge between these areas.

Education is a domain in which misconceptions are widespread, and intuition cannot substitute for an evidence-based approach. This thesis has outlined several areas of evidence from cognitive psychology that are directly applicable to teaching, and students can benefit from learning these skills, while also developing their metacognitive understanding of how learning works, and thus being better able to learn successfully in the future.

Misconceptions are also a problem for the professional; desirable difficulties are counterintuitive techniques which appear not to be adopted spontaneously. While there is some consensus over what techniques are effective, engaging with research evidence directly (rather than via policy mandates) can empower the professional, allowing research to take its rightful place as an integral component of professional expertise.

8

References

Abel, M., & Roediger, H. L. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition*, *45*(1), 81–92. https://doi.org/10.3758/s13421-016-0641-8

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D.J., & Sanford, R. N. (1950). *The authoritarian personality*. Harper.

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning?. *Journal of Educational Psychology, 111*(2), 189–209. https://doi.org/10.1037/edu0000282

Agarwal, P. K., & Bain, P. M. (2019). *Powerful teaching: Unleash the science of learning.* John Wiley & Sons.

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*, 437–448. http://dx.doi.org/10.1007/s10648-012-9210-2

Alexander, R. (2014). Evidence, policy and the reform of primary education: A cautionary tale. *Forum, 56*(3), 349–375.

Allison, P. D. (2001). *Missing data* (Vol. 136). Sage.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Author.

American Psychological Association (APA) (2015). *20 top principles for PreK-12 teaching and learning*. APA website. https://www.apa.org/ed/schools/teaching-learning/top-twenty-principles.pdf

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*, 5–11. https://doi.org/10.3102/0013189X025004005

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete edition). Longman.

Anderson, R. C. (1984). Some reflections on the acquisition of knowledge. *Educational Researcher, 13*(9), 5–10. https://doi.org/10.2307/1174873

Andrews, D., Nonnecke, B., & Preece, J. (2003). Electronic survey methodology: A case study in reaching hard-to-involve Internet users. *International Journal of Human-computer Interaction, 16*(2), 185–210. https://doi.org/10.1207/S15327590IJHC1602_04

Arya, D. J., & Maul, A. (2012). The role of the scientific discovery narrative in middle school science education: An experimental study. *Journal of Educational Psychology, 104*, 1022–1032. https://doi.org/10.1037/a0028108

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men; research in human relations* (p. 177–190). Carnegie Press.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 33–53. https://doi.org/10.1037/0278-7393.14.1.33

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.) *The psychology of learning and motivation, Vol. 2* (pp. 89–195). Academic Press.

Baddeley, A. D. (1966). Short term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology, 18*, 362–5. https://doi.org/10.1080/14640746608400055

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.) *The psychology of learning and motivation: Vol 8*. (pp. 47–89). Academic Press.

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*, 627–635. https://doi.org/10.1080/00140137808931764

Baer, J. (2016). Creativity doesn't develop in a vacuum. *New Directions for Child and Adolescent Development, 151*(1), 9–20. https://doi.org/10.1002/cad.20151

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*, 296–308. https://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*, 316–321. https://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*(4), 566–577. https://doi.org/10.1016/j.jml.2005.01.012

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: mind wandering facilitates creative incubation. *Psychological Science, 23*, 1117–1122. https://doi.org/10.1177/0956797612446024

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 129–163). Cambridge University Press.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Battig, W. F. (1979). The flexibility of human memory. In L. S. Cermak, & F. I. M. Craik (Eds.), *Levels of processing and human memory* (pp. 23–44). Lawrence Erlbaum.

Bauernschmidt, A. (2017). *GUEST POST: Two Examples Are Better Than One*. Learning Scientists blog. https://www.learningscientists.org/blog/2017/5/30-1

BBC (2013, November 7th). *Hans Rosling: How much do you know about the world?* https://www.bbc.co.uk/news/magazine-24836917

Bebeau, M. J., & Monson, V. E. (2012). Professional identity formation and transformation across the life span. In McKee, A. & Eraut, M. (Eds.) *Learning trajectories, innovation and identity for professional development* (pp. 135–162). Springer.

Beck, A., Wall, K., Firth, J., Tonner, P., Arnott, L. (2020, August 24–28). Developing a practitioner enquiry approach to school-university research partnership [Conference paper]. EERA Conference, Glasgow, United Kingdom. https://eera-ecer.de/ecer-2020-glasgow/ (Conference cancelled).

Benassi, V. A., Overson, C. E., & Hakala, C. M. (Eds.) (2014). *Applying science of learning in education: Infusing psychological science into the curriculum*. Society for the Teaching of Psychology. http://teachpsych.org/ebooks/asle2014/index.php

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Erlbaum.

Bennett, T. (2017). There are no ninjas: Why the research revolution might rescue teaching. In L. Rycroft-Smith & J. L. Dutaut (Eds.) *Flip the system UK: A teachers' manifesto* (pp. 7–14). Routledge.

Betts, K., Miller, M., Tokuhama-Espinosa, T., Shewokis, P. A., Anderson, A., Borja, C., Galoyan, T., Delaney, B., Eigenauer, J. D., & Dekker, S. (2019). *International report: Neuromyths and evidence-based practices in higher education.* Online Learning Consortium.

Bickman, L. (1974). Clothes make the person. *Psychology Today, 8*(4), 48–51.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Biesta, G. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory, 57*(1), 1–22. https://doi.org/10.1111/j.1741-5446.2006.00241.x

Biesta, G., Priestley, M., & Robinson, S. (2015). The role of beliefs in teacher agency. *Teachers and Teaching, 21*(6), 624–640. https://doi.org/10.1080/13540602.2015.1044325

Biggs, J. B., & Collis, K. (1982). *Evaluating the quality of learning: the SOLO taxonomy*. Academic Press.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science, 12*, 587–625. https://doi.org/10.1016/0364-0213(88)90014-6

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31*, 635–650.

Birnbaum, M. S. (2013). *Understanding and optimizing the inductive learning of categories and concepts*. Unpublished doctoral dissertation, University of California Los Angeles.

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

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review, 30*, 703–725. https://doi.org/10.1007/s10648-018-9434-x

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. L. Solso (Ed.) *Information processing and cognition: The Loyola symposium (pp.* 123–144). Lawrence Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura, (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A. (2011). On the symbiosis of remembering, forgetting, and learning. In A. S. Benjamin (Ed.) *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 1–22). Psychology Press.

Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: looking back at Schmidt and Bjork (1992). *Perspectives on Psychological Science*, *13*, 146–148. https://doi.org/10.1177/1745691617690642

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. Pomeranz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 35– 67). Lawrence Erlbaum.

Bjork, R. A., & Bjork, E. L. (2019). Forgetting as the friend of learning: implications for teaching and self-regulated learning. *Advances in Physiology Education, 43*, 164–167. https://doi.org/10.1152/advan.00001.2019

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21. https://doi.org/10.1177/003172170408600105

Black, P., & Wiliam, D. (2006) Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and Learning* (pp. 81–100). Sage.

Blakemore, S. J., & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry, 47*(3–4), 296–312. https://doi.org/10.1111/j.1469-7610.2006.01611.x

Blakemore, S. J., & Mills, K. L. (2014). Is adolescence a sensitive period for sociocultural processing?. *Annual Review of Psychology, 65*, 187–207. https://doi.org/10.1146/annurev-psych-010213-115202

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I. Cognitive domain*. David McKay.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106*(3), 849–858. https://doi.org/10.1037/a0035934.

Borenstein, M. H., Hedges, L. V., Higgins, J. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Borenstein, M. H., Hedges, L. V., Higgins, J. T., & Rothstein, H. R. (2014). *Comprehensive meta analysis version 3*. Biostat.

Boser, U. (2019). *What do teachers know about the science of learning? A survey of educators on how students learn.* The Learning Agency. https://www.the-learning-agency.com/insights/what-do-teachers-know-about-the-science-of-learning.

Bower, G. H., & Clark, M. C. (1969). Narrative stories as mediators for serial learning. *Psychonomic Science, 14*(4), 181–182. https://doi.org/10.3758/BF03332778

Boyle, J. (2012). Understanding the nature of experiments in real world educational contexts. In B. Kelly, & D. Perkins (Eds.), *Handbook of implementation science for psychology in education* (pp. 54–67). Cambridge University Press.

Braasch, J. L. G., & Goldman, S. R. (2010). The role of prior knowledge in learning from analogies in science text. *Discourse Processes, 47,* 447–479. https://doi.org/10.1080/01638530903420960

Brainerd, C. J., Reyna, V. F., Howe, M. L., Kingma, J., & Guttentag, R. E. (1990). The development of forgetting and reminiscence. *Monographs of the Society for Research in Child Development, 55*(3/4), i+iii+v-vi+1–109. https://doi.org/10.2307/1166106

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school*. National Academy Press.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermack & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Erlbaum.

British Psychological Society (2014). *Code of human research ethics (2nd ed).* BPS Official Website. https://www.bps.org.uk/sites/bps.org.uk/files/Policy/Policy%20-%20Files/BPS%20Code%20of%20Human%20Research%20Ethics.pdf

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20*, 493–523. https://doi.org/10.1016/0010-0285(88)90014-X

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

Bruner, J. S. (1990). *Acts of meaning: Four lectures on mind and culture*. Harvard University Press.

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

Bryce, T. G. K. (2018). Assessment. In T. G. K. Bryce, W. M. Humes, D. Gillies, & A. Kennedy (Eds.), *Scottish Education* (5th ed., pp. 748–767). Edinburgh University Press.

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., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2015). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*, 474–478. https://doi.org/10.3758/BF03194092

Carpenter, S. K., Witherby, A., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2019.12.009

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29*(1), 3–20. https://doi.org/10.2307/749715

Carter, A. (2015). *Carter review of initial teacher training* (ITT). UK Department for Education.

Carvalho, P. F., & Albuquerque, P. B. (2012). Memory encoding of stimulus features in human perceptual learning. *Journal of Cognitive Psychology, 24*, 654–664. https://doi.org/10.1080/20445911.2012.675322

Carvalho, P. F., & Goldstone, R. L. (2014a). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42*, 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2014b). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology, 5*, 936. https://doi.org/10.3389/fpsyg.2014.00936

Carvalho, P. F., & Goldstone, R. L. (2015a). What you learn is more than what you see: what can sequencing effects tell us about inductive category learning?. *Frontiers in Psychology, 6*, 505. https://doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2015b). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*, 281–288. https://doi.org/10.3758/s13423-014-0676-4

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1699–1719. https://doi.org/10.1037/xlm0000406

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*, 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., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Chartered College of Teaching (2020). *Chartered teacher programme: 2020–2021 cohort.* https://chartered.college/wp-content/uploads/2019/11/Chartered-Teacher-booklet-2020-cohort-1.pdf

Chi, M. T. H., Glaser, R., and Rees, E. (1982). Expertise in problem solving. In R. S. Sternberg (Ed.). *Advances in the psychology of human intelligence, Vol. 1* (pp. 1–75). Erlbaum.

Chi, M. T., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*(1), 1–61. https://doi.org/10.1111/j.1551-6709.2011.01207.x

Chin, D. B., Chi, M., & Schwartz, D. L. (2016). A comparison of two methods of active learning in physics: inventing a general solution versus compare and contrast. *Instructional Science, 44*, 177–195. https://doi.org/10.1007/s11251-016-9374-0

Cho, K. W., & Neely, J. H. (2013). Null category-length and target–lure relatedness effects in episodic recognition: A constraint on item-noise interference models. *Quarterly Journal of Experimental Psychology*, *66*(7), 1331–1355. https://doi.org/10.1080/17470218.2012.739185

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1403–1414. https://doi.org/10.1037/0278-7393.25.6.1403

Clapper, J. P. (2015). The impact of training sequence and between-category similarity on unsupervised induction. *Quarterly Journal of Experimental Psychology, 68*, 1370–1390. doi:10.1080/17470218.2014.981553

Clapper, J. P., & Bower, G. H. (1991). Learning and applying category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation*, (Vol. 27, pp. 65–108). Academic Press.

Coe, R. (2020). *Does research on retrieval practice translate into classroom practice?* Chartered College of Teaching. https://impact.chartered.college/article/does-research-retrieval-practice-translate-classroom-practice/

Coe, R., Aloisi, C., Higgins, S., & Major, L. E. (2014). *What makes great teaching? Review of the underpinning research*. Sutton Trust website. http://www.suttontrust.com/wp-content/uploads/2014/10/What-makes-great-teaching-FINAL-4.11.14.pdf

Cook, M. P. (2006). Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education, 90*(6), 1073–1091. https://doi.org/10.1002/sce.20164

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research, 76*(1), 1–62. https://doi.org/10.3102/00346543076001001

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104*, 163–191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–94. https://doi.org/10.1037/0096-3445.104.3.268

Cruwys, T., Gaffney, A. M., & Skipper, Y. (2017). Uncertainty in transition: The influence of group cohesion on learning. In Mavor, K. I., Platow, M. J., & Bizumic, B. (Eds.), *Self and social identity in educational contexts (pp. 207–222)*. Routledge.

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law.* Cambridge University Press.

Czeisler, C., Johnson, M. P., Duffy, J. F., Brown, E. N., Ronda, J. M., & Kronauer, R. E. (1990). Exposure to bright light and darkness to treat physiologic maladaption to night work. *New England Journal of Medicine, 322*, 1253–9. https://doi.org/10.1056/NEJM199005033221801

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Darling-Hammond, L., Hyler, M. E., Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science, 24*(2), 97–140. https://doi.org/10.1080/10888691.2018.1537791

Davis, S. D., Chan, J. C., & 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

Deans for Impact (2015). *The science of learning*. Author.

Dellarosa, D., & Bourne, L. E. (1985). Surface form and the spacing effect. *Memory & Cognition, 13*(6), 529–537. https://doi.org/10.3758/BF03198324

Dement, W., & Kleitman, N. (1957). The relation of eye movements during sleep to dream activity: An objective method for the study of dreaming. *Journal of Experimental Psychology, 53*, 339–46. https://doi.org/10.1037/h0048189

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*, 309–330. https://doi.org/10.1007/BF01320097

Di Vesta, F. J., & Peverly, S. T. (1984). The effects of encoding variability, processing activity, and rule-examples sequence on the transfer of conceptual rules. *Journal of Educational Psychology, 76*, 108–119. https://doi.org/10.1037/0022-0663.76.1.108

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. Stuss & R. Knight (Eds.), *Principles of frontal lobe function (pp.* 466–503). Oxford University Press.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Didau, D. (2018). *Teaching to make children cleverer – Part 2*. Learning Spy website. https://learningspy.co.uk/psychology/teaching-make-children-cleverer-part-2/

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicised): Effects of disfluency on educational outcomes. *Cognition, 118*(1), 111–115. https://doi.org/10.1016/j.cognition.2010.09.012

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. http://doi.org/10.1016/j.jml.2007.11.004

Doane, D. P., & Seward, L. E. (2011). Measuring skewness. *Journal of Statistics Education, 19*(2), 1–18. https://doi.org/10.1080/10691898.2011.11889611

Dobson, J. L. (2011). Effect of selected "desirable difficulty" learning strategies on the retention of physiology information. *Advances in Physiology Education, 35*, 378–383. https://doi.org/10.1152/advan.00039.2011

Docter, P. (Director). (2015). *Inside out* [Motion picture]. Pixar Studios.

Donaldson, G. (2011). *Teaching Scotland's future: Report of a review of teacher education in Scotland*. Scottish Government.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

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*, 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Dovemark, M., Kosunen, S., Kauko, J., Magnúsdóttir, B., Hansen, P., & Rasmussen, P. (2018). Deregulation, privatisation and marketisation of Nordic comprehensive education: social changes reflected in schooling. *Education Inquiry, 9*(1), 122–141. https://doi.org/10.1080/20004508.2018.1429768

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology, 5*(3), 322–337. https://doi.org/10.1016/0010-0285(73)90040-6

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology, 1*(1), 72–78. https://doi.org/10.1037/stl0000024

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, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In Zanna, M. & Olson, J. (Eds.), *Advances in experimental social psychology, Vol 44*. (pp. 247–296). Academic Press.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Dover. (Original work published 1885).

Educational Endowment Foundation (EEF) (2016). *Testing the impact of project-based learning in secondary schools*. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/project-based-learning

Educational Endowment Foundation (EEF) (2018). *Teaching and learning toolkit.* https://educationendowmentfoundation.org.uk/public/files/Toolkit/complete/EEF-Teaching-Learning-Toolkit-October-2018.pdf

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Egan, K. (1985). Teaching as story-telling: A Non-mechanistic approach to planning teaching. *Journal of Curriculum Studies, 17*(4), 397–406. https://doi.org/10.1080/0022027850170405

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

Eglington, L. G., & Kang, S. H. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*, 475–485. https://doi.org/10.1016/j.jarmac.2017.07.005

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 397–417. https://doi.org/10.1037/0278-7393.7.6.397

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23. https://doi.org/10.1111/1467-8721.00160

Enser, M. (2019). *How useful is cognitive load theory for teachers*? Times Education Supplement. https://www.tes.com/news/how-useful-cognitive-load-theory-teachers

Ericsson, K. A. (2017). Expertise and individual differences: the search for the structure and acquisition of experts' superior performance. *WIREs Cognitive Science, 8*(1–2), e1382. https://doi.org/10.1002/wcs.1382

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7/8), 114–21.

Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing, 12*(1), 77–84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Feddern, L., Belham, F. S., & Wilks, S. (2019). *Retrieval, interleaving, spacing and visual cues as ways to improve independent learning outcomes at scale*. *Impact, volume 2.* https://impact.chartered.college/article/feddern-retrieval-interleaving-spacing-visual-cues-independent-learning/

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed). Sage.

Finch, J. (1987) Research note: The vignette technique in survey research. *Sociology, 21*(1), 105–14. https://doi.org/10.1177/0038038587021001008

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

Firth, J. (2017). Experts in learning. In L. Rycroft-Smith & J. L. Dutaut (Eds.) *Flip the system UK: A teachers' manifesto* (pp. 20–28). Routledge.

Firth, J. (2018a). Is it all just memorisation? *The Profession: The Annual Publication for Early Career Teachers, 1*, 31–35.

Firth, J. (2018b). *How to learn: Effective study and revision methods for any course*. Arboretum Books.

Firth, J. (2018c). Teachers' beliefs about memory: What are the implications for in-service teacher education? *Psychology of Education Review, 42*(2), 15–22.

Firth, J. (2019a). *The teacher's guide to research: Engaging with, applying and conducting research in the classroom*. Routledge.

Firth, J. (2019b). *National 5 & CfE Higher Psychology student book* (2nd Ed.). Leckie & Leckie.

Firth, J. (2020). Teacher classroom reflections – tackling flawed metacognition and memory. *Impact, 8*, 78–80.

Firth, J., Rivers, I., & Boyle, J. (2019). A systematic review of interleaving as a concept learning strategy: A study protocol. *Social Science Protocols, July 2019*, 1–7. http://dx.doi.org/10.7565/ssp.2019.2650

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Journal of Educational Psychology, 59*(4), 290–296. https://doi.org/10.1037/h0020055

Foot-Seymour, V., Foot, J., & Wiseheart, M. (2019). Judging credibility: Can spaced lessons help students think more critically online?. *Applied Cognitive Psychology, 33*(6), 1032–1043. https://doi.org/10.1002/acp.3539

Forsyth, S. R., Odierna, D. H., Krauth, D., & Bero, L. A. (2014). Conflicts of interest and critiques of the use of systematic reviews in policymaking: an analysis of opinion articles. *Systematic Reviews, 3*(1), 122. https://doi.org/10.1186/2046-4053-3-122

Furnham, A. (2018). Myths and misconceptions in developmental and neuro-psychology. *Psychology, 9*(02), 249–259. https://doi.org/10.4236/psych.2018.92016

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review, 26*(1), 9–25. https://doi.org/10.1007/s10648-014-9249-3

Gathercole, S. E., Alloway, T. P., Willis, C., & Adams, A. M. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology, 93*(3), 265–281. https://doi.org/10.1016/j.jecp.2005.08.003

General Teaching Council for Scotland (GTCS) (2012). *The Standards for Registration: mandatory requirements for Registration with the General Teaching Council for Scotland*. http://www.gtcs.org.uk/web/FILES/the-standards/standards-for-registration-1212.pdf

General Teaching Council for Scotland (GTCS) (n.d.). *Professional update. https://www.gtcs.org.uk/professional-update/professional-update.aspx*

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355. http://dx.doi.org/10.1016/0010-0285(80)90013-4

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93. https://doi.org/10.3758/BF03197714

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences, 14*, 69–110. https://doi.org/10.1207/s15327809jls1401_4

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729. https://doi.org/10.1037/bul0000043

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24*, 608–628. https://doi.org/10.3758/BF03201087

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*, 213–228. https://doi.org/10.1080/02671520701296189

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews?. *European Journal of Clinical Investigation, 48*(6). https://doi.org/10.1111/eci.12931

Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics, 95*, 97–101. https://doi.org/10.1111/j.1949-8594.1995.tb15736.x

Guzman-Munoz, F. J. (2017) The advantage of mixing examples in inductive learning: a comparison of three hypotheses. *Educational Psychology, 37*, 421–437. https://doi.org/10.1080/01443410.2015.1127331

Hamilton, R. (1990). The effect of elaboration on the acquisition of conceptual problem-solving skills from prose. *Journal of Experimental Education, 58*, 5–17. https://doi.org/10.1080/00220973.1990.10806547

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*, 267–271. https://doi.org/10.1257/aer.100.2.267

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition, 17*(5), 627–638. https://doi.org/10.3758/BF03197085

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. (2013). *Visible learning: A Synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition, 3*, 153–160. https://doi.org/10.1016/j.jarmac.2014.03.003

Heider, F. (1958). *The psychology of interpersonal relations.* Wiley.

Higgins, E., & Ross, B. (2011). Comparisons in category learning: How best to compare for what. In *Proceedings of the Cognitive Science Society, 33*, No. 33.

Higgins, J. P. T., & Thomas, J. (Eds.) (2019). *Cochrane handbook for systematic reviews of interventions* (V6). The Cochrane Collaboration. https://training.cochrane.org/handbook

Hofling, C. K., Brotzman, E., Dalrymple, S., Graves, N., & Pierce, C. M. (1966). An experimental study in nurse-physician relationships. *Journal of Nervous and Mental Disease, 143*, 171–180.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295–355. http://dx.doi.org/10.1207/s15516709cog1303_1

Hood, M. (2016). *Beyond the plateau: The case for an institute for advanced teaching.* Institute for Public Policy Research. https://www.ippr.org/files/publications/pdf/beyond-the-plateau_July2016.pdf

Horvath, J. C., & Lodge, J. M. (2016). A framework for organising and translating science of learning research. In Horvath, J. C., Lodge, J. M., & Hattie, J. (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 7–20). Routledge.

Howard-Jones, P. A. (2014) Neuroscience and education: Myths and messages. *Nature Reviews Neuroscience, 15*(12), 817–824. https://doi.org/10.1038/nrn3817

Hunter, M. (1979). Diagnostic teaching. *The Elementary School Journal, 80*(1), 41–46.

Husmann, P. R., & O'Loughlin, V. D. (2018). Another nail in the coffin for learning styles? Disparities among undergraduate anatomy students' study strategies, class performance, and reported VARK learning styles. *Anatomical Sciences Education, 12*(1), 6–19. https://doi.org/10.1002/ase.1777

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology, 82*(3), 472–481. https://doi.org/10.1037/h0028372

Institute of Education Sciences (n.d.). *WWC: What works clearing house*. https://ies.ed.gov/ncee/wwc/

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In. P. R. Krishnaiah (Ed.), *Handbook of statistics* (pp. 199–236). North-Holland Publications.

Jaeger, A. J., Taylor, A. R., & Wiley, J. (2016). When, and for whom, analogies help: The role of spatial skills and interleaved presentation. *Journal of Educational Psychology, 108*(8), 1121–1139. https://doi.org/10.1037/edu0000121

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. http://doi.org/10.1016/j.jml.2007.11.007

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory, 18*(4), 394–412. https://doi.org/10.1080/09658211003702171.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kang, S. H. (2016a). The benefits of interleaved practice for learning. In Horvath, J. C., Lodge, J. M. & Hattie, J. (Eds.) *From the laboratory to the classroom: Translating the science of learning for teachers* (pp. 79–93). Routledge.

Kang, S. H. (2016b). Spaced repetition promotes efficient and effective learning policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 12–19. https://doi.org/10.1177/2372732215624708.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*, 97–103. https://doi.org/10.1002/acp.1801

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. (2016). *A powerful way to improve learning and memory: Practicing retrieval enhances long-term, meaningful learning.* Psychological Science Agenda. http://www.apa.org/science/about/psa/2016/06/learning-memory.aspx

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., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. Ross (Ed.) *Psychology of Learning and Motivation, Vol 61* (pp. 237–284). Academic Press.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education, 106*, 166–171. https://doi.org/10.1016/j.compedu.2016.12.006

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kliegl, O., Bjork, R. A., & Bäuml, K. H. T. (2019). Feedback at test can reverse the retrieval-effort effect. *Frontiers in Psychology, 10*, 1863. https://doi.org/10.3389/fpsyg.2019.01863

Koedinger, K. R., & Roll, I. (2012). Learning to think: Cognitive mechanisms of knowledge transfer. In: Holyoak K. J. & Morrison R. G. (Eds.), *Oxford handbook of thinking and reasoning, 2nd ed* (pp. 789–806). Cambridge University Press.

Kohlberg, L. (1963). The development of children's orientations toward a moral order. *Human Development*, *6*(1–2), 11–33. https://doi.org/10.1159/000112530

Kokkotas, P., Rizaki, A., & Malamitsa, K. (2010). Storytelling as a strategy for understanding concepts of electricity and electromagnetism. *Interchange, 41*(4), 379–405. https://doi.org/10.1007/s10780-010-9137-9.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition 9*, 149–171. https://doi.org/10.1006/ccog.2000.0433

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297–1317. https://doi.org/10.1002/acp.1537

Kornell, N. (2015). If it is stored in my memory I will surely retrieve it: anatomy of a metacognitive belief. *Metacognition and Learning, 10*(2), 279–292. https://doi.org/10.1007/s11409-014-9125-z

Kornell, N., Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224. https://doi.org/10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449–68. https://doi.org/10.1037/a0017350.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. Psychology and Aging, 25, 498–503. https://doi.org/10.1037/a0017807

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Sage.

Küpper-Tetzel, C. (2017). *On the potential limitations of spacing and retrieval practice in the classroom (and the need for more applied research)*. Learning Scientists blog. https://www.learningscientists.org/blog/2017/7/20-1

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science, 42*(3), 373–388. https://doi.org/10.1007/s11251-013-9285-2

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 & Cognition, 42*(5), 729–741. https://doi.org/10.3758/s13421-014-0394-1

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). Academic Press.

LaPiere, R.T. (1934). Attitudes vs. actions. *Social Forces, 13*, 230–7. https://doi.org/10.2307/2570339

Lave, J. (1991). Situating learning in communities of practice. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 63–82). American Psychological Association.

Leat, D. (Ed., 2017). *Enquiry and project-based learning: Students, school and society*. Routledge.

Lilienfeld, S. O., Lynn, S. J., Ruscio, J., & Beyerstein, B. L. (2011). *50 great myths of popular psychology: Shattering widespread misconceptions about human behavior.* Wiley.

Lin, T. H. (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality and Quantity, 44,* 277–287. https://doi.org/10.1007/s11135-008-9196-5

Linderholm, T., Dobson, J., & Yarbrough, M. B. (2016). The benefit of self-testing and interleaving for synthesizing concepts across multiple physiology texts. *Advances in Physiology Education, 40*, 329–334. https://doi.org/10.1152/advan.00157.2015

Loftus, E. F. (2000). Suggestion, imagination, and the transformation of reality. In A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman & V. S. Cain (Eds.) *The science of self-report: Implications for research and practice* (pp. 201–210). Lawrence Erlbaum.

Loftus, E. F. (2019). Eyewitness testimony. *Applied Cognitive Psychology, 33*(4), 498–503. https://doi.org/10.1002/acp.3542

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behaviour, 13*, 585–9. https://doi.org/10.1016/S0022-5371(74)80011-3

Logie, R. H. (2016). Retiring the central executive. *Quarterly Journal of Experimental Psychology, 69*(10), 2093–2109. https://doi.org/10.1080/17470218.2015.1136657

Lowden, K., Hall, S., Bravo, A., Orr, C., & Chapman, C. (2019). *Scottish education system: knowledge utilisation study*. Edinburgh: Scottish Government.

Loyd, D. L., Kern, M. C., & Thompson, L. (2005). Classroom research: Bridging the ivory divide. *Academy of Management Learning & Education, 4*(1), 8–21. https://doi.org/10.5465/amle.2005.16132533

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language, 58*(2), 480–494. https://doi.org/10.1016/j.jml.2007.04.004

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

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.) *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

McDaniel, M. A., & Einstein, G. O. (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed.), *Decade of behavior. Experimental cognitive psychology and its applications* (p. 73–85). American Psychological Association. https://doi.org/10.1037/10895-006

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

Macdonald, K., Germine, L., Anderson, A., Christodoulou, J., & McGrath, L. M. (2017). Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in Psychology, 8*, 1314. https://doi.org/10.3389/fpsyg.2017.01314

MacKendrick, A. (2015). *Interleaved effects in inductive category learning: The role of memory retention*. Unpublished doctoral dissertation, University of South Florida.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9*(3), 241–289. https://doi.org/10.1016/0167-9457(90)90005-X

Magnussen, S., & Melinder, A. (2012). What psychologists know and believe about memory: A survey of practitioners. *Applied Cognitive Psychology, 26*(1), 54–60. https://doi.org/10.1002/acp.1795

Magnussen, S., Wise, R. A., Raja, A. Q., Safer, M. A., Pawlenko, N., & Stridbeck, U. (2008). What judges know about eyewitness testimony: A comparison of Norwegian and US judges. *Psychology, Crime and Law, 14*, 177–188. https://doi.org/10.1080/10683160701580099

Manly, C. A., & Wells, R. S. (2015). Reporting the use of multiple imputation for missing data in higher education research. *Research in Higher Education, 56*(4), 397–409. https://doi.org/10.1007/s11162-014-9344-9

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, *13*(2), 125–139. https://doi.org/10.1016/S0959-4752(02)00016-6

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

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455. https://doi.org/10.1037/a0028085.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. Hematology/Oncology Clinics of North America, 14(4), 745–760. https://doi.org/10.1016/S0889-8588(05)70309-9

Melinder, A., & Magnussen, S. (2015). Psychologists and psychiatrists serving as expert witnesses in court: what do they know about eyewitness memory?. *Psychology, Crime & Law, 21*(1), 53–61. https://doi.org/10.1080/1068316X.2014.915324

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. (2015). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 978–984. https://doi.org/10.1037/xlm0000216

Michalski, R. S. & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 331–364). Tioga Publishing.

Milgram, S. (1963). Behavioural study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–8. https://doi.org/10.1037/h0040525

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction?. *Psychological Bulletin, 82*(2), 213–225. https://doi.org/10.1037/h0076486

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory, 24*(2), 257–271. https://doi.org/10.1080/09658211.2014.1001992

Mori, K., & Arai, M. (2010). No need to fake it: Reproduction of the Asch experiment without confederates. *International Journal of Psychology, 45*(5), 390–397. https://doi.org/10.1080/00207591003774485

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25*(6), 1159–1168. https://doi.org/10.1177/0956797614524581

Mulligan, N. W., & Osborn, K. (2009). The modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 564–571. https://doi.org/10.1037/a0014524

Mulrow, C.D., Thacker, S.B., & Pugh, J.A. (1988). A proposal for more informative abstracts of review articles. *Annals of Internal Medicine, 108*, 613–615. https://doi.org/10.7326/0003-4819-108-4-613

Murata, A. (2011). Introduction: Conceptual overview of lesson study. In L. Hart, A. S. Alston, & A. Murata (Eds.), *Lesson study research and practice in mathematics education* (pp. 1–12). Springer.

Murdock Jr, B. B. (1967). Recent developments in short-term memory. *British Journal of Psychology, 58*(3–4), 421–433. https://doi.org/10.1111/j.2044-8295.1967.tb01099.x

Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. (2007). Adaptive memory: survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 263–273. https://doi.org/10.1037/0278-7393.33.2.263

Neff, J. A. (1979). Interaction versus hypothetical other: The use of vignettes in attitude research. *Sociology and Social Research*, *64*, 105–125.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). MIT Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition, 155*, 23–29. https://doi.org/10.1016/j.cognition.2016.06.007

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know?. *Psychological Bulletin, 131*(5), 763–784. https://doi.org/10.1037/0033-2909.131.5.763

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Science Education, 15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Norris, S. P. (1985). The philosophical basis of observation in science and science education. Journal of Research in Science Teaching, 22(9), 817–833. https://doi.org/10.1002/tea.3660220905

Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–83. https://doi.org/10.1037/h0043424

Ost, J., Easton, S., Hope, L., French, C. C., & Wright, D. B. (2017). Latent variables underlying the memory beliefs of chartered clinical psychologists, hypnotherapists and undergraduate students. *Memory*, *25*(1), 57–68. https://doi.org/10.1080/09658211.2015.1125927

Papadatou-Pastou, M., Gritzali, M., & Barrable, A. (2018). The learning styles educational neuromyth: Lack of agreement between teachers’ judgments, self-assessment, and students’ intelligence. *Frontiers in Education, 3*, 105. https://doi.org/10.3389/feduc.2018.00105

Papadatou-Pastou, M., Haliou, E., & Vlachos, F. (2017). Brain knowledge and the prevalence of neuromyths among prospective teachers in Greece. *Frontiers in Psychology, 8*, 804. https://doi.org/10.3389/fpsyg.2017.00804

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

Peca, K. (2000). *Positivism in education: Philosophical, research and organizational assumptions*. U.S. Department of Education.

Peterson, L. R. & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*, 193–198. https://doi.org/10.1037/h0049234

Peterson, D. J., & Wissman, K. (2020). Using tests to reduce mind-wandering during learning review. *Memory,* 1–6. https://doi.org/10.1080/09658211.2020.1748657

Piaget, J. (1926). *The language and thought of the child* (M. Gabain, trans.). Routledge & Kegan Paul.

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, trans.). International Universities Press.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice, 41*(4), 219–225. https://doi.org/10.1207/s15430421tip4104_3

Popper, K. (1963), *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology, 84*(4), 892–897. https://doi.org/10.1111/1365-2656.12382

Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy, 70*(19), 699–711. https://doi.org/10.2307/2025108

Racsmány, M., Szőllősi, A., & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory & Cognition*. https://doi.org/10.3758/s13421-020-01041-5

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough?. *Journal of Experimental Psychology: General, 140*(3), 283–302. https://doi.org/10.1037/a0023956

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review, 27*, 483–504. https://doi.org/10.1007/s10648-014-9273-3

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Sarling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33.

Reynolds, K.J, Subašić, E., Lee, E., Bromhead, D., & Tindall, K. (2015). Does education really change us? The impact of school-based social processes on the person. In Reynolds, K.J. & Branscome, N.R. (eds.), *Psychology of change: Life contexts, experiences, and identities*. Psychology Press.

Riazat, N., & Firth, J. (2020). Memories that stick. In Chartered College of Teaching (Ed.), *The early career framework handbook,* (pp. 45–54). Sage.

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

Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology, 26*(2), 234–242. https://doi.org/10.1002/acp.1814

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology, 101*(4), 836–852. https://doi.org/10.1037/a0016026

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Robinson, O. C. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative Research in Psychology*, *11*(1), 25–41. https://doi.org/10.1080/14780887.2013.801543

Robinson, P., & Lowe, J. (2015). Literature reviews vs systematic reviews. *Australian and New Zealand Journal of Public Health, 39*(2), 103–103. https://doi.org/10.1111/1753-6405.12393

Robson, C. (2002). *Real world research: A resource for social scientists and practitioner-researchers* (2nd Ed.). Blackwell.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242–248. https://doi.org/10.1016/j.jarmac.2012.09.002

Röer, J. P., Bell, R., & Buchner, A. (2013). Is the survival-processing memory advantage due to richness of encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1294–1302. https://doi.org/10.1037/a0031214

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology, 35*(9), 677–688. https://doi.org/10.1037/0022-3514.35.9.677

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24*, 355–367. https://doi.org/10.1007/s10648-012-9201-3

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science, 16*(4), 183–186. https://doi.org/10.1111/j.1467-8721.2007.00500.x

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39*, 406–412. https://doi.org/10.3102/0013189X10374770

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

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*(6), 481–498. https://doi.org/10.1007/s11251-007-9015-8

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*, 900–908. https://doi.org/10.1037/edu0000001

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*, 341–357. https://doi.org/10.1037/0033-295X.96.2.341

Rumelhart, D. E. (1991). Understanding understanding. In W. Kessen, A. Ortony & F. Craik (Eds.), *Memories, thoughts and emotions: Essays in honor of George Mandler* (pp. 257–275). Lawrence Erlbaum.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications.

Sachs, J. (2016). Teacher professionalism: why are we still talking about it?. *Teachers and Teaching, 22*(4), 413–425. https://doi.org/10.1080/13540602.2015.1082732

Salomon, G. (1992). The changing role of the teacher: From information transmitter to orchestrator of teaching. In F. K. Oser, A. Dick, & J.-L. Patry (Eds.), *Effective and responsible teaching: The new synthesis* (pp. 37–49). Jossey-Bass.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist*, *24*, 113–142. https://doi.org/10.1207/s15326985ep2402_1

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology, 109*, 84–98. https://doi.org/10.1037/edu0000119

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit?. *Journal of Applied Research in Memory and Cognition, 7*, 361–369. https://doi.org/10.1016/j.jarmac.2018.05.005

Schalk, L., Schumacher, R., Barth, A., & Stern, E. (2018). When problem-solving followed by instruction is superior to the traditional tell-and-practice sequence. J*ournal of Educational Psychology, 110*(4), 596–610. https://doi.org/10.1037/edu0000234

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schooler, C. (1989). Social structural effects and experimental situations. In K. W. Schaie & C. Schooler (Eds.), *Social structure and aging: Psychological processes* (pp. 1–21). Erlbaum.

Schwartz, B. L., & Efklides, A. (2012). Metamemory and memory efficiency: Implications for student learning. *Journal of Applied Research in Memory and Cognition, 1*(3), 145–151. https://doi.org/10.1016/j.jarmac.2012.06.002

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–522. https://doi.org/10.1207/s1532690xci1604_4

Scottish Government (2018). *Pupils in Scotland 2018 (Supplementary Data).* https://www2.gov.scot/Topics/Statistics/Browse/School-Education/dspupcensus/dspupcensus18

Scottish Qualification Authority (SQA) (2018). *Higher psychology course specification.* https://www.sqa.org.uk/files_ccc/HigherCourseSpecPsychology.pdf

Scottish Qualification Authority (SQA) (2019). *Higher Psychology course report 2019.* https://www.sqa.org.uk/files_ccc/2019HCourseReportPsychology.pdf

Scutt, C. (2020, January 20th). *How much do teachers really need to know about the science of learning?* https://schoolsweek.co.uk/how-much-do-teachers-really-need-to-know-about-the-science-of-learning/

Sears, D.O. (1986). College sophomores in the laboratory: Influences of a narrow data base on psychology's view of human nature. *Journal of Personality and Social Psychology, 51*, 513–30. https://doi.org/10.1037/0022-3514.51.3.515

Senior, V., Weinman, J., & Marteau, T. M. (2002). The influence of perceived control over causes and responses to health threats: A vignette study. *British Journal of Health Psychology*, *7*(2), 203–211. https://doi.org/10.1348/135910702169448

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591–611. https://doi.org/10.2307/2333709

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

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences, 3*, 153–160.

Short, D. (2012). Teaching scientific concepts using a virtual world - Minecraft. *Journal of the Australian Science Teachers Association*, *58*(3), 55–58.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology, 70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Simon, H. A., & Chase, W.G. (1973). Skill in chess. *American Scientist, 61*, 394–403.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. *PloS one, 6*(8), e22757. https://doi.org/10.1371/journal.pone.0022757

Simons, D. J., & Chabris, C. F. (2012). Common (mis)beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PloS one*, *7*(12), e51876. https://doi.org/10.1371/journal.pone.0051876

Slone, L. K., & Sandhofer, C. M. (2017). Consider the category: The effect of spacing depends on individual learning histories. *Journal of Experimental Child Psychology, 159*, 34–49. https://doi.org/10.1016/j.jecp.2017.01.010

Smith, M. (2018). *The emotional learner.* Routledge. https://doi.org/10.4324/9781315163475

Smith, M. & Firth, J. (2018). *Psychology in the classroom: A teacher's guide to what works.* Routledge. https://doi.org/10.4324/9781315163420

Smith, C. D., & Scarf, D. (2017). Spacing repetitions over long timescales: A review and a reconsolidation explanation. *Frontiers in Psychology, 8*, 962. https://doi.org/10.3389/fpsyg.2017.00962

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

Soderstrom, N.C. and Bjork, R.A. (2015). Learning versus performance: An integrative review. P*erspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition, 33*(6), 1116–1129. https://doi.org/10.3758/BF03193217

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

Sonnenberg, S. J. (2017). Student identity and the marketisation of higher education. In Mavor, K. I., Platow, M. J., & Bizumic, B. (Eds.), *Self and social identity in educational contexts* (pp. 257–276). Routledge. https://doi.org/10.4324/9781315746913-ch15

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*(5), 315–317. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Steinberg, L., & Monahan, K. C. (2007). Age differences in resistance to peer influence. *Developmental Psychology, 43*(6), 1531–1543. https://doi.org/10.1037/0012-1649.43.6.1531.

Stewart, W. (2015). *Leave research to the academics, John Hattie tells teachers*. Times Educational Supplement online. https://www.tes.com/news/leave-research-academics-john-hattie-tells-teachers

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Sullivan, G. M. (2011). Getting off the "gold standard": randomized controlled trials and education research. *Journal of Graduate Medical Education, 3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1

Surma, T., Vanhoyweghen, K., Camp, G., & Kirschner, P. A. (2018). The coverage of distributed practice and retrieval practice in Flemish and Dutch teacher education textbooks. *Teaching and Teacher Education, 74*, 229–237. https://doi.org/10.1016/j.tate.2018.05.007

Sweller, J., Ayres, P. & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Taber, K. (2013). *Classroom-based research and evidence-based practice: An introduction*. Sage. https://doi.org/10.4135/9781849208734

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223*, 96–105.

Tanaka, J. W., & Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science, 12*(1), 43–47. https://doi.org/10.1111/1467-9280.00308

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study?. *Psychonomic Bulletin & Review, 20*(2), 356–363. https://doi.org/10.3758/s13423-012-0319-6

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: the (lack of) effect of Sans Forgetica on memory. *Memory,* 1–8. https://doi.org/10.1080/09658211.2020.1758726

Taylor, J. S., DeMers, S. M., Vig, E. K., & Borson, S. (2012). The disappearing subject: exclusion of people with cognitive impairment and dementia from geriatrics research. *Journal of the American Geriatrics Society*, *60*(3), 413–419. https://doi.org/10.1111/j.1532-5415.2011.03847.x

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*, 837–848. https://doi.org/10.1002/acp.1598

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press.

Toscano, J. C., Buxó-Lugo, A., & Watson, D. G. (2015, January). Using game-based approaches to increase level of engagement in research and education. *Teachercraft* (pp. 139–151). ETC Press.

Tsabet, L. (2018). *Why an 'interleaving' curriculum could improve knowledge retention.* TES Online. https://www.tes.com/news/why-interleaving-curriculum-could-improve-knowledge-retention

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). Academic Press.

Tulving, E. (2008). On the law of primacy. In Gluck, M. A., Anderson, J. R., & Kosslyn, S. M. (Eds.), *Memory and mind: A festschrift for Gordon. H. Bower* (pp. 31–48). Lawrence Erlbaum.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Underwood, J. D. (2003). Student attitudes towards socially acceptable and unacceptable group working practices. *British Journal of Psychology*, *94*(3), 319–337. https://doi.org/10.1348/000712603767876253

Valdez, P., Ramírez, C., & García, A. (2014). Circadian rhythms in cognitive processes: implications for school learning. *Mind, Brain, and Education, 8*(4), 161–168. https://doi.org/10.1111/mbe.12056

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Vehkalahti, K., Puntanen, S., & Tarkkonen, L. (2006). Estimation of reliability: a better alternative for Cronbach’s alpha. Helsinki: Department of Mathematics and Statistics, University of Helsinki.

Verkoeijen, P., & Bouwmeester, S. (2014). Is spacing really the “friend of induction”?. *Frontiers in Psychology, 5*, 259. https://doi.org/10.3389/fpsyg.2014.00259

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children’s memory and category induction. *Cognition, 109*, 163–167. https://doi.org/10.1016/j.cognition.2008.07.013

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*, 750–763. https://doi.org/10.3758/s13421-010-0063-y

Wall, K. (2018). Building a bridge between pedagogy and methodology: emergent thinking on notions of quality in practitioner enquiry. *Scottish Educational Review, 50*(2), 3–22.

Wall, K., Beck, A., & Firth, J. (2018). A practical introduction to practitioner enquiry. [Conference paper]. Engage Strathclyde Conference, University of Strathclyde, Glasgow. https://doi.org/10.13140/RG.2.2.22864.20480

Wall, K., & Hall, E. (2016). Teachers as metacognitive role models. *European Journal of Teacher Education, 39*(4), 403–418. ttps://doi.org/10.1080/02619768.2016.1212834

Wall, K., & Hall, E. (2019). *Research methods for understanding professional learning*. Bloomsbury.

Ward, C. L., & Wilson, A. E. (2015). Implicit theories of change and stability moderate effects of subjective distance on the remembered self. *Personality and Social Psychology Bulletin, 41*(9), 1167–1179. https://doi.org/10.1177/0146167215591571

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness. *Sociological Methods & Research, 2*1(1), 52–88. https://doi.org/10.1177/0049124192021001003

Webb, H. (2019). *Revision techniques: Interleaving and spacing.* SecEd – The Voice for Secondary Education. https://www.sec-ed.co.uk/best-practice/revision-techniques-interleaving-and-spacing/

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, Y., Sumeracki, M., & Caviglioli, O. (2018). *Understanding how we learn: A visual guide*. Routledge.

Wiliam, D. (2010, March). *Teacher quality: why it matters, and how to get more of it.* The Spectator. http://www.vcsta.org/wp-content/uploads/2013/03/Spectator-talk.pdf

Wiliam, D. (2019). *Dylan Wiliam: Teaching not a research-based profession*. Times Education Supplement online. https://www.tes.com/news/dylan-wiliam-teaching-not-research-based-profession

Willingham, D. (2004). *Ask the cognitive scientist: Practice makes perfect—but only if you practice beyond the point of perfection*. American Federation of Teachers. https://www.aft.org/periodical/american-educator/spring-2004/ask-cognitive-scientist-practice-makes-perfect

Willingham, D. T. (2017). A mental model of the learner: Teaching the basic science of educational psychology to future teachers. *Mind, Brain, and Education, 11*(4), 166–175.

Wilson, K., & Korn, J. H. (2007). Attention during lectures: Beyond ten minutes. *Teaching of Psychology, 34*(2), 85–89. https://doi.org/10.1080/00986280701291291

Wilson, S. (2016). Divergent thinking in the grasslands: thinking about object function in the context of a grassland survival scenario elicits more alternate uses than control scenarios. *Journal of Cognitive Psychology, 28*(5), 618–630. https://doi.org/10.1080/20445911.2016.1154860

Wise, R. A., & Kehn, A. (2020). Can the effectiveness of eyewitness expert testimony be improved?. *Psychiatry, Psychology and Law, 27*(2), 315–330. https://doi.org/10.1080/13218719.2020.1733696

Wojcik, S., & Hughes, A. (2019). *Sizing up Twitter users*. Pew Research Centre. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/04/twitter_opinions_4_18_final_clean.pdf

Wright, D. B., London, K., & Field, A. P. (2011). Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology*, 2(2), 252–270. https://doi.org/10.5127/jep.013611

Xu, J., & Metcalfe, J. (2016). Studying in the region of proximal learning reduces mind wandering. *Memory & Cognition, 44*(5), 681–695. https://doi.org/10.3758/s13421-016-0589-8

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General, 145*(7), 918–33. https://doi.org/10.1037/xge0000177

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. Lodge, & J. Hattie (Eds). *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 61–78). Routledge.

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied, 23*, 403–416. https://doi.org/10.1037/xap0000139

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*, 229–241. https://doi.org/10.3758/s13421-012-0255-8

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction, 46*, 12–20. https://doi.org/10.1016/j.learninstruc.2016.08.002

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*, 41–44. https://doi.org/10.3758/BF03329756

Zeigarnik, B. (1927). Das Behalten erledigter und unerledigter Handlungen. *Psychologische Forschung, 9*, 1–85.

Zhao, Y. (2017). What works may hurt: Side effects in education. *Journal of Educational Change, 18*(1), 1–19. https://doi.org/10.1007/s10833-016-9294-4

Zulkiply, N. (2013). Effect of interleaving exemplars presented as auditory text on long-term retention in inductive learning. *Procedia – Social and Behavioral Sciences, 97*, 238–245. https://doi.org/10.1016/j.sbspro.2013.10.228

Zulkiply, N. (2015). The role of bottom-up vs. top-down learning on the interleaving effect in category induction. *Pertanika Journal of Social Sciences & Humanities, 23*, 933–944.

Zulkiply, N., & Burt, J. S. (2013a). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*, 16–27. https://doi.org/10.3758/s13421-012-0238-9

Zulkiply, N., & Burt, J. S. (2013b). Inductive learning: Does interleaving exemplars affect long-term retention?. *Malaysian Journal of Learning and Instruction, 10*, 133–155.

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction, 22*, 215–221. https://doi.org/10.1016/j.learninstruc.2011.11.002

Appendix 1: Article for student teachers/NQTs

The following article appeared in the magazine for student and probationary teachers published by the Chartered College of Teaching, entitled 'The Profession' in 2018, and was then republished for the 2019 edition. Firth, J. (2018). Is it all just memorisation? The Profession: The Annual Publication for Early Career Teachers, 1, 31–35.

Is It All Just Memorisation?

by J. Firth

As a memory researcher and teacher educator, I sometimes get a negative reaction when I talk about improving teachers' understanding of human memory, and making educational practice more evidence-based. Am I suggesting that school learning is just a matter of memory - and isn't memorisation a bad thing? I think that using memory effectively isn't a bad thing at all, and in fact it's inevitable - all learning involves memory on some level. In this article I explore what cognitive researchers mean when they talk about memory, as well as discussing some promising research findings that have begun to be applied to education.

Memory as viewed by cognitive science

While recognising that the foundation of learning is the brain (on one level, all learning must involve changes to the physical structure or interconnection of neurons), cognitive science explains learning in terms of thought processes and behaviour. When something new is learned, a memory is stored or altered in a way that will impact on a learner's ability to think or act. For example, practicing spelling can make mistakes less likely to occur in future, while studying a concept in class should make it easier for learners to understand examples of this concept that they might see in an exam, or when reading.

Learning without understanding would of course be problematic, but in practice, understanding and memory are closely intertwined. It’s very difficult to memorise something that you don’t understand at all - imagine trying to memorise a text in an unfamiliar language! Research dating back several decades has shown that the more meaningful a learning task and the more deeply we think about new information, the better we remember. For example, if people are shown a list of words and asked how much they like each item on the list, they are much more likely to remember them than if they are asked to do something superficial, such as state whether each word contains a letter ‘e’ (Hyde & Jenkins, 1973). This *depth of processing* principle can easily be applied to teaching, by providing learners with tasks which prompt them to respond to ideas, rather than simply copying them down.

Interconnected understanding

Many new teachers may rightly be worried that trying to improve memory is too simplistic - we also want learners to understand. Cognitive psychologists look at understanding in terms of interconnections between words, experiences and ideas. This helps us to see why learning facts in a single context does not lead to understanding - or to successful memory. If a new concept is only connected to one other thing then the overall structure of what has been learned is fragile, prone to being forgotten, and provides a poor foundation for later learning. For example, imagine trying to teach a science class about chemical bonding, if they don’t fully understand how atoms work. It’s perhaps not difficult to see why they might misunderstand and rapidly forget such information.

In contrast, the key feature of something that is well learned and understood is that learning is connected to existing knowledge, and can be used by learners in multiple situations. Psychologists call a structure of interconnected knowledge a *schema*. You have a schema for a classroom, for example, but this means a lot more than just being able to define it - you also know who you would typically find in a classroom, what kind of events occur there, where people sit or stand, and so forth. Understanding of a concept therefore includes more than just the abstract 'fact', but instead represents how this fact is linked to other things.

The last point relates to another issue - flexible, well-structured memories are necessary for *transfer*. By transfer, we mean how well a learner is able to apply their learning to new situations. For example, a learner who has studied the solar system may find out about the planetary system of another star, and make the (correct) inference that the planets of this new system revolve around the star.

Another implication of looking at learning in terms of schemas is that all new learning should connect to prior learning. This helps to explain why we, as teachers, can give a clear explanation of something, only to find that several members of the class are giving us blank looks. The problem is not always with what we do, but with what they currently know. We may need to find out more about the learners' existing knowledge, and link content to their background and interests (Moll & Gonzalez, 1994).

Evidence-informed techniques

It's not enough just to know how memories are structured - researchers also want to find ways to help learning progress more effectively. Clearly learning begins with new information being presented to learners, for example via reading, a video, an experience or a lecture. However, it is now well understood that presentation of information is insufficient. As discussed above, the new tasks should involve a meaningful context where they can connect information to what they already know.

It’s also important to prevent learners from forgetting, and - counterintuitively - one way to do so is to prompt recall of concepts from memory. This *retrieval practice* (which could occur when answering a question or when writing or speaking about the concept) appears to significantly reduce forgetting, but the effect is only obvious over time delays of a few days or more (Karpicke et al., 2013). The benefits are greatest when the retrieval is spaced out over time, with gaps between study and revision (Dunlosky & Rawson, 2015) - a benefit usually termed the *spacing effect*.

When I was a school pupil myself, we were frequently asked to copy sections of text from a book or screen into jotters. This sort of task does not require retrieval practice, because pupils are not asked to remember anything for more than a few seconds. A better alternative would be to give them an explanation with their pens down, and at the end of the explanation, ask them to write down everything they can remember. To ensure accuracy, this could then be checked against a textbook or information sheet. Any review that took place could be done after a delay (e.g. the next day), making use of the spacing effect.

It might seem self-evident that looking at examples of the same type of thing (e.g. presenting several examples of industrialisation in History or Economics) would aid learning. However, studying examples of different concepts mixed together results in better memory for those concepts, especially when the concepts are similar to each other and therefore easily confused (Carvalho & Goldstone, 2015). This technique is known as *interleaving*. In one experiment, researchers compared interleaving of real-world examples of social science concepts versus showing learners multiple examples of the same concept. They found that interleaving led to better memory for the examples, as well as a superior ability to correctly identify previously unseen examples (Rawson et al., 2015). It seems that comparing easily-confused concepts side-by-side makes it easier for learners to notice and remember key differences between them.

What is less useful?

While studying effective learning techniques, researchers have also scrutinised traditional teaching practices, and in some cases found them less effective than might be assumed. One example is *overlearning* - engaging in repeated practice of concepts beyond the point where a learner has already 'got' the key idea (often termed the point of *mastery*). Textbooks tend to use this approach - who hasn't worked through a page of very similar arithmetic problems?

However, it appears that overlearning is useless over the long term. For example, a study by Rohrer and Taylor (2006) gave two groups maths problems to work though. One group did three problems, and the other did nine problems. The average accuracy after the third problem was around 90% for both groups, meaning that problems 4-9 constituted overlearning. When tested on similar tasks four weeks later, there was no difference in test scores between the two groups - suggesting that the later practice problems had essentially been a waste of time. Once mastery has been achieved, it would be more efficient for learners to stop and move onto something else, perhaps practising additional items days or even weeks later (i.e. making use of the spacing effect).

Re-reading is another popular technique which seems to have limited benefits. By reading something two or more times, the learner acquires a sense of familiarity with the material, and therefore feels like they understand it. However, this feeling can be an illusion - it might not translate into better memory or transfer. A more effective strategy would be for the learner to summarise key points and then test themselves, thereby making use of retrieval practice and prompting deep, meaningful processing. And contrary to what might be assumed, study methods that involve self-testing don’t make pupils more anxious - they actually reduce anxiety, as learners become more confident in their own knowledge (Agarwal et al., 2014).

Conclusion

Memory plays a key role throughout learning, but can’t be reduced to the memorisation of isolated facts - our memories are interrelated, and depend on deep and active processing. Cognitive scientists are now able to recommend educational techniques that can help with developing durable, well-structured memories that will last over the long term, and to discourage the use of more inefficient methods like re-reading and overlearning. Evidence-based teaching which makes use of retrieval practice, the spacing effect and interleaving can help with memory as well as with transfer to new situations, helping pupils over the longer term.

Key messages

- Our memories are interrelated and are the basis of developing an understanding, not just learning individual facts.

- Evidence-based teaching which makes use of retrieval practice and interleaving can help with memory, and with transfer to new situations.

- Some traditional education practices such as ‘overlearning’ are not supported by evidence from cognitive science.

References

Agarwal PK, D’Antonio L, Roediger III HL, McDermott KB, and McDaniel MA (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.

Carvalho PF and Goldstone RL (2015) What you learn is more than what you see: what can sequencing effects tell us about inductive category learning? *Frontiers in Psychology, 6*: 505.

Dunlosky J and Rawson KA (2015) Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students’ learning. *Scholarship of Teaching and Learning in Psychology, 1*(1): 72–78.

Hyde TS and Jenkins JJ (1973) Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior, 12*(5): 471–480.

Karpicke JD, Lehman M and Aue WR (2014) Retrieval-based learning: An episodic context account. In: Ross B (ed) *Psychology of Learning and Motivation, Vol 61* (pp. 237–284). Waltham, MA: Academic Press.

Moll LC and Gonzalez N (1994) Lessons from research with language-minority children. *Journal of Reading Behavior, 26*(4): 439–456.

Rawson KA, Thomas RC and Jacoby LL (2015) The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review, 27*: 483–504.

Rohrer D and Taylor K (2006) The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*: 1209–1224.

Appendix 2: PROSPERO Registration

This is the record of the pre-registration of the systematic review (Chapter 3). Reference: Jonathan Firth, Ian Rivers, James Boyle. A systematic review of interleaving as a concept learning strategy. PROSPERO 2018 CRD42018093814 Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42018093814

A systematic review of interleaving as a concept learning strategy

by J. Firth, I. Rivers and J. Boyle

Review question

For a population of learners in mainstream education, is presenting examples of concepts in an interleaved order a more successful learning strategy than presenting examples blocked by topic, in terms of learners' ability to remember examples and transfer to novel examples.

Searches

One researcher will conduct data extraction, using the PsycINFO, Web of Science, BEI, AEI and ERIC databases. Search terms will focus on the research variable interleaving (interleav*, with possible synonyms "contextual interference", shuffl*, intermix*) and on the outcome variable (learning/"conceptual knowledge", inducti*), for records from 2008-present. Where necessary, database journal categories will be used to exclude items from irrelevant domains or on the basis of exclusion criteria below. Other search methods will include hand searching of relevant journals, and reference chasing from existing narrative review articles Rohrer (2012), Carvalho & Goldstone (2015), and Kang (2016).

Types of study to be included

Inclusion criteria are as follows:

1. Age 13-65, must be a neurotypical sample.

2. Experimental or quasi-experimental designs only.

3. Studies must have collected primary data.

4. One of the primary research variables must be interleaving as it relates to learning/memory/understanding.

In addition to excluding studies that don't meet all four inclusion criteria, two further exclusion criteria will be applied:

5. Exclude neurological/fMRI-based studies.

6. Exclude studies with outcome variables that don't directly demonstrate concept learning (e.g. studies of attention, mind-wandering, visual perception, motor learning, language learning).

Inclusions will be cross-checked with the other researchers. Any discrepancy will be solved via discussion. Domain (visual v's verbal, and relevance to specific subject discipline) will be assessed.

Condition or domain being studied

The term interleaving refers to immediate variation within a set of tasks or example items, whereby each item is immediately followed and preceded by an example of a different category/concept rather than appearing in blocks of the same type of item repeatedly (which is termed a 'blocked' arrangement). It could arise due to a randomisation or 'shuffling' of the order of items, or a more deliberate alternation of items. Interleaving has been investigated in maths learning, physical skill learning and the inductive learning of patterns and pictures. It has the potential to benefit the learning of concepts in a way that transfers to future learning and use. This review will investigate evidence for its use in concept learning in education-relevant contexts.

Participants/population

Neurotypical population, representative of those in formal education at school or university. This will comprise adolescents or adults, but excluding older adults (65+) on the basis that this population may have memory issues that differ from the bulk of students in education.

Intervention(s), exposure(s)

Interleaving. This is defined as mixing or alternating the order of presentation such that examples of concept ‘A’, concept ‘B’ an concept ‘C’ would be presented in an order such as ABCABCABC.

Comparator(s)/control

Blocked presentation, which involves concepts being presented via multiple examples of the same concept (e.g. AAA, BBB, CCC for the concepts labelled above). Other than the order of presentation, control conditions should be identical to the intervention condition to avoid confounding variables.

Context

Formal education, but studies on equivalent adults from the general population (e.g. recruited via Amazon Mechanical Turk) need not be excluded.

Main outcome(s)

Research studies must measure learning via memory recall (correctly identifying trained examples), or transfer (correctly identifying novel examples), or both. This has a direct connection with health in two main domains. One is the mental health of school/university pupils at exam time, where it has been shown that cognitive processes interact with anxiety levels. The other is the psychological wellbeing of adults who are experiencing memory loss. Relevant studies on these health outcomes will be eligible for inclusion in the review, provided that other criteria are also met.

Additional outcome(s)

None

Data extraction (selection and coding)

One researcher will conduct the initial selection, and inclusions will be cross-checked with the other researchers. Any discrepancy will be solved via discussion. Key data to be included are name and year of study, experiment number (where there are multiple experiments in the same publication), outcome measure (memory v's transfer), mean score and SD for each condition, F values or t values, population, sample size for each condition, domain tested (visual v's verbal v's other), relevance to specific subject discipline (e.g. maths, science). These will be assessed again by the first researcher, and coding will be cross-checked among the research team.

Risk of bias (quality) assessment

Possible sources of bias include participant demand characteristics, incomplete or selective reporting of outcomes, lack of sample diversity, conflicts of interest, other. These will be assessed again by the first researcher, and judgements will be cross-checked among the research team. A funnel plot will be used to assess publication bias.

Strategy for data synthesis

For meta-analysis, we will use Cohen’s d, or calculate this via F values.

The robustness of findings could be affected by publication bias, and this will be assessed via funnel plot visual inspection

It is unlikely that experimental studies will have very small sample sizes, but if so, Hedge’s g will be used to minimise bias from such studies.

Relevance to specific subject disciplines will be assessed via descriptive statistics.

Analysis of subgroups or subsets

We plan will conduct subgroup analyses in order to evaluate the impact of moderators on pooled effect sizes, including the key research question (transfer v’s memory), type of category studied (which will be divided by verbal v’s pictorial), and interleaved learning v’s practice.

Contact details for further information

Jonathan Firth

jonathan.firth@strath.ac.uk

Organisational affiliation of the review

University of Strathclyde

www.strath.ac.uk

Review team members and their organisational affiliations

Mr Jonathan Firth. University of Strathclyde

Professor Ian Rivers. University of Strathclyde

Professor James Boyle. University of Strathclyde

Type and method of review

Intervention, Meta-analysis, Systematic review

Anticipated or actual start date

23 April 2018

Anticipated completion date

30 June 2018

Funding sources/sponsors

None

Conflicts of interest

None

Language

English

Country

Scotland

Stage of review

Review Ongoing

Subject index terms

Concept Formation; Humans; Learning

Date of registration in PROSPERO

20 June 2018

Date of publication of this version

20 June 2018

Details of any existing review of the same topic by the same authors

n/a

The record owner confirms that the information they have supplied for this submission is accurate and complete and they understand that deliberate provision of inaccurate information or omission of data may be construed as scientific misconduct.

The record owner confirms that they will update the status of the review when it is completed and will add publication details in due course.

Versions

20 June 2018

Appendix 3: Systematic Review Protocol Article

The following article was prepared and published in as follows. Reference: Firth, J., Rivers, I., & Boyle, J. (2019). A systematic review of interleaving as a concept learning strategy: A study protocol. Social Science Protocols, July 2019, 1-7. http://dx.doi.org/10.7565/ssp.2019.2650. I am very grateful to journal editor Dr. Laura A. Cariola of the University of Edinburgh for her suggested revisions.

A systematic review of interleaving as a concept learning strategy: A study protocol

by J. Firth, I. Rivers and J. Boyle

Abstract

Background: Education Scotland’s (2018) framework for interventions for equity supporting the Scottish Attainment Challenge highlights the promotion of high quality learning and the effective use of evidence and data. This study protocol outlines the methodology of a systematic review of the literature into the use of interleaving to facilitate the effective learning and teaching of new concepts. *Methods:* The systematic review has been pre-registered with PROSPERO, an international database of prospectively registered systematic reviews. The review will investigate whether presenting examples of to-be-learned concepts in an interleaved order is a more effective learning strategy than presenting examples blocked by topic, in terms of learners’ ability to remember examples and to transfer learning to novel examples. *Discussion:* Interleaving is widely recommended as an evidence-based approach to teaching with considerable potential as a strategy for learners experiencing difficulties in working memory functioning and conceptual learning, but to date there has not been a comprehensive review of the evidence base. The review will address this gap. It will synthesize primary research studies from the past decade, investigate boundary conditions and variables that interact with interleaving, and will include a meta-analysis of recent studies. This protocol provides the details of the rationale of the review, and details the inclusion criteria and approaches to data extraction.

Keywords: Interleaving, memory, transfer, concept learning, education, spacing, attainment challenge, working memory

Background

The term interleaving refers to variability within a set of tasks or example items such that each item is immediately followed and preceded by an example of a different type or concept rather than appearing in blocks of the same type of item repeatedly; the latter is termed a 'blocked' sequence. For example, in Figure 1, presentations of item types '1', '2' and '3' are shown blocked (Example A), or with the three item types interleaved (Example B):

Example A: blocked sequence: 111112222233333

Example B: interleaved sequence: 123123123123123

Figure 1. Comparison of interleaved and blocked sequences.

The benefit of interleaving is sometimes termed the interleaving effect, and has been investigated in numerous contexts, many of which are of direct relevance to education. It has been found to benefit maths learning (e.g., Rohrer et al., 2015), the conceptual learning of science categories and examples (Eglington & Kang, 2017; Rawson et al., 2015), and the inductive learning of images of animal species and modern art paintings (Birnbaum, Kornell, Bjork & Bjork, 2013; Kornell & Bjork, 2008). If learners were being taught about species of bird, for example, blocking would involve showing multiple examples of the same species of bird consecutively, while interleaving would involve showing examples of one species of bird, then another, then yet another, and so on.

Interleaving has been described as a desirable difficulty, in that it can make learning slower at first but more durable over the long term (Yan, Bjork & Bjork, 2016). Its benefits were discovered by William Battig in the 1960s via his research into the learning of word pairs, and he believed that the interference caused by interleaving can make learning more resilient — an idea consistent with the role of varied environmental contexts in learning (e.g., see Smith, Glenberg & Bjork, 1978). As such, Battig (1979) tended to describe the effect in terms of 'contextual interference', a term which emphasises the role of the broader context rather than just the interleaved items themselves (Battig, 1979; Magill & Hall, 1990). This term is still widely used in the domain of motor learning. However, recent research into concept learning has suggested that it may be the relationship between interleaved items that leads to the interleaving benefit rather than simply the interference which results from the format, with interleaving making it easier for learners to compare and contrast items and thereby notice subtle conceptual differences between them (Birnbaum et al., 2013). In one study which helped to revive recent interest in applying interleaving to education, Kornell and Bjork (2008) found that interleaving sets of artwork by different artists led to more effective learning than spending the same amount of time looking at blocked examples of paintings by the same artist. Viewing paintings by different artists in an interleaved order improved learners' ability to later identify the style of these artists as indicated by their ability to correctly identify novel example paintings, perhaps due to the greater ease of learning when contrasting items are seen consecutively — an idea known as the discrimination-contrast hypothesis (Birnbaum et al., 2013). This fits with a body of research into categorisation which tends to show that highlighting differences (i.e. making discriminative contrast more salient) has a more beneficial effect on category learning than highlighting similarities (Higgins & Ross, 2011).

Supporting this idea, Hausmann and Kornell (2014) mixed the study of Indonesian vocabulary with the learning of biology terms, and did not find a benefit of interleaving, presumably because the two sets of material were too conceptually distant to productively be compared or organised. Similar items do tend to show a benefit of interleaving, especially when the difference between them is subtle (Carvalho & Goldstone, 2014), possibly because subtle differences would be very hard for learners to notice if viewed during separate study sessions. Interestingly, though, Eglington and Kang (2017) did not find that explicitly highlighting differences during the learning phase impacted on the benefits of interleaving compared to blocking.

There is an unavoidable connection between interleaving and the spacing effect (i.e. distributed practice), and the two are often conflated. This is because interleaving items inevitably increases the gaps between one example item and the next (as can be seen in Figure 1). In Kornell and Bjork's (2008) study of interleaving and modern art paintings, the researchers initially attributed their findings to spacing, an effect which was already well established at the time (e.g., see Dempster, 1996). However, Kang and Pashler (2012) carried out a replication where spacing was held constant, using filler images to increase the temporal space between one blocked item and the next, and found that the interleaved condition was nevertheless superior to the spaced or blocked conditions. In a similar study which used trivia questions as filler items, Birnbaum et al. (2013) concluded that the benefit to inductive learning of visual items was largely due to interleaving rather than spacing, while Taylor and Rohrer (2010) found that for mathematics practice, both spacing and interleaving have separate beneficial effects. It could be argued that failing to control for spacing (as was typical in studies prior to 2012) leads to a confound between the two variables.

Interleaving and spacing therefore tend to be seen as separate phenomena today, and both are widely recommended by sources which advocate applying cognitive science to classroom teaching, for example the UK's Chartered College of Teaching (https://chartered.college), the Learning Scientists (learningscientists.org), as well as numerous recent books on teaching practice.

A practical difference between spacing and interleaving is that rather than re-studying the same material on separate occasions, the procedure of interleaving studies tends to involve presenting several different examples of a category during a learning phase, and then presenting novel items during a test phase (instead of or as well as testing memory for the original items). This has educational implications; spacing is likely to be beneficial when items studied are exactly the same, such as for the practice of foreign language vocabulary. Interleaving, in contrast, stands to benefit knowledge transfer following the learning of varied prior examples. It is likely to be useful in situations where learners may have to identify novel instances of previously studied concepts, such as identifying signs of glaciation in a previously unseen landscape, or recognising social psychology phenomena during an everyday encounter.

Although the potential of interleaving to help learners to compare and contrast exemplars has been noted, another theoretical explanation of the benefit is that learners tend to pay more attention to interleaved items. This is known as the attention-attenuation hypothesis (Wahlheim, Dunlosky & Jacoby, 2011). It is an idea which fits with evidence that blocked presentations tend to lead to more mind-wandering (Metcalfe & Xu, 2016). It also fits with recent findings that working memory capacity does not play a major role in the interleaving effect (Guzman-Munoz, 2017; Sana, Yan, Kim, Bjork, & Bjork, 2018) — if contrast between current and previous items is the key factor in the effect then a larger working memory capacity should increase the benefit, but this does not appear to be the case. This finding suggests that interleaving could generalise to younger learners whose working memory is still developing, or to pupils with additional support needs or adults with impaired working memory.

The present study is a systematic review and meta-analysis of research into interleaving, focusing on the work over the ten years since Kornell and Bjork's seminal study of interleaved learning of art paintings. It aims to provide a much-needed overview of the evidence base for interleaving as an educational technique, as well as an indication of the effect size (if any) of the technique as an intervention, and any important interactions that may emerge. Given that many of the studies cited thus far were conducted on undergraduate populations and with specific tasks (art, maths, etc), the review will also aim to shed light on whether the evidence base – such as it is – can support recommendations for applying interleaving to other educational domains. It may also give an indication of what kind of school tasks are most likely to benefit from this intervention, and an idea of where further research is needed. The review question is as follows: For a population of learners in mainstream education, is presenting examples of concepts in an interleaved order a more successful learning strategy than presenting examples blocked by type, in terms of learners' ability to remember examples and transfer to novel examples?

2. Methods/Design

2.1 Searches

One researcher will conduct data extraction, using the PsycINFO, Web of Science, BEI, AEI and ERIC databases. Search terms will focus on the research variable interleaving (interleav*, with possible synonyms "contextual interference", shuffl*, intermix*) and on the outcome variable (learning/"conceptual knowledge", inducti*), for records from 2008-present. Where necessary, database journal categories will be used to exclude items from irrelevant domains or on the basis of exclusion criteria below. Other search methods will include hand searching of relevant journals, and reference chasing from existing narrative reviews by Rohrer (2012), Carvalho & Goldstone (2015), and Kang (2016).

2.2 Domain and context

The focus will be on learning and memory. More specifically, the review will focus on the application of interleaving as an intervention for learning tasks, such that it could be applied to schools or other educational contexts. The review will focus on studies where participants are in formal education, but studies on equivalent adults from the general population (e.g., recruited via Amazon Mechanical Turk) need not be excluded.

2.3 Participants

As a first step, the review will look at studies of mainstream education populations, representative of those in formal education at school or university. This can comprise adolescents or adults or both, but will exclude older adults (65+) on the basis that this population may have memory issues that differ from those experienced by the bulk of students in education.

2.4 Intervention and comparator

The review will search for studies of interleaving versus blocking, where interleaving refers to immediate variation within a set of tasks or example items, whereby each item is immediately followed and preceded by an example of a different category/concept, such that examples of concept ‘1’, concept ‘2’ and concept ‘3’ would be presented in an order such as 123123123. Interleaving could arise either due to a randomisation or ‘shuffling’ of the order of items, or a more deliberate alternation of items.

Blocked presentation involves concepts being presented via multiple examples of the same concept (e.g., 111, 222, 333 for the concepts labelled above). It is defined as presentations where studied items appear in blocks of the same type of item repeatedly.

Other than the order of presentation, control conditions should be identical to the intervention condition to avoid confounding variables; studies where spacing is a potential confound will be included, but the potential effect of this issue will be considered when reviewing the evidence.

2.5 Types of study to be included

Inclusion criteria are as follows:

- • Age 13-65, must be a typically (or assumed typically) developing sample without known memory problems.
- • Experimental or quasi-experimental designs only.
- • Studies must have collected primary data.
- • One of the primary research variables must be interleaving as it relates to learning/memory/understanding.

In addition to excluding studies that don't meet all four inclusion criteria, two further exclusion criteria will be applied:

- • Exclude neurological/fMRI-based studies.
- • Exclude studies with outcome variables that don't directly demonstrate concept learning (e.g., studies of attention, mind-wandering, visual perception, motor learning, language learning).

Inclusions will be cross-checked within the research team. Any discrepancies will be solved via discussion. Domain (visual v's verbal, and relevance to educational contexts) will be assessed.

2.6 Main outcome(s)

Research studies must measure learning via memory recall (correctly identifying trained examples), or transfer (correctly identifying novel examples), or both.

It is notable that such results do not exist in isolation in the real world, but instead link to other domains. One is the mental health of school/university pupils at exam time, where it has been shown that cognitive processes interact with anxiety levels. The other is the psychological wellbeing of learners who are experiencing memory difficulties. Relevant studies on these health outcomes will be eligible for inclusion in the review, provided that other criteria are also met.

2.7 Data extraction

Key data to be included are name and year of study, experiment number (where there are multiple experiments in the same publication), design (within or between participants), outcome measure (memory v's transfer), mean score and SD for each condition, F values or t values, population, sample size for each condition, domain tested (visual v's verbal v's other), relevance to specific subject discipline (e.g., maths, science). These will be assessed again by the first researcher, and coding will be cross-checked among the research team.

2.8 Risk of bias (quality) assessment

Possible sources of bias include participant demand characteristics, incomplete or selective reporting of outcomes, lack of sample diversity, or conflicts of interest. These will be assessed again by the first researcher, and judgements will be cross-checked among the research team. Funnel plots, fail-safe N, Begg and Mazumdar's Test and Eggers' Test (see Borenstein et al., 2014) will be used to assess publication bias.

2.9 Strategy for data synthesis

For meta-analysis, we will use we will calculate Hedge's *g* standardised effect size, which minimises bias from studies with small sample size. Representation of tasks which link to specific-subject education disciplines will also be assessed.

2.10 Analysis of subgroups

We will conduct subgroup analyses in order to evaluate the impact of moderators on pooled effect sizes, including the key research question (transfer v's memory), type of category studied (which will be divided by science/factual v's art/images), and within v's between participant designs.

3. Discussion

This PROSPERO-registered protocol outlines the search and analysis strategy for a systematic review into interleaving as it applies to education. Conducting the review will provide a much clearer picture of the interleaving effect, including some of the variables with which interleaving may interact, and any relevant boundary conditions. It will help to clarify whether interleaving is useful as a memory intervention, or if it is beneficial for transfer of prior learning to new contexts, or both.

At a time when interleaving is increasingly being advocated as a strand of evidence-based teaching practice, it will be useful to gain an objective summary of the evidence for the effect, and also to investigate its potential as a strategy to support learners who struggle with concept learning, or who have reduced working memory. This will include finding out what type of school-based tasks it appears to be best suited to, with a view to supporting pupils with additional support needs.

Interleaving promotes the contrast of real-world examples and thereby facilitates learners' induction of new concepts. Given its potential for use with visual examples, interleaving could boost the development of conceptual knowledge learning among pupils with reading disabilities, thereby supporting their subsequent understanding of texts (Willingham, 2006). It could also be a suitable technique for helping pupils who struggle with traditional approaches to inductive learning such as discovery or problem-based learning, due to the working memory demands of such tasks (Kirschner et al., 2006).

References

Battig, W. F. (1979). The flexibility of human memory. In L. S. Cermak, & F. I. M. Craik (Eds), *Levels of processing and human memory.* Hove, UK: Lawrence Erlbaum.

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.

Borenstein, M., Hedges, L. V., Higgins, J. T., & Rothstein, H. R. (2014). *Comprehensive Meta Analysis Version 3*. Englewood, NJ: Biostat.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495.

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*(1), 281–288.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 318-339). San Diego, CA: Academic Press.

Education Scotland (2018). Interventions for equity. Retrieved from *https://education.gov.scot/improvement/self-evaluation/ Interventions%20for%20Equity*

Eglington, L. G., & Kang, S. H. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 475–485.

Guzman-Munoz, F. J. (2017) The advantage of mixing examples in inductive learning: a comparison of three hypotheses. *Educational Psychology, 37*(4), 421–437.

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in memory and Cognition, 3*(3), 153–160.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1388–1393). Austin, TX: Cognitive Science Society.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*(1), 97–103.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86.

Kornell, N. & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science, 19*, 585–592.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science*, *9*(3–5), 241–289.

Metcalfe, J., & Xu, J. (2015). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 978–984.

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review, 27*(3), 483–504.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*, 900–908.

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit?. *Journal of Applied Research in Memory and Cognition, 7*(3), 361–369.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342–353.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*(6), 837–848.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750–763.

Willingham, D. T. (2006). How knowledge helps. *American Educator, Spring*, 30-37.

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. J. C. Horvath, J. Lodge, & J. A. C. Hattie (Eds). *From the laboratory to the classroom: Translating the learning sciences for teachers*. Abingdon, UK: Routledge.

Appendix 4: Blog post, 'What is evidence-based education'

This blog post was written for and posted on my own blog (May 30 2017) *and shared with educators in an attempt to contribute critically to the debate over evidence-based educators.*

What is evidence-based education? A primer on the 'what works' debate, with key sources and a discussion of its pros and cons.

by J. Firth

I recently joined and met with SURE - 'School and University Research Enquiry', a research group which has put several schools in the Glasgow area in contact with the University of Strathclyde in order to exchange knowledge and conduct new research. The ultimate aim is to promote a more evidence-informed approach to educational decision making and practice.

With this in mind, I thought it would be helpful to write a brief overview of the field of evidence-based education, including some of the main publications and debates.

What is it?

Firstly, evidence-based education is the idea that research of various kinds should be used to inform decisions about teaching and learning. It is conceived of as an alternative to teaching practice that is guided by intuition and/or experience.

An educator's job includes a huge amount of decision making. For example, what should be taught today? What about tomorrow? What type of homework should be set, and when?

How can a teacher maintain discipline effectively, and engage their pupils? Evidence-based education aims to tackle these questions pragmatically on the basis of past findings, and is sometimes referred to as a ‘what works’ approach.

Focusing on the example of homework, a traditional view might be that the teacher should allocate whatever they judge to be useful, or whatever is just ‘the way it’s done’ (or whatever is lying around the office, is quick to mark, or is in the textbook/revision guide!). An evidence-based alternative would be to look at this issue from the perspective of research which has shown that some strategies lead to more effective/durable learning than others - the cognitive psychology of memory tells us that learners remember more if there is a delay before they practice material that they have mastered in class, and that they remember more if they do a closed-book test rather than copying from notes. The teacher may therefore decide to set a practice test, and to do so after a one-week delay rather than on the same day as the material was done in class.

The above example relates to memory, and the what works approach as a whole usually refers to techniques or interventions that boost attainment (as measured by some form of test or exam), but evidence could inform many other types of decision too. For example, when considering an issue such as student motivation, evidence could be evaluated to help determine the most effective way to proceed.

As a model, this borrows from the philosophy behind evidence-based medicine. We would probably take it for granted that a doctor should select a treatment that has been shown by reliable (and replicated) research to be the most effective, rather than being guided by tradition (leeches, anyone?) or their individual gut feeling about what ought to work. In the same way, it is argued, teachers should look to the evidence rather than relying on their personal preferences or even on classroom experience. Insisting on evidence may have the incidental advantage of making educational practices less vulnerable to fads, such as the learning styles myth.

Sounds great! So everyone agrees with this…?

No! It has many critics, and their points are well worth taking on board. Firstly, the idea that education can derive a model of effective practice from medicine is open to doubt. Learning is not really like curing an illness - it's cumulative, has no clearly defined end point, and there are important subtleties such as how well it can be transferred to new situations. The entire approach could therefore be seen as over-simplistic.

Secondly, what works for one group might not work for all. To take one example, Kalyuga (2007) has described the 'expertise reversal effect' whereby tasks that are effective with beginners become ineffective or at least inefficient when used with more advanced learners. Another example, much discussed in recent years, is that homework appears to be more effective for secondary students than for primary (Cooper et al., 2006). This is not a killer blow to the idea of evidence-based practice, but it does suggest that the use of evidence must be cautious and thoughtful - we can't apply one-size-fits-all solutions.

Thirdly, there are concerns about the validity of some of the evidence used. Education is a notoriously tricky area to research - for ethical reasons it is often necessary to rely on correlations and secondary data, leaving some findings open to confounding variables. Meanwhile, a lot of the research evidence from cognitive psychology in areas such as working memory and learning is based on laboratory studies with university students. That doesn't make it inherently bad research, but does mean that we should be cautious about generalising it to school pupils.

Finally - and linked to the previous point - some people argue that the evidence referred to in this approach is often positivist in its underlying scientific philosophy, whereas many educators and learning researchers subscribe to a social constructivist view of learning.

Key literature

There is a lot of literature in this field, including both empirical research studies and reviews. For anyone who is new to this area, these are a few very useful publications to get you started. In the main they come from proponents of the idea, but I've also included some key critiques:

American Psychological Association's 'top 20' ways to apply psychology in the classroom

Broader than most, the APA's guide includes such issues as creativity, classroom management, and growth mindset, as well as strategies that impact on learning more directly.

Biesta (2007)

Gert Biesta here criticises evidence-based practice and also questions the broader assumption that closely-controlled lab work has ever contributed much to society (!). He argues that it tends to link to top-down approaches where administrators and governments say that strategies work on the basis of lab research, when they may not work in a specific context.

Additionally, the notion of something working doesn’t address philosophical issues of who it works for, and to what social end.

Coe et al. - the Sutton Trust report

Coe et al. (2014)’s report ‘What makes Great Teaching?’ is useful in that it goes beyond the cognitive evidence and considers such issues as classroom climate, teacher knowledge levels, and how teachers can improve. Otherwise, it draws on a similar body of research to Dunlosky et al. (2013; see below). The Sutton Trust also back the Education Endowment Foundation’s ‘Teaching and Learning Toolkit’, which provides a useful (if rather undiscriminating) visual guide to evidence-based strategies in terms of cost, lasting impact and the security of the supporting research.

Dunlosky et al. (2013)

The authors are psychologists and memory researchers, and this paper reviews a number of different findings from cognitive psychology. In particular, it endorses the use of retrieval practice (the ‘testing effect’) and distributed practice (the ‘spacing effect’), while noting that techniques such as re-reading and highlighting are generally ineffective as study strategies.

Hattie’s taxonomy

Australian researcher John Hattie is probably the biggest name in this field; he has synthesised numerous meta-analyses of educational research and built up a list of interventions together with their average statistical effect size. He takes an effect size of 0.4 as a ‘hinge point’ above which interventions fall into (roughly) the top half, i.e. they are among the more effective interventions - but the higher the effect size, the better. The work is also helpful in identifying some interventions that have tended not to make a large impact. It has its flaws, both conceptual and statistical, but it’s a useful starting place for finding out about several important strategies.

The Learning Scientists

An excellent blog run by four cognitive psychologists who study learning and memory. It is aimed at students and teachers, and makes the science highly accessible without dumbing it down.

Marzano’s top ten

It’s useful to be aware of the work of Marzano et al. (2001), one of the earlier evidence-based summaries of effective teaching interventions. The strategies they endorse include analogies and metaphors, student-generated study notes, and feedback/formative assessment. There have been important new findings and some of the key research questions have moved on a bit since it came out, however, so it is a bit dated.

NCEE

The National Center for Education Evaluation and Regional Assistance (NCEE) in the USA offers the ‘What Works Clearinghouse’. It usefully reviews studies of efficacy in terms of learning, but the focus tends to be on large-scale programmes, for example the “Great Explorations in Math and Science® (GEMS®) Space Science Sequence” curriculum, rather than on specific techniques that teachers could use in class. This makes their findings less immediately applicable.

Zhao (2017)

In his paper ‘What works may hurt’, Zhao refers back to the analogy of evidence-based medicine and borrows a further concept - that of side effects. From this perspective, an intervention may ‘work’ from a learning point of view, but it could have any number of side effects. Just as with a drug, any benefits must be evaluated in that context. For example, an intervention that boosts learning over the short-term could also harm motivation over the longer term.

Is all of this a threat to teachers?

It is worth considering: does all of this amount to self-proclaimed experts telling us what to do (or what not to do)? At times that might be a valid concern, but the entire nature of making education more evidence based is that that evidence is (or can be) open to scrutiny. You may not agree with all of the conclusions from the sources above, but their arguments are probably backed up by a more thorough factual base than the opinion of a staffroom colleague. And if you are unsure, then you are free to scrutinise and evaluate the sources.

A problem, certainly, lies with teachers’ access to information. If teachers can’t or won’t access the evidence themselves, this puts a lot of power in the hands of central institutions who may try to push inappropriate programmes and interventions. Teachers (and schools more broadly) are in a stronger position to ward this off if they not only learn about the evidence but are also aware of its limitations.

For this to happen, practitioners require journal access, CPD time, and also the skills to critique the research methods and statistics used. How can that be achieved? This BERA report sets out a vision of schools and colleges as "research-rich environments in which to work" (p.5). It’s a radical idea, and one that asks us to reconsider the very nature of what teacher professionalism involves.

References

Biesta, G. (2007). Why “what works’’ won’t work: Evidence-based practice and the democratic deficit in educational research. Educational Theory, 57(1), 1–22.

Coe, R., Aloisi, C., Higgins, S. and Major, L.E. (2014). What makes great teaching? Review of the underpinning research. Accessed 14 May 2017 at http://www.suttontrust.com/wp-content/uploads/2014/10/What-makes-great-teaching-FINAL-4.11.14.pdf

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. Review of Educational Research, 76(1), 1–62.

Hattie, J. (2013). Visible learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement. London: Routledge.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educ Psychol Rev, 19, 509–539. doi: 10.1007/s10648-007-9054-3

Marzano, R. J., Pickering, D., & Pollock, J. E. (2001). Classroom Instruction That Works: Research-Based Strategies for Increasing Student Achievement. Alexandra, V.A.: ASCD.

Zhao, Y. (2017). What works may hurt: Side effects in education. Journal of Educational Change, 18(1), 1–19.

Appendix 5: Systematic review, table of all studies by constituent experiments (n = 56), with materials, design, and WoE ratings.

Author/study	WoE A quality	WoE B method	WoE C utility	Overall weight of evidence (WoE D)	Materials used	Number of categories per condition (with exemplars used in study phase)	Comments on methodology and/or findings
Birnbaum (2013) Expt 1a	High	High	Medium	High	Images of butterflies	16 (4)	This is the same study as published in Birnbaum et al., 2013 (Expt 3)
Birnbaum (2013) Expt 1b	High	High	Medium	High	Images of butterflies	16 (4)	Spacing was beneficial, in addition to contrast. However, the benefits of spacing may have an upper limit in terms of inter-item delay becoming too difficult.
Birnbaum (2013) Expt 2	High	High	Medium	High	Art images (paintings)	12 (6)	Control condition was replication of Kornell & Bjork Expt 1a. Other condition included pre-training of category names.
Birnbaum (2013) Expt 3	High	High	Medium	High	Art images (paintings)	6 (6)	Focused on incidental learning, with reduced number of categories. Interleaving benefited incidental learning of categories.
Birnbaum (2013) Expt 4	High	High	Medium	High	Art images (paintings) and eras (impressionism, romanticism, renaissance, and baroque)	12 (5) but grouped into 4 superordinate categories.	Incidental learning of categories but directed learning of superordinate categories (art eras). Separate-blocked led to best results, perhaps because art eras have high within-category variability.

Birnbaum (2013) Expt 5[38]	High	High	High	High	Images of butterflies	8 (4)	Testing for transfer of compare/contrast learning strategy to a different interleaved task. Interleaved superior, but no improvement or decline across parts 1 & 2.
Birnbaum et al. (2013) Expt 1	High	High	High	High	Images of birds	8 (4)	Used materials taken from Wahlheim et al. (2011) but a different number of categories; inserted trivia to interfere with contrast and this reduced interleaving benefit.
Birnbaum et al. (2013) Expt 2	High	High	High	High	Images of butterflies	16 (4)	Introduced spacing as well as interleaving via filler materials, and found that for spaced out items, interleaving and blocking were equivalent, but interleaving was superior with contiguous items.
Birnbaum et al. (2013) Expt 3	High	High	High	High	Images of butterflies	16 (4)	Showed items in groups whereby space didn’t interrupt contrast. Large spacing was superior to small spacing.
Carvalho & Goldstone (2014a) Expt 1	High	High	Medium	High	‘Blob’ figures	3 (8)	Compared similarity of categories; low similarity led to a blocking advantage, while high similarity led to an interleaving advantage.
Carvalho & Goldstone (2014a) Expt 2	High	High	Medium	High	‘Blob’ figures	3 (8)	Blocking advantage disappears without memory component, i.e. with simultaneous presentation of new and old items.
Carvalho & Goldstone (2014a) Expt 3	High	High	Medium	High	‘Blob’ figures	3 (8)	Compares simultaneous and successive presentation for low-similarity categories only.
Carvalho & Goldstone (2014b) Expt 1a	High	High	Medium	High	‘Blob’ figures	3 (8)	Same mats as Carvalho and Goldstone (2014a); investigated 24-hour delay. Schedule and similarity level didn’t interact with delay.
Carvalho & Goldstone (2014b) Expt 1b	High	High	Medium	High	‘Blob’ figures	3 (8)	Removed ‘refresher’ study session and found very low recall rates.
Carvalho & Goldstone (2015b) Expt 1	High	High	Medium	High	“Fribble” objects	3 (4)	Interleaving beneficial for active learning but blocking superior for passive.

[38] Birnbaum (2013) Expts 6a and 6b did not meet inclusion criteria in that they did not test interleaving as an IV.

Carvalho & Goldstone (2015b) Expt 2	High	High	Medium	High	“Ziggerins”	3 (4)	Replicated findings of Expt 1 with different materials.
Carvalho & Goldstone (2017) Expt 1	High	High	Medium	High	“Alien” cartoon creatures	2 (9)	Varied five specific stimulus dimensions (arms, legs, eyes, mouth, and antenna of cartoon figure), and blocking led to classification when features stayed the same across examples.
Carvalho & Goldstone (2017) Expt 2	High	High	Medium	High	“Alien” cartoon creatures	2 (9)	Added delay to test how learning progresses. No overall memory differences, but blocking led to more encoding of similarities, interleaving to more encoding of differences.
Carvalho & Goldstone (2017) ext 3	High	High	Medium	High	“Alien” cartoon creatures	2 (9)	Eye-tracker technology suggested that interleaving led to more visual attention on items that differ from previous example.
Dobson (2011)	Medium	High	High	High	Verbal (physiology information)	2 (8-9)	Interleaving showed no benefit to memory for verbal scientific information.
Eglington & Kang (2017) Expt 1	High	High	High	High	Science images (chemical molecules)	5 (12)	Interleaving benefit extends to STEM categories (chemistry molecules).
Eglington & Kang (2017) Expt 2	High	High	High	High	Science images (chemical molecules)	4 (12)	Replicated interleaving benefit with more difficult categories; found no effect of intra-category similarity.
Eglington & Kang (2017) Expt 3	High	High	High	High	Science images (chemical molecules)	4 (12)	Replicated findings of Expt 2 with the differences highlighted in red.
Eglington & Kang (2017) Expt 4	High	High	High	High	Science images (chemical molecules)	8 (12)	Replicated findings with additional categories added.
Guzman-Munoz (2017) Expt 1	High	High	Medium	High	Art images (paintings)	12 (6)	Pilot study with just 23 students. Interleaving superior, correlation with working memory capacity small and non-significant. All three experiments refer to ‘spacing’ but the manipulation is interleaved.
Guzman-Munoz (2017) Expt 2	High	High	Medium	High	Art images (paintings)	12 (6)	Replicated Experiment 1 with larger sample and greater spacing obtained by showing pictures for longer; concluded that mixing advantage is mainly due to interleaving, not spacing.

Guzman-Munoz (2017) Expt 3	High	High	Medium	High	Art images (paintings)	12 (6)	Replicated Experiment 2 using arithmetic problems to add spacing. Interleaving superior, with a marginal interaction with working memory (slightly greater advantage for high WM individuals).
Kang & Pashler (2010) Expt 1	High	High	Medium	High	Art images (paintings)	3 (24)	Added temporal spacing between items, and found that this eliminated the interleaving effect.
Kang & Pashler (2010) Expt 2	High	High	Medium	High	Art images (paintings)	3 (10)	Replicated Expt 1 with fewer items and a ‘simultaneous’ condition, the latter was equally as effective as interleaving, and both were superior to blocking.
Kornell & Bjork (2008) Expt 1a	High	High	Medium	High	Art images (paintings)	12 (6)	Interleaved (they refer to ‘spacing’ but the manipulation is interleaved as later defined) presentation is superior.
Kornell & Bjork (2008) Expt 1b	High	High	Medium	High	Art images (paintings)	12 (6)	Replication of 1a with between-participants design.
Kornell & Bjork (2008) Expt 2	High	High	Medium	High	Art images (paintings)	12 (6)	Replication of 1a with participants simply asked to identify items as familiar/unfamiliar artist.
Kornell et al. (2010)	High	High	Medium	High	Art images (paintings)	12 (6)	Used same methodology as Kornell & Bjork in a study that also tested older adults.
Linderholm et al. (2016)	Medium	High	High	High	Exercise physiology texts.	n/a	There was no test of transfer/categorisation. Interleaving was beneficial for learning themes of a text, and was boosted by retrieval practice.
MacKendrick (2015) Expt 1	Medium	High	Medium	Medium	Art images (paintings)	20 (4)	Interleaving superior for low-between/high-within (LBHW) similarity categories, though just a non-significant trend in favour of interleaving for HBHW, suggesting memory is overloaded with 20 categories.
MacKendrick (2015) Expt 2	Medium	High	Medium	Medium	Art images (paintings)	20 (4)	Interleaving superior for both HBHW and LBLW categories where cues about key features were given.
Metcalfe & Xu (2016)	High	High	Medium	High	Art images (paintings)	12 (12, 15 or 18)	Novel selection of paintings; also looked at mindwandering via a probe, with more mindwandering reported in the blocked condition.
Noh et al. (2016) Expt 1	High	High	Medium	High	Maths images (patterns of lines and shapes)	4 (16)	Interleaving was better for ‘information integration’ categories, while blocking was better for ‘rule-based categories’.
Noh et al. (2016) Expt 2	High	High	Medium	High	Maths images (patterns of lines and shapes)	4 (16)	Added another dimension of complexity. The findings followed the same pattern as Experiment 1.

Rawson et al. (2015) Expt 2[39]	High	High	High	High	Verbal (psychology concepts)	10 (10)	Interleaving superior for both studied and novel examples, but effect disappeared if definitions of concept provided, with non-significant trends in the other direction.
Sana et al. (2017) Expt 1	High	High	High	High	Verbal (statistics concepts)	3 (6)	Interleaving superior to blocking for statistics problems, but difference disappears for Ps with the highest WM scores.
Sana et al. (2017) Expt 2	High	High	High	High	Verbal (statistics concepts)	3 (6)	Inserting delays (cartoons) between example problems was helpful for the blocking condition but harmful for the interleaving condition.
Sana et al. (2017) Expt 3	High	High	High	High	Verbal (statistics concepts)	3 (6)	Simultaneous presentation of problems was helpful to blocked sequences, perhaps due to reduced memory load making comparison easier.
Verkoeijen & Bouwmeester (2014)	High	High	Medium	High	Art images (paintings)	12 (6)	Replicated findings of Kornell & Bjork (2008, Expt 2).
Wahlheim et al. (2011) Expt 1	High	High	Medium	High	Animal images (birds)	12 (6)	Interleaving effect found; simultaneous presentation of examples did not have an effect.
Wahlheim et al. (2011) Expt 2	High	High	Medium	High	Animal images (birds)	12 (6)	Self-paced version of Expt 1, showed evidence of reduced attention on later trials.
Yan et al. (2017) Expt 1	High	High	Medium	High	Art images (paintings)	12 (6)	Test of hybrid blocked-to-interleaved schedule, interleaved-to-blocked schedule, and 'mini-blocks' where blocks are subdivided into two. Interleaving superior to blocking.
Yan et al. (2017) Expt 2[40]	High	High	Medium	High	Art images (paintings)	12 (12)	Extended Expt 1 with more options for gradation due to more exemplars; mini-blocks were now in four groups. Interleaving was better than blocking and was equivalent to both mini blocks and blocked-to interleaved. Interleaved-to-blocked was not superior to pure blocking.
Zulkiply (2013)	Medium	High	High	High	Verbal (psychological disorders)	6 (3)	Interleaving of aurally presented texts was beneficial over short and long term.
Zulkiply (2015)	High	High	Medium	High	Art images (paintings)	12 (6)	Same images as Kornell & Bjork (2008). Compared rule-based learning (via prior factual information about each artist) with inductive. Interleaving superior over both conditions.

[39] Rawson et al. Expts 1a, 1b and 3 did not meet inclusion criteria in that they did not test interleaving as an IV.

[40] Yan et al. Expts 3 & 4 did not meet inclusion criteria in that they did not test learning/retention as a DV.

Zulkiply & Burt (2013a) Expt 1	High	High	Medium	High	Art images (paintings)	12 (6)	Same images as Kornell & Bjork (2008). Insert 30s delay, and found that interleaving was better immediate than temporally spaced, though the difference disappeared by the 4th test block.
Zulkiply & Burt (2013a) Expt 2	High	High	Medium	High	Abstract digital images	12 (6)	Designed high-discriminability (HD) v's low-discriminability (LD) materials. Interleaving best with LD, but effect reversed with HD.
Zulkiply & Burt (2013b) Expt 1	High	High	Medium	High	Art images (paintings)	12 (6)	Same images as Kornell & Bjork (2008). Interleaving advantage persists over a long-term retention condition (7 days).
Zulkiply & Burt (2013b) Expt 2	High	High	High	High	Verbal (psychological disorders)	6 (3)	Interleaving advantage persists over a long-term retention condition (7 days), this time with verbal materials.
Zulkiply et al. (2012) Expt 1	High	High	High	High	verbal (psychological disorders)	6 (3)	Interleaving advantage generalises to visually presented texts.
Zulkiply et al. (2012)	High	High	High	High	verbal (psychological disorders)	6 (3)	Interleaving advantage generalises to aurally presented texts.

Appendix 6: Systematic review, table of items included in meta-analysis (n = 32)

No.	Study/experiment	Design (B = between, W = within)	Transfer/ memory?	Materials - category	Sample	WoE D	Effect size
1	Birnbaum, Kornell, Bjork & Bjork (experiment 2)	W	Transfer	Science (animal images)	114 undergraduates	High	0.44
2	Birnbaum (experiment 2)	W	Transfer	Art (paintings)	62 adults (MTurk)	High	0.33
3	Dobson (all – 1 experiment)	B	Memory	Science (verbal–biology)	189 undergraduates	High	0.17
4	Eglington & Kang (experiment 1)	B	Transfer	Science (chemistry images)	60 undergraduates	High	0.60
5	Eglington & Kang (experiment 2)	B	Transfer	Science (chemistry images)	60 undergraduates	High	0.61
6	Eglington & Kang (experiment 3)	B	Transfer	Science (chemistry images)	60 undergraduates	High	0.55
7	Eglington & Kang (experiment 4)	B	Transfer	Science (chemistry images)	60 undergraduates	High	0.71
8	Guzman-Munoz (experiment 2)	W	Memory	Art (paintings)	118 undergraduates	High	0.73
9	Guzman-Munoz (experiment 2)	W	Transfer	Art (paintings)	118 undergraduates	High	0.84
10	Guzman-Munoz (experiment 3)	W	Transfer	Art (paintings)	118 undergraduates	High	0.75
11	Kang & Pashler (experiment 1)	B	Transfer	Art (paintings)	88 undergraduates	High	0.74
12	Kang & Pashler (experiment 2)	B	Transfer	Art (paintings)	90 undergraduates	High	0.55
13	Kornell & Bjork (experiment 1a)	W	Transfer	Art (paintings)	120 undergraduates	High	0.80
14	Kornell & Bjork (experiment 1b)	B	Transfer	Art (paintings)	72 undergraduates	High	1.13
15	Kornell & Bjork (experiment 2)	W	Transfer	Art (paintings)	80 undergraduates	High	0.36
16	Kornell, Castell, Eich & Bjork	W	Memory	Art (paintings)	64 undergraduates	High	0.58
17	Kornell, Castell, Eich & Bjork	W	Transfer	Art (paintings)	64 undergraduates	High	0.58
18	MacKendrick	B	Transfer	Art (paintings)	120 undergraduates	Medium	1.65

	(experiment 1)						
19	Metcalfe & Xu	W	Transfer	Art (paintings)	66 undergraduates	High	1.08
20	Noh, Yan, Bjork & Maddox (experiment 1)	B	Memory	Science (maths images)	132 adults (MTurk)	High	0.35
21	Noh, Yan, Bjork & Maddox (experiment 2)	B	Transfer	Science (maths images)	132 adults (MTurk)	High	0.29
22	Rawson, Thomas & Jacoby (experiment 1b)	B	Memory	Science (verbal–psychology)	197 undergraduates	High	0.68
23	Rawson, Thomas & Jacoby (experiment 1b)	B	Transfer	Science (verbal–psychology)	197 undergraduates	High	0.87
24	Sana, Yan & Kim (experiment 1)	B	Transfer	Science (verbal–statistics)	126 undergraduates	High	0.68
25	Sana, Yan & Kim (experiment 2)	B	Transfer	Science (verbal–statistics)	137 undergraduates	High	0.35
26	Sana, Yan & Kim (experiment 3)	B	Transfer	Science (verbal–statistics)	135 undergraduates	High	0.43
27	Verkoeijen & Bouwmeester	W	Transfer	Art (paintings)	123 adults (MTurk)	High	0.37
28	Wahlheim, Dunlosky & Jacoby (experiment 1)	W	Memory	Science (animal images)	48 undergraduates	High	0.46
29	Wahlheim, Dunlosky & Jacoby (experiment 1)	W	Transfer	Science (animal images)	48 undergraduates	High	0.60
30	Zulkiply (2013)	W	Transfer	Science (verbal–psychology)	40 undergraduates	High	1.00
31	Zulkiply, McLean, Burt & Bath (experiment 1)	W	Transfer	Science (verbal–psychology)	40 undergraduates	High	0.53
32	Zulkiply, McLean, Burt & Bath (experiment 2)	W	Transfer	Science (verbal–psychology)	40 undergraduates	High	0.28

Appendix 7: Article on teacher reflections (Firth, 2020)

This piece was accepted for Impact, the professional peer-reviewed journal of the Chartered College of Teaching. It focuses on the common practice of classroom observation, applying the science of memory and metacognition to highlight how our reflections of what we observe and recall can often be flawed. This is the submitted draft prior to editing. Reference: Firth, J. (2020). Teacher classroom reflections – tackling flawed metacognition and memory. Impact, 8, 78-80.

Teacher classroom reflections – tackling flawed metacognition and memory

by J. Firth

"How do you think it went?" Reflection on one's own lessons is typically viewed as a key aspect of professional development, and a stepping stone towards effective teaching practice. Teachers are exhorted to be reflective practitioners, and this – it is assumed – will make them better at their job, a principle often seen as an instrument of system-wide improvement (e.g. Donaldson, 2011).

But how accurate are such reflections, particularly when it comes to the subtle processes of learning and memory? Can we rely on a teacher's ability to accurately judge learning when they see it, or even to correctly remember what actually happened in a lesson? To answer these questions, I will draw on two main areas of research. One is the research into metacognition. The other is the cognitive psychology involved in memory for events, specifically the flaws and biases affecting our recollections of what we have witnessed. My specific focus is on reflective discussions such as those between a teacher and their mentor or departmental manager, but the points largely apply to independent reflections, too.

Metacognition

What exactly is metacognition? Metacognition involves “thinking about thinking”, and the term is often used in education to refer learners’ ability to monitor and guide their own learning processes. In principle it can relate to any reflective aspect of cognitive function, such as thinking about problem solving, or the memory of another memory.

While we tend to focus on *learners’* cognitive processes about their learning, we can also consider *teacher* metacognition, including their beliefs about how learning works. How accurate are these beliefs; how closely does the teacher’s mental model of learning mirror reality? It’s an important question; by analogy, if a mechanic believed that all faults in a car engine were the result of divine intervention then they would be poorly equipped to suggest appropriate remedies. As Kornell and Bjork (2007) stated with respect to student study choices, “when metacognitive judgments are faulty, study decisions based on such judgments will be faulty as well” (p. 220). The same argument could be made about teacher decision making, too.

Perhaps surprisingly, there is evidence that educators do not always make accurate judgements about learning. Soderstrom and Bjork (2015) note that there is a tendency to mistake performance for learning; a teacher may therefore think that learning has been accomplished due to students having successfully completed of a set of tasks in class, failing to take account of later forgetting. Flawed beliefs about learner individual differences are common, too, with over 90% of teachers endorsing the ‘learning styles’ myth in some surveys (Howard-Jones, 2014). And in a study by McCabe (2018), judgement of learning scenarios by academic support centres at universities revealed moderate to low support for effective, evidence-based strategies such as spacing. In lab studies, too, experimental participants tend to misjudge the value of effective learning techniques, even when given a chance to try them out (Yan et al., 2017).

More broadly, people do not seem to have accurate beliefs about how memory works. Simons and Chabris (2011) surveyed the general public, and found that many agreed or strongly agreed with statements such as the following:

- Human memory works like a video camera, accurately recording the events we see and hear so that we can review and inspect them later (63%).

- Once you have experienced an event and formed a memory of it, that memory does not change (48%).

In stark contrast, neither of these statements were endorsed by a panel of cognitive psychologists – zero percent in both cases!

Unfortunately, it appears that we cannot rely on experience and reflection alone to improve this situation; research into various professions has shown that misconceptions about memory can persist even at high-levels of seniority (Melinder & Magnussen, 2015). My own research has found that a large proportion of teachers agree with statements like 'the best way to learn something is to hear or read it repeatedly within the same hour' (which runs contrary to the spacing effect and neglects the impact of forgetting); the level of agreement does not reduce in line with years of classroom experience (Firth, 2018).

Memory for events

The issue of how we conceptualise memory brings me to the research literature on how accurately witnesses remember events. While historically it might have been assumed that a confident eyewitness would provide the best form of evidence, it is now recognised that witnesses often misremember what they have seen, and that their confidence is not at all a good guide to the accuracy of what they say (Wells & Olson, 2003). We can fail to observe apparently obvious changes in our surroundings; for example, Davies and Hine (2007) showed participants a film of a burglary where the actor changed mid-way through, and 61% did not notice the change. Even the duration of witnessed events can be inaccurately recalled – in one study, estimates were five times as long as actual durations (Wright & Loftus, 2008).

Teachers, of course, witness pupil performance and classroom activities, and then later reflect on these. What might impact on the accuracy of this process? Psychological research suggests four key factors:

- As now recognised by law enforcement, memory is easily corrupted by information received after an event, such as leading questions. This information can merge with or replace our memories of the original event. A teaching implication is that lesson observers may unintentionally distort teachers' memories through their questioning.
- Social pressure: memory can be affected when people hear others giving a conflicting account, demonstrating the effect of social pressure on recall. The perceptions of a mentor or manager when discussing a lesson could represent significant social pressure.
- Emotion/anxiety: teachers often feel an increased level of stress when their lessons are observed. The Yerkes-Dodson law suggests that we perform best at a medium level of anxiety, whereas intense stress harms memory; teachers may therefore remember lessons more accurately when not observed.
- Assumptions: there is a tendency, according to schema theory, for people to 'fill in the blanks' of anything they can't entirely remember, using past knowledge and assumptions to do so. Teachers may do this when reflecting on a lesson – and may also be unaware that they are doing so.

In addition, even when we do remember things accurately, we are subject to biases in how we interpret and explain them, in particular the self-serving bias. This is where ambiguous issues of cause and effect tend to be interpreted in ways that are flattering to the observer.

Remedies and conclusion

So far, the evidence I have presented has been rather negative in what it says about the accuracy of teacher reflections. Taken together, these two areas of research suggest that the recollection of a lesson will be subject to both flawed beliefs and inaccurate recall of events. However, there are ways of mitigating the effects described above.

I suggest two main approaches to addressing the issue. Firstly, it's not advisable to ask learners to reflect on their learning without providing them with some kind of tools or a framework for the task (EEF, 2018), and the same logic can be extended to teachers. However teachers are rarely provided with conceptual tools with which to scaffold their reflection (Mockler, 2011), with feedback discussions often drawing on a simple set of prompts or questions, if anything at all. Some specific strategies to support reflection might include:

- Starting a reflective discussion with a short free recall of the lesson by the teacher, followed by prompting by the observer which is kept as neutral as possible ("and did you see anything else…?"). This is the way that investigators conduct 'cognitive interviews' with eyewitnesses.
- Using categories as prompts for reflection on the content of the lesson – a technique which has been found to improve the quality of brainstorming sessions (Deuja et al., 2014).
- A visual proforma, in which the teacher's own professional goals could be set side by side with other information such as the lesson plan, and perhaps (if possible) a

recording or transcript of the lesson.

- Structuring reflection as a series of discrete stages or questions, for example as recommended by Korthegen’s levels of reflection model (e.g. Korthagen & Vasalos, 2005).
- A lesson study approach, where groups of teachers plan and then peer-observe lessons, providing specific and thorough prompting in a relatively non-threatening way.

However, even systematic approaches to reflection such as those suggested above may still be affected by misremembering and biases. Therefore, as a second strand, preparation of new teachers could include an overview of relevant areas of memory and misconceptions, to raise their awareness of this issue. Understandably, teachers often take a ‘common sense’ approach to learning and memory, which – as we have seen – can be flawed, and out of line with scientific understanding. Memory is a fundamentally counterintuitive area of study, and we often simply don’t know what we don’t know about it; as David Dunning (best known for the Dunning-Kruger effect) put it, “people are destined not to know where the solid land of their knowledge ends and the slippery shores of their ignorance begin” (Dunning, 2011, p. 250).

I am not suggesting that reflection is a bad thing or should be reduced – far from it. But as with other areas of teaching, it can benefit from being informed by evidence. Much of our everyday cognition is relatively inaccessible to intuition, and we cannot expect reflection alone to improve the education system without providing suitable guidance about how to navigate its potential pitfalls.

References

Davies G and Hine S (2007) Change blindness and eyewitness testimony. Journal of Psychology 141(4): 423–434.

Deuja A, Kohn NW, Paulus PB and Korde RM (2014) Taking a broad perspective before brainstorming. Group Dynamics: Theory, Research, and Practice 18(3): 222–236.

Donaldson G (2011) Teaching Scotland's Future: Report of a Review of Teacher Education in Scotland. Edinburgh: Scottish Government.

Dunning D (2011) The Dunning–Kruger effect: on being ignorant of one's own ignorance. In: Zanna M and Olson J (eds) Advances in experimental social psychology, Vol 44. New York: Academic Press, pp.247–296.

Education Endowment Foundation (EEF) (2018). Metacognition and self-regulated learning: Guidance report. Available at: https://educationendowmentfoundation.org.uk/public/files/Presentations/Publications/Metacognition/EEF_Metacognition_and_self-regulated_learning.pdf (accessed 7 June 2019).

Firth J (2018) Teachers' beliefs about memory: What are the implications for in-service teacher education? Psychology of Education Review 42(2): 15–22.

Howard-Jones PA (2014) Neuroscience and education: Myths and messages. Nature Reviews Neuroscience 15(12): 817–824.

Kornell N and Bjork RA (2007) The promise and perils of self-regulated study. Psychonomic Bulletin & Review 14(2): 219–224.

Korthagen F and Vasalos A (2005) Levels in reflection: Core reflection as a means to enhance professional growth. Teachers and Teaching 11(1): 47–71.

Melinder A and Magnussen S (2015) Psychologists and psychiatrists serving as expert witnesses in court: What do they know about eyewitness memory?. Psychology, Crime & Law 21(1): 53–61.

Mockler N (2011) Beyond ‘what works’: Understanding teacher identity as a practical and political tool. Teachers and Teaching 17(5): 517–528.

Simons DJ and Chabris CF (2011) What people believe about how memory works: A representative survey of the US population. PloS one 6(8): e22757.

Soderstrom NC and Bjork RA (2015) Learning versus performance: An integrative review. Perspectives on Psychological Science 10(2): 176–199.

Wells GL and Olson EA (2003) Eyewitness testimony. Annual Review of Psychology 54(1): 277–295.

Wright DB and Loftus EF (2008) Eyewitness memory. In: Cohen G and Conway MA (eds) Memory in the Real World. Hove, UK: Psychology Press, pp.91–105.

Yan VX, Soderstrom NC, Seneviratna GS, Bjork EL and Bjork RA (2017) How should exemplars be sequenced in inductive learning? Empirical evidence versus learners’ opinions. Journal of Experimental Psychology, Applied 23(4): 403–416.

Appendix 8: Report on school psychology teaching (Bohan et al., 2017)

This report was prepared by a working group of the British Psychological Society, Scotland in October 2017, of which I am a member. It summarised the current state of pre-tertiary psychology teaching in Scotland at the time, and the priorities for improvements.

Teaching Psychology in Scottish Schools and Colleges: the Future

by J. Bohan, J. Firth, K. Russell and M. Williamson

Introduction

This report describes the ongoing work of the British Psychological Society Scottish Branch's (BPS-S) Pre-Tertiary Working Group1, which was set up in 2016. The remit of the group is to gather information on all aspects of pre-tertiary Psychology education in Scotland and to identify areas of concern and where support is needed. To date this group has completed two major pieces of work during 2017: an audit of pre-tertiary education providers on their perception of current and future Psychology education provision, and a one-day conference of key stakeholders involved in Psychology education in Scotland, held in Edinburgh on 1st June. In this report we summarise the key findings of the audit, the main themes raised by stakeholders at the conference, and proposals for future actions.

Overview of pre-tertiary Psychology teaching in context

In Scotland, Psychology has been taught in further education colleges since the early 1990s and in schools since 1999, and its popularity reflects the pattern of growth internationally over the last few decades; Psychology is popular in schools in at least ten other European countries, and is long-established in US high schools.

Prior to 2016, the only teacher training in Psychology was the in-service TQFE (Teaching Qualification for Further Education) for college lecturers. Many Psychology teachers in schools therefore initially trained in and taught other school subjects before starting to teach Psychology. Some have added Psychology to their teaching qualification via GTCS' Professional Registration process. A new Professional Graduate Diploma in Education (PGDE, secondary) course in Psychology teaching began in the academic year 2016-17 at the University of Strathclyde, allowing newly qualified teachers (NQTs) in Psychology to enter the profession for the first time. The Working Group are unaware of any plans for courses at other institutions, although, anecdotally, demand from Psychology graduates appears to be high.

In Scottish schools, the main courses taught are the National Qualifications (NQs) provided by the Scottish Qualifications Authority (SQA): National 5 Psychology (or 'N5') and Higher Psychology. There is currently no Advanced Higher (AH) or National 4 ('N4'). Both N5 and Higher involve a practical research assignment and an exam, both externally assessed. In 2017, 622 candidates sat the N5, and 3666 sat the Higher qualifications; entry figures have increased steadily since the first exams in 2000. These courses are also taught at most FE colleges, and in addition, many college students study for the Higher National Certificate (HNC) and/or Higher National Diploma (HND) in Social Sciences; these courses are at SCQF levels 7 and 8, and include elements of Psychology and research methods (as well as Sociology, History and Politics). Psychology is cited as being the third most popular subject studied at UK universities (Trapp et al., 2011) and the SQA figures show that, in a Scottish context, this popularity extends to pre-tertiary education, with room for further expansion.

Teachers of Psychology in schools and colleges have formed a strong community, with activities including events, sharing of resources, and peer support. Most teachers engage in either the Association for the Teaching of Psychology Scotland (ATPS, c.60 members), or the PsychEmail network (c.170 participants), or both.

However, whilst studying Psychology in Scotland's schools and colleges is increasingly popular there has been little data collected on educational leaders' views as to the current provision and future growth of Psychology as a subject. As such, in 2016 the BPS-S established a Pre-Tertiary Working Group to address this issue. Their remit was to gather information on all aspects of pre-tertiary education in Scotland and to identify areas of concern and where support is needed.

Initially the Working Group conducted an audit of Scottish education providers and asked a range of questions related to their perception of the subject and its place in the curriculum, as well as how the subject is delivered at their institution. In addition, the Working Group organised an open conference on the state of pre-tertiary Psychology education in Scotland and invited key stakeholders to the event. This provided an opportunity for individuals and organisations across the education sectors to discuss the issues they see as important to the continuing future growth of Psychology in pre-tertiary education. The outcomes of both are summarised in the following sections.

Audit of provision of Pre-tertiary Psychology Education

An online questionnaire was emailed to 429 Scottish school head teachers and 17 FE college principals in March/April 2017. Of these, 36 schools and 6 FE colleges completed the survey.

81% of the school responses revealed that pupils were currently studying Psychology, with 39% responding that Psychology was taught in-school by staff members, and 42% taught via their local FE College. 19% reported that Psychology was not available to pupils. However, these percentages should be treated with caution due to the size of the sample, and the possibility that centres which currently offer Psychology were more likely to respond.

This popularity of the subject was also reflected with 76% heads estimating that there was a high-to-medium demand for N5/Higher Psychology from pupils and parents. As well as its inherent interest value, the popularity of the subject may well relate to the skills it provides; 91% of head teachers viewed it as a useful subject for developing skills for life, work and future study.

The survey also asked questions related to Psychology's place within the wider curriculum. There was general agreement that Psychology was considered to be a science comparable to the natural sciences (68%). But unlike other science subjects there was less agreement on whether it should be a core subject in the curriculum; only 24% agreed / strongly agreed, 35% neither agreed nor disagreed, 41% disagreed / strongly disagreed. So, whilst the subject is popular and considered valuable for assisting personal and academic development, the majority were less sure that it should be a core subject in the curriculum.

Other questions related to who should be teaching Psychology. Head teachers were asked if Psychology NQ courses "could potentially be taught well in my centre by a teacher of another subject, even if they have not studied Psychology at degree level". 63% disagreed / strongly disagreed with this statement, with only 24% agreeing / strongly agreeing. This suggests that most head teachers see the value in employing Psychology teachers who have training and a degree specific to the subject. In open responses, some head teachers pointed out that employing a Psychology teacher was a luxury in the current financial situation, whilst others suggested that Psychology teachers could positively contribute across the broader curriculum.

College responses must be treated with great caution as only six of the 17 FE colleges responded; those who did respond may well be those that already view the subject favourably, therefore these results should not be treated as indicative of the state of Psychology in the FE sector in general.

All college respondents said they provided a range of Psychology courses: all offered Higher Psychology, and all but one offered Psychology National 5. Five of the six delivered Psychology within HNC/HND courses and almost all provided Psychology units as elements of other courses. Modes of provision included daytime classes, evening classes and school-college partnerships.

Similar to the data from schools, the popularity of the subject was again evident in this sector, in that all but one college respondent estimated demand for Psychology to be 'high'. As for perceptions of the value and status of the subject in the curriculum, there was strong agreement that a Higher or N5 Psychology compared well to other subjects in developing useful skills for life, work and future study, and also that Psychology is a research-based, scientific subject. This positive evaluation of the subject was also reflected in the consensus view that, in order to ensure it is well taught, lecturers should have studied Psychology at degree level. In addition, most respondents agreed or strongly agreed that Psychology should be taught in all colleges at multiple levels, suggesting that colleges may see Psychology as a core curriculum element to a greater extent than schools. There was substantial agreement that Psychology lecturers can contribute to the broader college curriculum, such as cross-curricular promotion of wellbeing, interdisciplinary projects, and in-house CPD on teaching and learning, and young people's social-emotional development.

In sum, data from both schools and colleges, though limited, show high demand from students for Psychology, positive perceptions of its value and status as a subject in the overall curriculum, and agreement that it should be taught by those with an appropriate degree-level qualification in the discipline. Based on the responses submitted, one apparent difference between the sectors is that colleges see Psychology as having greater importance in their curriculum than do schools. We might speculate that this may be, at least in part, attributable to the sectors' different histories of delivering Psychology: many colleges have a long-established tradition of Psychology education, having offered 'NC' Units from the early 1990s onwards; they were thus relatively well-prepared for teaching the new NQ courses in 1999, in terms of degree-qualified staff and resources. In addition, many college staff teach Psychology at both HE and FE level, strengthening their ability to provide Psychology education. For schools the process has been more recent and more complex: Psychology was not taught in schools before 1999, and although demand from pupils for the new courses was evident from the start, a general lack of school staff with a Psychology qualification led to a more chequered pattern of provision, including school-college partnerships and delivery by teachers of other subjects, or indeed non-provision; many schools still do not offer Psychology.

Event: Teaching Psychology in Scottish Schools and Colleges - the Future

Teaching Psychology in Scottish Schools and Colleges - The Future was an open event organised by the British Psychological Society Scottish Branch (BPS-S) Pre-Tertiary Working Group which took place on 1st June 2017 in Edinburgh. It brought together organisations and individuals from a range of backgrounds and perspectives – school teachers, college lecturers, university academics, student teachers, representatives from national education agencies - who are all involved in Psychology education in Scotland, to discuss the subject's future at a time of major changes in Scottish education. Through bringing people together in this way, it was intended that a broad network or coalition for pre-tertiary Psychology would be established; this was a longer-term aim of the event.

Dr Scott Hardie, Chair of BPS-S, welcomed delegates and introduced a range of speakers including a keynote address by Joe Walker, Senior Education Officer with a Psychology remit, from Education Scotland. The event included a presentation of the BPS-S survey results presented by Morag Williamson and Kirsten Russell on behalf of the Working Group, and there were also presentations by PGDE Psychology course staff and by Psychology educators from school, college and university sectors. A final plenary discussion session led by Jonathan Firth gave delegates the opportunity to discuss issues relevant to pre-tertiary Psychology education. Below we summarise the key themes that delegates identified.

Key themes from the day

Following a review of the presentations, group discussion, and post-event feedback, the following themes have been identified by the Working Group:

1. Links between schools/colleges and the higher education (HE) sector

In his talk, Dr Jason Bohan (Glasgow University) drew attention to the fact that the number of 1st year Psychology undergraduates who have previously studied the subject in some form has increased dramatically (to around 70% in his own experience); this was potentially problematic in terms of the student's transition to HE, as research suggests that students can become demotivated if their prior learning is not taken into consideration (see Kitching and Hulme, 2013) and this could have the unwanted consequence of students switching honours subjects if they do not feel challenged or engaged in their studies. In discussion, delegates felt that cross-sector links were beneficial in several respects but were generally under-developed and the BPS-S could help to facilitate future discussions. It was also noted that preparation for studying Psychology at university was not the sole aim of pre-tertiary Psychology and that the subject provided broader benefits to students' personal and academic development at school or college.

2. Psychology in the 'Broad General Education' (BGE)

Several speakers raised the issue of Psychology's role in the 'Broad General Education' phase which lasts from pre-school/primary up to approximately S3 (age 13-15) of secondary school. The keynote speaker, Joe Walker, argued for the value of psychological skills to future citizens. Val Martin (University of Strathclyde) described how some PGDE trainees had recently taught aspects of Psychology to younger pupils whilst on placement and had developed materials on topics that appealed to the children, including non-verbal communication. Psychology education projects for primary school children are increasingly being conducted (e.g. Rhodes, 2017). In discussion, delegates were generally in favour of teaching Psychology to younger age groups, but it was recognised that teacher expertise was required, especially in terms of ethics, for example when teaching sensitive topics.

3. Psychology across the curriculum

Both the keynote speaker and several other delegates pointed out that Psychology already exists in various guises in several areas of the school/college curriculum, such as biology, RME/RMPS, PSE, etc. It was felt that such content should be explicitly identified as Psychology. Likewise, cross-curricular projects on aspects of learning and general well-being, for example on motivation, 'mindsets', 'emotional intelligence' etc, were essentially psychological and required appropriate expertise. Psychology teachers would likely be well-placed to be involved in such projects, and/or to evaluate the quality of any external provision.

In addition, these are areas where collaboration with educational psychologists would be helpful, but delegates believed this did not often happen.

4. Teacher training and continuing professional development (CPD)

The event included a description of the new Psychology PGDE course by Val Martin and Norrie McKay (both University of Strathclyde), and several of their students attended the event. Norrie McKay described the difficulties of placing both students and NQTs due to the relatively restricted numbers of schools in the Glasgow area which offer Psychology. Students were placed as far afield as North Berwick and Arran this year. It was acknowledged that this situation may change over time, as the course is accepting a smaller cohort in the coming year, and the number of schools offering Psychology is gradually increasing.

5. Quality and demand of Psychology courses

There was some discussion of whether the content of SQA's suite of courses is appropriate, and whether it prepares learners well for their future work and studies. Some delegates felt that Higher Psychology was too easy, while others felt that it was very demanding, with coursework which is more comparable to that in AH courses in other subjects, and a lower-than-average pass rate due to the stronger emphasis on analysis and evaluation. Generally the content was seen to be relevant and interesting to young learners. The emphasis on research skills including assessed practical research was seen as a strength that should not be watered down by future changes. The SQA is currently reviewing the Higher syllabus, and it was suggested that BPS-S and HE stakeholders could assist in this process. It was also noted that there are other courses available, and that if the current courses do not meet pupils' and centres' requirements there could be a shift in uptake, for example with more online study of introductory-level HE courses by school learners.

6. Role of Psychology education in meeting national policy objectives

The keynote speaker asked delegates to consider the potential of Psychology education in raising attainment and reducing the ‘attainment gap’, which are major current education objectives of the Scottish government. In order to realise such potential, Psychology might be taught (in some appropriate form) to much younger children, i.e. at Primary/pre-school level, and even to parents. Support for parents exists and often focuses on psychological processes such as attachment, mental health and learning, but is generally outwith the education system. The government regards engagement with parents as crucial in raising attainment.

Recommendations for future actions

Unless otherwise stated, these recommendations refer to proposed future actions for BPS-S and/or the Working Group. Some recommendations are dependent on funding being obtained.

1. Links between schools/colleges and the HE sector

It is recommended that BPS-S facilitate communication between HE and the school/FE sector by collating and publicising information on existing links and collaborative projects, providing information/advice to support the development of new links, and holding further events. This could help to improve articulation between undergraduate 1st year content and

Higher/A-level/HN courses. It is also important to investigate ways of informing HE stakeholders about the content and value of pre-tertiary Psychology courses.

2. Psychology in the BGE

It is recommended that BPS-S investigate ways to facilitate support materials being collated or produced for the Primary and S1-3 age groups, and offers professional support via CPD for teachers working with this age group as a collaborative venture, such as a one-off course involving an HE institution, ATPS, and educational psychologist(s). SQA has been asked to introduce N4 Psychology, and BPS-S could write to SQA supporting this request.

3. Psychology across the curriculum

BPS-S could establish links with researchers who study psychological processes relevant to the broader pre-tertiary curriculum, and facilitate a dialogue between this group of researchers, Education Scotland, GTCS and SQA to ensure that relevant curriculum areas such as PSE and Health & Wellbeing are based on psychological evidence and are delivered by knowledgeable professionals.

4. Teacher training and CPD

BPS-S could play a coordinating role in ongoing CPD provision for staff, in partnership with other organisations, to help ensure that a range of affordable CPD options are available throughout the year and for colleagues at different stages of their teaching careers. This could include supporting the current PGDE course which is still in its infancy, as well as investigating the possibility of further courses being launched, especially in areas that are unlikely to benefit from the current Glasgow-based course e.g. the north-east of Scotland.

5. Quality and demand of Psychology courses

It is recommended that BPS-S prepare a briefing document for SQA's revision/review of Higher Psychology in late 2017/early 2018, in order to offer input regarding course content and assessment, including information on BPS pre-tertiary policies, availability of expertise in the discipline of Psychology, and the key findings from this report. See also Conclusions for an update on this point (below). Over the longer term, research could be undertaken to look at the content and skills included in Psychology courses and how they compare in terms of demand to other pre-tertiary courses.

6. Role of Psychology education in meeting national policy objectives

The psychological basis of Scottish Government policy objectives such as the Pupil Equity Fund - and consequent decision-making within institutions - could be evaluated, and findings circulated. BPS-S could contact STEMEC, the Scottish Government's advisory group on STEM education (see http://bit.ly/2hSV0PY) to promote Psychology's role in future developments around STEM education. One or more BPS-S or Working Group members could attend/contribute to the Holyrood conference on STEM education in March 2018 (http://stem.holyrood.com).

Conclusions

Psychology education in Scotland's schools and colleges is healthy. There is strong demand from students and parents and education providers recognise the value of the subject for the personal and academic development of learners. There is also acknowledgement that Psychology should be taught by subject specialists, and that Psychology teachers can positively contribute across the curriculum. Overall, it appears that Psychology is perceived to have a positive and growing place within the school and college curriculum, however, certain issues are recognised which need to be addressed. At the same time, there are promising areas of untapped potential.

The Working Group has therefore set out the above recommendations in order to pursue the aim of further developing and supporting pre-tertiary Psychology in Scotland. An important focus is on facilitating links and collaborative action amongst key stakeholders; indeed the June 2017 event in Edinburgh, and the subsequent dissemination of this report, constitute first steps in establishing a supportive, collaborative network. The BPS-S is committed to supporting such activity and will organise future forums to that end.

One specific action already initiated is BPS-S representation at SQA N5/Higher review meetings (see recommendation 5a above). Pre-tertiary Working Group member Dr Jason Bohan has agreed to act as this representative and will report on progress in appropriate forums, such as the annual ATPS CPD event (autumn 2017) and future BPS-S education events.

The recommended actions above are not definitive; the Working Group will continue to explore other avenues in support of pre-tertiary Psychology in Scotland, especially collaborative activities, and we encourage all organisations and individuals who share our objectives to contact us with comments and suggestions for future actions (contact details are below).

References

Hulme, J. A., & Kitching, H. J. (2013). Bridging the gap: facilitating students' transition from pre-tertiary to university psychology education. *Psychology Teaching Review, 19*(2), 15–30.

Rhodes, E. (2017) After-school psychology club. *The Psychologist, 30*(8), 16–17. Retrieved from https://thepsychologist.bps.org.uk/volume-30/august-2017/after-school-psychology-club

Trapp, A., Banister, P., Ellis, J., Latto, R., Miell, D., & Upton, D. (2011). The future of undergraduate psychology in the United Kingdom. *Higher Education Academy*. Retrieved from www.bps.org.uk/sites/default/files/documents/the_future_of_undergraduate_psychology_in_the_uk.pdf.

Appendix 9: Examples of the four skills used for Study 2.

Skill is stated in brackets after each example. Examples derive from the following research studies, all of which are included in the mandatory content for Higher Psychology: Milgram (1963); Mori and Arai (2010); Dement and Kleitman (1959); Czeisler et al. (1990), as well as the following studies which are very widely taught: Asch (1951); Bickman 1974; Hofling et al. (1966); Tajfel (1970). In the latter three examples the studies featured only in the test phase, and unfamiliarity with the studies should not in principle have been a barrier to identifying the correct skill.

Pre-test phase

Asch's study was a lab experiment, and the task did not resemble real life situation.

Asch's study used groups of actors. This allowed Asch to decide in advance what the majority group was going to say, and to put pressure on the real participant.

Asch's study was different from previous research studies into conformity such as the work of Jenness, as it used a task where the answer was clear and unambiguous. This allowed Asch to demonstrate normative social influence.

Main phase

This 1963 study was a lab experiment. It used 40 participants, all of whom were male (description)

The study involved an overnight stay in a sleep laboratory, which may have caused some participants to sleep less well than normal, or may have affected the content of their dreams (description)

In this study the participants went home to sleep, and it wasn't clear that they were getting a good night sleep at home, which may have caused them to be more tired (description)

A follow up to the study used a condition where the experimenter left the room, and there was a lower level of obedience when the authority figure was not physically present (description)

The study used groups of four, and all of the participants were fellow students, meaning that it was a reasonably realistic setting (description)

As all of the participants in this study were male, it is impossible to generalise the results of the study to women, limiting the value of the original findings (evaluation)

By forcing participants to sleep in a sleep laboratory, the study may have influenced their sleep quality or affected the content of their dreams, making the results less valid – sleep may be different in everyday life (evaluation)

Although this sleep study monitored participants overnight, they slept at home, and variations in their sleeping conditions such as noise could lead to error in the results, making the findings less reliable (evaluation)

By testing the effect of the authority figure leaving the room, the study made an accurate assessment of how important it is for an authority figure to be physical present, strengthening earlier findings (evaluation)

As the study used groups of fellow students, the experiment was realistic setting, meaning that its findings can be generalised to everyday situations where conformity occurs (evaluation)

The researchers used a set of participants who were all told to give the wrong answer on certain occasions. This meant that the true participant sometimes heard the majority giving a wrong answer (explain)

The study measured participants hormone levels. This was done so that they could find out whether the participants' body clocks were gradually changing across the six 'nightshifts' (explain)

The study used polysomnography, which includes an EEG measure of a person's brain activity. This allowed the researchers to see which stage of sleep a participant was in, and to wake them up at the right points (explain)

In order to make the deception convincing, the researcher had recorded a person crying out in pain, and this was played after the 'shock' switches were pressed. This allowed the researcher to ensure that every participant heard the same thing (explain)

The study used filter glasses because the researchers needed participants to see different line lengths (explain)

In this study, participants heard other people giving a clearly wrong answer. The study was designed this way because the researcher wanted to test normative influence, and it had to be clear that people were conforming in order to be liked, and not due to uncertainty (analysis)

The study measured participants' attention by giving them frequent short tests during the night. This is because cognitive psychologists understand that attention levels reduce when we get tired (analysis)

The use of polysomnography, which includes an EEG measure of a person's brain activity, is typical of research in the biological approach to psychology, in which physical process of the brain are assumed to link to thoughts and behaviour (analysis)

In the study, the researcher frequently said that he took responsibility for any harm that came to the victim. This aspect of the design links to agency theory, which says that obedience is more likely if an authority figure takes responsibility (analysis)

Previous studies had used actors to provoke conformity, but this used sets of filter glasses to make participants to see different line lengths, and as a result, it is likely that people's answers sounded more natural than was the case in older research (analysis)

Test phase, set 1 (items 1–12)

Milgram's (1963) study was an artificial lab experiment. This limits the validity of the results, and makes it hard to be sure whether similar obedience levels would be found in real life (evaluation)

Asch's study was different from previous research studies into conformity such as the work of Jenness, as it used a task where the answer was clear and unambiguous. This allowed Asch to demonstrate normative social influence (analysis)

One strength of the Mori & Arai (2010) study was that it had a large sample of over 100 participants (description)

The Czeisler et al. (1990) study was done on volunteers, and the findings can't be generalised to actual shift workers (evaluation)

The experimenter in the room frequently said that he took responsibility for any harm that came to the victim. This links to agency theory, which says that obedience is more likely if an authority figure takes responsibility (analysis)

The glasses used in the Mori and Arai (2010) study allowed the researchers to control who saw what. This meant that they could ensure that lines looked longer to one participant out of four (explanation)

A field experiment by Hofling et al. (1966) tested whether nurses would follow an order to give a drug overdose. This study found an even higher level of obedience than Milgram had, suggesting that Milgram's results were not due to the artificiality of his setting (analysis)

Asch's study was a lab experiment, and the task did not resemble real life situation (description)

One problem with the Tajfel (1970) study was that the task was highly artificial and short-term, therefore making it hard to generalise the findings to real cases of prejudice (evaluation)

Bickman's (1974) study of obedience used actors in different types of clothing to give instructions, and this allowed the researcher to study the effect of uniform on obedience level (explanation)

A flaw in Dement and Kleitman's (1957) study of sleep was that it used a very small sample - just 9 people (description)

A study of circadian rhythms investigated the sleep-wake cycle of an individual who remained in near-total darkness for several weeks. This allowed the researchers to find out what happens when we have no external zeitgebers to affect our body clock (explanation)

Test phase, set 2 (items 13–24)

Asch's study used groups of actors. This allowed Asch to decide in advance what the majority group was going to say, and to put pressure on the real participant (explanation)

The ideas from research by Milgram and Hofling et al. can be applied in the real world, such as by increasing the level of obedience to authority in emergency situations by ensuring that all emergency workers are in uniform (analysis)

By having participants sleep in a laboratory, the Dement & Kleitman study may have influenced their sleep quality or affected the content of their dreams, making the results less valid (evaluation)

An issue with the sample in Czeisler et al's (1990) study of sleep was that the participants were not shift workers, they were ordinary members of the public (description)

Mori and Arai (2010) used participants who knew each other - students from the same college. This contrasts with some earlier studies such as the work of Asch, where conformity was tested in groups of strangers (analysis)

Sherif et al's 'Robber's cave' study investigated prejudice in a very limited group - their participants were all white Protestant schoolboys (description)

An important aspect of the setup of the Milgram (1963) study was that the true participant and Mr Wallace drew lots for the positions of teacher and learner. This was faked, but it made it look to the participant as if they could have been in the 'learner' role (explanation)

Asch's (1951) study was a lab experiment with other variables kept constant, allowing a clear conclusion cause-and-effect conclusion to be drawn and ensuring that the results were valid (evaluation)

Czeisler used a small sample of young male participants, meaning that the study was able to gain rich detail and control, but making it harder to generalise the results to women or to male workers from different age groups (evaluation)

The use of an EEG by Dement and Kleitman (1957) is typical of research in the biological approach to psychology, in which physical process of the brain are assumed to link to thoughts and behaviour (analysis)

To ensure that the situation seemed real, participants heard a recording of a person crying out in pain after each 'shock' switch was pressed. This also allowed the Milgram to ensure that every participant heard the same thing (explanation)

Czeisler’s study used several measures of circadian changes, relying not just on attention scores but also measuring blood hormone levels and body temperature (description)

Appendix 10: Experts in learning (Chapter in an edited book) (Firth, 2017)

This piece was accepted for a book about professionalism and professional status in teaching. Reference: Firth, J. (2017). Experts in learning. In L. Rycroft-Smith & J. L. Dutaut (Eds.) Flip the system UK: A teachers' manifesto (pp. 20–28). Routledge.

Experts in Learning

by J. Firth

Flipping the system means putting teachers back in control of educational practice, and this book is therefore acutely concerned with the locus of judgements made within teaching. The gradual centralisation of education since the late 1980s, with decisions over content and pedagogy increasingly being taken by government departments, can be seen as a direct challenge to teacher professionalism. This question seems more urgent than ever: do we want our teachers to be homogeneous 'delivery agents', or to be autonomous and diverse professionals who are experts in learning?

Within this debate, a 'what works' approach to teaching practice - whereby research evidence is used to guide teaching practice – is sometimes seen as a threat to teachers' professional judgement on the basis that externally set standards (often based on out-of-date or cherry-picked research results) are imposed on teachers from the top down. In this chapter I argue that research-based knowledge about learning, in particular the intricate and often counterintuitive functioning of human long-term memory, is actually a key tool for the emancipation of teachers and for flipping the system towards greater teacher agency. Armed with this knowledge and an understanding of their own learning context, teachers will always be in a better position to say what works for their learners than any external authority.

Criticisms of evidence-based practice

It is important first of all to set out why evidence-based practice has been seen as a threat. One factor is that it implies an imposition of teaching practices with little or no regard to whether these practices are appropriate to the setting - a criticism of centralisation rather than of the use of evidence per se. However, there is a further concern about the nature of the desired improvements, well expressed here by Biesta (2007, p.5):

"*…evidence-based education seems to favor a technocratic model in which it is assumed that the only relevant research questions are questions about the effectiveness of educational means and techniques, forgetting, among other things, that what counts as "effective" crucially depends on judgments about what is educationally desirable.*"

It is certainly true that efficacy in learning and memory is an empirical matter that does not address moral questions of purpose. However, that the idea that studying 'what works' is insufficient does not make it (in my view) any less worthwhile, nor does it impede discussion of values and purposes. In practice, there is already a great deal of discussion about what young people might need to learn for the workplace of the future or to enrich their lives, yet the technical question of how best to impart knowledge and skills tends to be neglected. It is entirely possible to consider how to apply learning science effectively in concert with the already rich and emotive debate over what should be taught.

Research and teacher agency

Most parts of the UK have witnessed an increasing level of interference in teaching practices as well as curriculum content under the general label of raising standards (Alexander, 2014). Such moves tend to be accompanied by heightened accountability measures, reducing teacher agency and prompting teachers to become compliant and unadventurous in their professional practice (Sachs, 2016).

In opposition to this, some voices have called for increased levels of professionalism and autonomy among the teaching profession. A positive example in recent years has been the promotion of teachers engaging with and carrying out their own research; the British Educational Research Association (BERA) stated that "teachers and teacher educators can be equipped to engage with and be discerning consumers of research…[and] may be equipped to conduct their own research, individually and collectively" (2014, p. 5). In his report on 'Teaching Scotland's Future', Donaldson (2011) recommends that narrow interpretations of the teacher's role must be challenged in order to facilitate engagement with research.

This perspective on professional practice has much to recommend it. Professionals who have more control are less stressed (Marmot *et al.*, 1991), while the ability to make meaningful changes is highly motivating and prompts creativity in the workplace (Amabile & Kramer, 2011). Teaching, from this point of view, is better viewed as a disciplined improvisation rather than as the fulfilment of a precisely programmed series of actions, a perspective that is supported by the finding that the level of detail in teachers' lesson planning reduces in line with their years of classroom experience (Sawyer, 2004).

Greater research engagement has the potential to lead to a more responsive form of accountability (Halstead, 1994) whereby practitioners continually analyse and modify their own professional processes. Doing so requires teachers to have control over what they do in the classroom, and so the movement towards teacher research engagement has an intrinsic link to teacher agency.

Knowledge and intuition

Debate about teacher agency is part of a broader discourse around professionalism. Teachers, of course, feel that they should be viewed, treated and remunerated as professionals. However, the quality of the decisions that any professional makes depends on their knowledge and skills.

The role of research evidence in professional practice is to suggest improvements and to modify harmful or ineffective practices – just as if our learners were using bad study strategies, we wouldn't hesitate to correct them (and they probably are - see Hartwig & Dunlosky, 2012). Initial teacher training can be conceptualized as a process where inexperienced teachers are guided in the development of these professional skills, while CPD activities aim to foster and extend them throughout the teaching career. The term 'skill' should be used with some caution, though. In an era where the politics of pedagogy has come to focus on technical skills rather than theoretical and pedagogical understanding, the importance of theory and values-based teacher education needs defending (Donaldson, 2011; Brown *et al.*, 2016).

The focus here is on professional practice that is grounded in an understanding of theory; one of the main contentions of this chapter is that it is important to teacher professionalism that we engage with research into human long-term memory. Our research knowledge must be better than the 'common sense' understanding of an intelligent non-teacher if it is to be considered professional knowledge at all. This is especially important given that assumptions about memory tend to encompass a range of inaccurate views, such as the widespread concept of 'permanent memory' – the idea that memory for a single experience, once well learned, does not change (Simons & Chabris, 2011).

Indeed, memory as a field of study has been described as inherently unintuitive (Bjork, 2011). There is good reason therefore to think that without adequate training, a teacher would be working with highly inaccurate assumptions about how information is used, taken in, consolidated and later retrieved. Memory is not the only research area that can affect teaching practice but it is a particularly fundamental one, given the pedagogical importance of learning and retention. Inaccurate assumptions could impact on a great many teacher decisions, from designing materials to deciding when to finish working on a task.

What about the role of experience? As governments increasingly push for schemes that place untrained graduates directly into teaching roles, it would be useful to know whether a more accurate understanding of memory can develop over time. Here, research into other professions such as lawyers and judges suggests that even years of experience cannot overcome memory biases and misconceptions (Magnussen *et al*., 2010). Psychological research into cognition and prejudice also suggests that experience alone cannot be relied on to overcome biases in thinking (pupils' flawed revision strategies may be further evidence of this!).

Our intuitive thought processes are rapid, and largely automatic. One example (Kahneman, 2002, p. 451) involves the application of basic arithmetic to a puzzle:

A bat and a ball cost £1.10 in total. The bat costs £1 more than the ball. How much does the ball cost?

This type of thinking may well give us a quick but inaccurate response - 10p. However, we also have a slower and more effortful mode of thinking, which is more sensitive to training (Kahneman, *op cit*.) When we scrutinise our own answer using this more deliberate system, we realise that the obvious assumption is incorrect - the ball must cost 5p, and the bat £1.05.

As teachers, a key metacognitive function of this more effortful analytical thinking is the ability to check our own intuitive reasoning, modifying judgements on the basis of our theoretical understanding. It’s not the case that intuition is always (or even often) inaccurate; these abilities have evolved through human pre-history for their survival value, and in some areas of interaction intuition serves us well - for example, snap judgements of personality and mood appear to be largely accurate (Ambady & Rosenthal, 1992). However other domains such as statistics seem to inherently run counter to human intuition, with assumptions leading to systematic errors (Tversky & Kahneman, 1974).

For reasons discussed above, there are good reasons to think that the domain of learning and memory is a highly unintuitive area where theoretical knowledge has an important role to play as part of pedagogical expertise. Yan, Bjork and Bjork (2016) have shown that even when effective learning strategies are tried out or their logical benefits explained (in a way analogous to teacher CPD), research participants tend to still stick to their inaccurate hunches. However, if people had both explicit training *and* practical experience of better strategies, they did adjust their behaviour and thinking. This research finding concords with a view of professionalism where neither experience not factual know-how alone can fully equip a teacher for their role, but a synthesis of theoretical understanding and practical experience can do so.

Key evidence - learning and memory

What, then, are the main evidence-based principles of memory that can inform our professional practice? Let us consider some relevant facts about human memory.

Memory is easily distorted

Memories of events are not discrete entities. Instead, they are interlinked in networks which psychologists and linguists refer to as schemas (or schemata). Taking in new information is not a neutral process, but is influenced by the state of prior knowledge and beliefs - new information can influence these schemas, but likewise old memories can distort new ones. Memories can also be distorted by later information and questioning, with people often finding it hard to retain with certainty the source of their memories (Schacter, 2001), and even just the process of recalling a memory changes it (Bjork, 2011). As has been recognised for many decades in psychology, taking in new information is a dynamic process of interpretation.

Learning should be distributed over time

Learners often focus on working on a set of tasks until they have ‘got it’, but immediate performance is an unreliable guide to long-term learning (Soderstrom & Bjork, 2015); the fact that a learner can do something today, tomorrow or even next week does not mean that they will be able to do it in six months’ time. Conversely, failure in the short term does not mean that they have not learned.

This dissociation between current performance and long-term learning links to a powerful strategy for learning and review - learning should be distributed over time, with larger rather than smaller gaps between initial learning and review tasks (Rohrer, 2015). More spaced-out learning and practice reduces forgetting and may help learners to mentally connect new information to a broader set of experiences.

Immediate feedback

It might be assumed that immediate corrective feedback is highly valuable to learning. However, a body of research has shown that delaying and minimising feedback can result in better long-term retention (Soderstrom & Bjork, 2013). This may link to the benefits of learners reflecting on problems themselves and overcoming difficulties independently. In addition, Hattie and Timperley (2007) note that corrective feedback is among the least effective options, with feedback based on highlighting the best elements of student work (but not generic praise) having a greater effect.

As with distributed learning, this area of research presents the counterintuitive conclusion that delaying and reducing teacher input can help learning over the long term.

Generation and retrieval

It is increasingly being recognized that repetition and re-reading are ineffective learning strategies (e.g. Hartwig & Dunlosky, 2012). So what is? The ‘generation effect’ suggests that words produced by a learner, for example when completing a gap fill, will be better recalled that those which are read - a finding that backs up certain forms of active learning.

A similar finding is the ‘testing effect’ (also called ‘retrieval practice’) - memory is improved when learners have to recall information, relative to repeated re-reading, even in the absence of any feedback (Roediger & Karpicke, 2006). This suggests that educators would benefit from looking at testing in a fundamentally different way - not as assessment, but as a technique for building new memories (Karpicke, 2016). It is important to clarify that retrieval from memory doesn’t need to involve a test - it could include writing or discussion, for example.

These areas represent matters of near-universal agreement in psychology, and although the best way of applying such principles to education is still a matter of ongoing research, they offer considerable potential benefits to the teacher. They are easily applied - techniques such as spacing out topics or delaying feedback needn't involve changes to materials or tasks. What's more, the resulting efficiency gains from these factors represent potential time savings which could ease some of the pressures of teacher workload.

It is worth adding that the discussion of the role of memory in education thus far should not be interpreted in the narrow sense of advocating simple memorisation. A pupil's memory plays a role in every aspect of learning and is used every time they have a conversation or read a text, and it underlies the development of well-integrated conceptual understandings.

Sources of research evidence

Finally, it is important to consider the sources from which teachers get research information. This, too, can represent a form of centralised control, or a means of empowerment of the practitioner. There is a developing grassroots movement among teachers seeking to engage with research-based practice; the teacher-led researchED conferences as well as Pedagoo local meet-ups are important means of sharing knowledge, as are social media and practitioner blogs. Publications such as TES are increasingly featuring research evidence based around memory, too.

The role of governments is more problematic. They could in principle give good-quality generic guidance but recent experience suggests that they often don't; instead, their use of evidence displays considerable distortions based on their ideology and agenda (Alexander, 2014). In a model which features increased teacher agency, the role of governments and other superordinate authorities would see the informed practitioner as the locus of decision making. Their role would be to support rather than dictate to the practitioner, for example by:

- funding teacher access to research journals, as happens for NHS doctors;
- maintaining high standards of initial teacher education;
- supporting teacher research projects with time and funding;
- facilitating and supporting equal-status cooperation between the teaching fraternity and other sectors of academia.

Of course, professional development does not happen in a vacuum but within learning communities, with teachers influencing and supporting one another. Some schools have introduced a 'research lead' position, facilitating dissemination of new educational research evidence among staff. Local school and/or college clusters may be helpful, drawing on a broader pool of knowledge. Pupils and parents, too, can be engaged in the discussion of effective learning, helping to democratise the educational process.

Conclusion

Engaging with research can empower the professional rather than being used to challenge or replace professional judgement. In effect we must flip the evidence-based learning discourse, allowing the research to take its rightful place as an integral component of our professional expertise. Failure to do so will weaken our professional standing at a time when it is already under threat.

Research evidence can benefit a range of domains, but human memory is one that is especially relevant to education. It is also one in which misconceptions are widespread, and intuition – or even years of classroom experience - cannot substitute for an evidence-based theoretical understanding. This chapter has outlined several areas of research from cognitive psychology that are directly applicable to teaching.

Evidence-based professionalism is more likely to emerge under the ideal conditions of good institutional support, teacher autonomy and access to appropriate resources and role models (Drew et al., 2015). But in the absence of such conditions, individual teachers can boost their expertise in learning as part of a grassroots movement, helping our profession as a whole to say to governments and other authorities, 'we understand research in learning and our work is evidence-based… is yours?'

References

Alexander, R. (2014). Evidence, policy and the reform of primary education: A cautionary tale. *Forum, 56*(3), 349–375.

Amabile, T. & Kramer, S. (2011). *The Progress Principle: Using Small Wins to Ignite Joy, Engagement, and Creativity at Work.* Cambridge, MA: Harvard Business Press.

Ambady, N. & Rosenthal, R. (1992). Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274. doi: 10.1037/0033-2909.111.2.256

Biesta, G. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory, 57*(1), 1–22.

BERA-RSA (2014). Research and the teaching profession: Building the capacity for a self-improving education system. Final report of the BERA-RSA inquiry into the role of research in teacher education. London: Author.

Bjork, R. A. (2011). On the symbiosis of remembering, forgetting, and learning. In A.S. Benjamin (Ed.) *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork* (pp. 1–22). New York: Psychology Press.

Brown, T., Rowley, H., and Smith, K. (2016). The beginnings of school led teacher training: New challenges for university teacher education. *School Direct Research Project Final Report*. Accessed 14/01/2017 at http://www.esri.mmu.ac.uk/resgroups/schooldirect.pdf

Donaldson, G. (2011). *Teaching Scotland's Future: Report of a Review of Teacher Education in Scotland*. Edinburgh: Scottish Government.

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology, 1*(1), 72–78.

Drew, V., Priestley, M., & Michael, M. K. (2016). Curriculum development through critical collaborative professional enquiry. *Journal of Professional Capital and Community, 1*(1), 92–106.

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.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel Prize Lecture, 8*, 351–401.

Karpicke, J. D. (2016). A powerful way to improve learning and memory: Practicing retrieval enhances long-term, meaningful learning. *Psychological Science Agenda*. Accessed 24 June 2016 at http://www.apa.org/science/about/psa/2016/06/learning-memory.aspx

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. Ross (Ed.) *Psychology of Learning and Motivation*, Vol 61 (pp.237–284). Waltham, MA: Academic Press.

Magnussen, S., Melinder, A., Stridbeck, U., & Raja, A. Q. (2010). Beliefs about factors affecting the reliability of eyewitness testimony: A comparison of judges, jurors and the general public. *Applied Cognitive Psychology, 24*(1), 122–133.

Marmot, M. G., Stansfeld, S., Patel, C., North, F., Head, J., White, I., ... & Smith, G. D. (1991). Health inequalities among British civil servants: the Whitehall II study. *The Lancet, 337*(8754), 1387–1393.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review, 27*, 635–643.

Sachs, J. (2016). Teacher professionalism: why are we still talking about it?. *Teachers and Teaching, 22*(4), pages 413–425.

Sawyer, R.K. (2004). Creative teaching: Collaborative discussion as disciplined improvisation. *Educational Researcher, 33*(2), 12–20.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. *PloS one, 6*(8), e22757.

Soderstrom, N.C. and Bjork, R.A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124

Webb, P. T. (2005). The anatomy of accountability. *Journal of Education Policy, 20*, 189–208.

Yan, V.X., Bjork, E.L. and Bjork, R.A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General, 145*(7), 918–933. doi: 10.1037/xge0000177

Appendix 11: Article on Teachers' beliefs about memory.

This article was accepted for a BPS journal, and the findings also form the basis of Study 1 in this thesis. Reference; Firth, J. (2018). Teachers' beliefs about memory: What are the implications for in-service teacher education? Psychology of Education Review, 42(2), 15-22.

Teachers' beliefs about memory:
what are the implications for in-service teacher education?

by J. Firth

Abstract

Memory plays a key role in learning, and it is therefore important that teachers understand its workings in order to make decisions that benefit learning. However, previous research has shown that memory is an area which is subject to misconceptions. This study used an online survey with a 5-item Likert scale to determine teachers' responses to statements about memory and forgetting, including counterintuitive phenomena such as the spacing effect. It was found that participants scored better on the items compared to studies of the general public, but there were notable misconceptions. Accuracy of memory beliefs didn't increase in line with self-reported number of years of experience. Teachers of psychology scored higher, suggesting that an understanding of cognition can reduce misconceptions. Although small scale, this survey addressed an under-researched area, and future directions for research and implications for teacher CPD are suggested.

Keywords: professional knowledge, metacognition, spacing effect, retrieval practice

Introduction

Teachers need an accurate understanding of human memory processes in order to present material effectively in the classroom, as they are responsible for guiding student learning. Several well-established psychological principles can inform this work. For example, incorporating a delay before restudying leads to more durable learning compared to more immediate restudy - the spacing effect (Cepeda et al. 2008). Retrieval practice means the active retrieval of information from memory, and is considered a more effective learning and revision strategy than popular choices such as re-reading (Karpicke et al., 2014; Dunlosky et al., 2013). Interleaving - mixing up different types of problems or examples - is another promising strategy.

However, memory is complex and its workings are not intuitively obvious (Bjork, 2011). Misconceptions about memory are rife; a survey by Simons and Chabris (2011, 2012) found the proposition that memory "works like a video camera, accurately recording the events we see and hear so that we can review and inspect them later" was endorsed by 63% of their sample of members of the public, but by 0% of memory researchers. Professionals serving as expert witnesses often endorse similar myths (Melinder & Magnussen, 2015). However, Ost et al. (2017) found that psychology PhD students did somewhat better on similar questions, suggesting that a theoretical understanding of human cognition can lead to more accurate memory beliefs; graduates scored better on the Simons and Chabris survey, too, suggesting that education more broadly could play a role.

Uncertainty about how memory works could lead to learners using flawed study strategies. In a survey of over 300 first year university students, Hartwig and Dunlosky (2012) found a preference for ineffective strategies such as re-reading, which correlate with lower grade point averages. Education has a relatively slow feedback loop, and learners tend to confuse performance in the here-and-now with permanent learning. This means that experience of using one's memory is often not sufficient to develop an accurate understanding of its functions.

If learners are unable to fathom memory for themselves, it would appear to fall to teachers to guide the learning process. However, teachers too appear to be prone to mistaking immediate performance for long-term learning (Soderstrom & Bjork, 2015), suggesting that judgements in educators cannot be based solely on experience or reflection. Learning advice given to students often is characterised by misconceptions; Morehead et al. (2016) found that 91% of teachers endorsed the 'learning styles' myth (the idea that learners can be categorised by their preferred sensory modality e.g. auditory or visual and should be taught mainly via this modality; see e.g. Kirschner, 2017). In the same study, teachers did endorse some more effective strategies such as self-testing – which affords an opportunity for retrieval practice - but their reasoning did not indicate an understanding of how memories are formed.

The present survey aimed to establish whether teachers' beliefs about human memory are subject to similar misconceptions as have been found among other populations. On the basis of memory being counterintuitive, it was predicted that years of experience would have little impact on accuracy of response. The study also looked at participants teaching subjects, to investigate whether background psychological knowledge would affect accuracy of responses.

Method

Design and Materials

A survey was undertaken using the website PsyToolkit (www.psytoolkit.org). It featured 24 statements relating to memory, each with a 5-point Likert scale labelled "strongly agree" to "strongly disagree". The first set of questions related to memory in general, with two drawn directly from Simons and Chabris (2011). The remaining questions focused on aspects of memory relevant to teaching and on which there is broad scientific agreement, for example the spacing effect, retrieval practice, and interleaving (see Table 1 for a full list of questions and summarised responses).

Sample and Procedure

Participants were recruited from a Scottish secondary school (n = 45) or via weblink shared on Twitter (n=34) using the author's account, alongside a tweet inviting teachers to take part. 58 participants were from the secondary sector, 8 from primary, and 14 self-reported as "other including FE and HE". The online survey featured an ethics statement and consent form.

Results

(this section featured a shorter version of the results shown in Section 5.2.3 of the thesis)

Discussion

An accurate understanding of human memory is likely to affect successful planning and teaching. The findings presented here suggest that teachers hold a better understanding of such issues than other professionals or the general public, perhaps because initial teacher education features input on learning theories. However, significant inaccuracies were also present, which showed in particular a lack of understanding of how the spacing effect and retrieval practice might be applied to teaching practice.

The finding that years of experience had little impact on the accuracy of teachers' beliefs supported the study's first hypothesis. The exact reasons why these misconceptions do not diminish with experience remain to be firmly established, but fits with the points made earlier about the counterintuitive nature of memory and the slow feedback loop when learning. Indeed, feedback may promote some teacher misconceptions - flawed strategies are sometimes better in the short-term, e.g. when cramming for an impending test (Kornell, 2009), and, as noted earlier, teachers may mistake current performance for learning.

It is possible that the findings are an artefact of other differences, such as generational differences between groups. Future research could include a longitudinal investigation via surveys with trainee/early career teachers at various stages.

Flaws in teachers' knowledge have implications for their professional learning. The current findings suggest that the CPD undertaken by teachers does not improve their understanding of memory, perhaps because this topic is absent from such training or because an in-depth understanding is not fostered. Teachers may benefit from more thorough training in memory and cognition, a conjecture that is supported by the finding that those teachers whose remit included psychology scored higher overall.

Higher scores among the sample compared to the general public may be attributable in part to their higher education level as qualified teachers, but there may also have been an effect of sampling bias. Twitter users are likely to engage more with the science of learning, e.g. by reading educational blogs which are easily shared through social media. Although a school sample was also used, this institution might not have a typical staff profile, and its location (Scotland) could affect the generalisability of the findings. Future work should feature a more representative sample.

The current study also did not provide baseline responses to the questions; this could be done by surveying beginner student teachers. Further research should also determine the practical effects (if any) of misconceptions about memory, perhaps via classroom observations. This would clarify whether flawed teacher beliefs have a detrimental impact in line with previous findings relating to learner beliefs. It would also be helpful to develop a taxonomy of memory-relevant educational tasks to inform the selection of future survey questions, and to survey memory experts (as in Simons & Chabris, 2011) to better establish the scientific consensus on the issues covered.

Conclusion

Although small scale, the present survey addressed an under-researched area, and can be used as the foundation for further work. It found evidence that teachers' misconceptions about memory are lower than those of the general population, do not change in line with experience, and are higher among those without a psychology teaching remit. To tackle erroneous beliefs, psychology-based theories about learning and memory could form a greater part of teacher education including CPD.

References

Bjork, R. A. (2011). On the symbiosis of remembering, forgetting, and learning. In A.S. Benjamin (Ed.) Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork (pp. 1–22). New York: Psychology Press.

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.

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, 14(1), 4–58.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. Harvard Business Review, 85(7/8), 114–21.

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.

Horvath, J. C. & Lodge, J. M. (2016). A framework for organising and translating science of learning research. In Horvath, J. C., Lodge, J. M., & Hattie, J. (Eds.), From the Laboratory to the Classroom: Translating Science of Learning for Teachers (pp. 7–20). Abingdon, Oxon: Routledge.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. Ross (Ed.), Psychology of Learning and Motivation, Vol. 61 (pp. 237–284). Waltham, MA: Academic Press.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. Computers & Education, 106, 166–171.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. Applied Cognitive Psychology, 23, 1297–1317.

Melinder, A., & Magnussen, S. (2015). Psychologists and psychiatrists serving as expert witnesses in court: what do they know about eyewitness memory?. Psychology, Crime & Law, 21(1), 53–61.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. Memory, 24(2), 257–271.

Ost, J., Easton, S., Hope, L., French, C. C., & Wright, D. B. (2017). Latent variables underlying the memory beliefs of chartered clinical psychologists, hypnotherapists and undergraduate students. Memory, 25(1), 57–68.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. PloS one, 6(8), e22757.

Simons, D. J., & Chabris, C. F. (2012). Common (mis)beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. PloS one, 7(12), e51876.

Soderstrom, N.C. and Bjork, R.A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, 10(2), 176–199.

Appendix 12: Materials used in Study 3

Questions, notional correct answers, and academic rationale.

Item No.	Item wording	Correct response	Rationale
1	People suffering from amnesia typically cannot recall their own name or identity.	Disagree	Media myth about amnesia.
2	Sometimes people who have committed murder have no memory for the crime because they have repressed the memory.	Disagree	Generally discredited theory.
3	Very high stress during an event can harm a person’s ability to remember the event accurately at a later date.	Agree	Eyewitness memory literature.
4	When small children describe events they have experienced, their accounts are usually more accurate than those of adults.	Disagree	Eyewitness memory literature.
5	A person’s perception and memory for an event may be affected by his or her attitudes and expectations.	Agree	Social perception and eyewitness memory literature.

6	Once you have experienced an event and formed a memory of it, that memory does not change.	Disagree	Eyewitness memory literature; memory retrieval literature.
7	Improvements in learning always require spending more time studying.	Disagree	Memory techniques can improve performance with study time held constant.
8	Most learners have a good idea of how practice/study will impact on their memory.	Disagree	Metacognitive literature on study habits - most students make flawed choices.
9	When reviewing a topic, it’s best to give learners open questions rather than multiple-choice questions or verbal summaries.	Agree	Literature on retrieval practice.
10	The best way to learn something is to go over it repeatedly within the same hour.	Disagree	Techniques such as retrieval practice and deep processing more effective than repetition.
11	Learners benefit from mixing up lots of different types of problems, rather than doing one type of task at a time.	Agree	Literature on interleaving as a study/practice technique.
12	The majority of information taught during a class will still be retained by learners 2-3 weeks later.	Disagree	Forgetting curve.

13	Learners are in the best position to judge what and how they should study.	Disagree	Metacognition; students' study habits.
14	It's always best to simplify things for learners in some way, because making something easier helps it to be processed into long-term memory.	Disagree	Literature on desirable difficulties - more difficult tasks e.g. retrieval, distributed practice can be more effective.
15	As a teacher, it is wise to wait until learners have almost forgotten things before you go over them again.	Agree	Spacing effect.
16	Multiple re-readings are more useful for learning than doing lots of tests.	Disagree	Retrieval practice is more effective than re-reading.
17	Good study advice for learners should include telling them to finding a place where they are comfortable and to do all their revision there.	Disagree	Literature on variable environmental context & memory.
18	A learner's current performance on a task is not a reliable guide to their long-term learning.	Agree	Performance v's learning distinction.
19	One of the best ways to remember something over the long term is to focus on its meaning and how it links to other things.	Agree	Deep processing; elaborative processing.

20	If a learner guesses and is not correct they may remember the wrong answer, so it’s best to avoid guessing/predictions during lessons.	Disagree	Pretesting effect; questions prior to study can boost later learning.
21	Including extra information or examples in a written passage makes it harder for learners to remember the main points.	Agree	Research into embellishment and extraneous details.
22	Ultimately, learners form new memories through frequent repetition.	Disagree	Retrieval practice more important than repetition.
23	It makes sense to do a homework task soon after the material is done in class.	Disagree	Spacing effect.
24	Once learners have got a question wrong and then been corrected, they will be able to predict whether they will get it right in future.	Disagree	Metacognition: feedback doesn’t always lead to accurate judgements of learning.

Appendix 13: Materials used in Study 4.

These are the vignette scenarios used in Study 4. Source: items 2 & 3 based on McCabe (2011; 2018); items 1 and 4-9 are novel. Note that vignettes are presented here in the order that they appeared to participants; labels (e.g. 'interleaving 1') can be used to match each vignette to the relevant data in Chapter 6, but these were not seen by participants. The overall instruction read as follows: "For each of the scenarios, please decide for yourself which answer you think is most likely, and then circle or highlight your response on the sheet. There are 9 scenarios."

Scenario 1 ('interleaving 1')

A senior Geography class is learning about lakes. Their teacher divides them into two teams. Team A looks at pictures of different types of lakes all mixed together (MIXED), each with a picture of the lake, its name, and a label saying what type of lake it is. Team B see the same information categorised by type (that is, GROUPED), so that they see all examples of one type of lake together, then the next type, and so on. Finally, the two teams are given the same test.

Please provide a prediction on the 7-point scale in reference to typical pupils.

i. Team A (MIXED) would gain much higher test scores.

ii. Team A (MIXED) would gain moderately higher test scores.

iii. Team A (MIXED) would gain slightly higher test scores.

iv. Test scores for teams A and B would be about EQUAL.

v. Team B (GROUPED) would gain slightly higher test scores.

vi. Team B (GROUPED) would gain moderately higher test scores.

vii. Team B (GROUPED) would gain much higher test scores.

Scenario 2 ('spacing 1')

Two educational psychologists are working with a group of P5 pupils to help them with maths. They have identified 10 target skills, and they have ten weekly half-hour sessions in which the pupils can practice these. In scenario A, the pupils spend an entire session focusing on just one skill, and then move on to the next skill the following week, and so on (that is, study is INTENSIVE). In scenario B, pupils look at a larger number of skills more briefly during each study session, and then return to these for further practice over the next few weeks (that is, study is SPREAD OUT). Pupils in both classes spend the same overall amount of time studying the skills. After the sessions are over, a maths test on the skills studied is given to pupils from both classes.

Provide a prediction on the following 7-point scale in reference to typical pupils.

i. Scenario A (INTENSIVE) will result in much higher test scores

ii. Scenario A (INTENSIVE) will result in moderately higher test scores

iii. Scenario A (INTENSIVE) will result in slightly higher test scores

iv. Test scores for Scenarios A and B will be about EQUAL

v. Scenario B (SPREAD OUT) will result in slightly higher test scores

vi. Scenario B (SPREAD OUT) will result in moderately higher test scores

vii. Scenario B (SPREAD OUT) will result in much higher test scores

Scenario 3 ('retrieval 1')

In two different secondary classes, a 275-word prose passage about a specific topic is presented. In Lesson A, students first study the passage for 5 minutes, and then are asked to write down from memory as much of the material from the passage as they can for a further 5-minute period (they take a TEST). In Lesson B, learners first study the passage for 5 minutes, and then are asked to study the passage again for another 5 minutes (they RESTUDY). After 1 week, all students are asked to recall as much of the passage as they can remember.

Provide a prediction on the following 7-point scale in reference to typical secondary pupils:

i. Lesson A (TEST) will result in much higher test scores

ii. Lesson A (TEST) will result in moderately higher test scores

iii. Lesson A (TEST) will result in slightly higher test scores

iv. Test scores for Lessons A and B will be about EQUAL

v. Lesson B (RESTUDY) will result in slightly higher test scores

vi. Lesson B (RESTUDY) will result in moderately higher test scores

vii. Lesson B (RESTUDY) will result in much higher test scores

Scenario 4 (‘interleaving 2’)

Two P7 classes are learning about mental health. Their teacher has prepared examples of teenagers who suffer from three key types of mental health problems. The school pupils are presented with these examples, together with suggested solutions. In class A, pupils look at examples of the same type of mental health problem consecutively (i.e., GROUPED). In class B, pupils see the examples of the three types in an intermingled fashion (that is, MIXED), such that an example of one type is followed by an example of a different type, until all examples have been presented. After viewing all the examples, the leaners are given a test with a selection of novel (previously unseen) case studies of individuals with mental health problems, and they are asked to identify suitable solutions.

Provide a prediction on the 7-point scale in reference to typical pupils.

i. Class A (GROUPED) will do much better on the test

ii. Class A (GROUPED) will do moderately better on the test

iii. Class A (GROUPED) will do slightly better on the test

iv. Test performance for Classes A and B will be about EQUAL

v. Class B (MIXED) will do slightly better on the test

vi. Class B (MIXED) will do moderately better on the test

vii. Class B (MIXED) will do much better on the test

Scenario 5 ('retrieval 2')

Two schools are running revision group ahead of end-of-year exams. In School X, learners spend their study periods reading over lesson notes, and looking at lesson slides (they RESTUDY). In School Y, learners spend each of their study periods testing themselves using flashcards (they take a TEST). A few weeks later, all pupils from both schools sit an identical exam during which they have to remember and apply the information, and they gain a percentage mark.

Please provide a prediction on the following 7-point scale in reference to typical secondary pupils:

i. School X (RESTUDY) will obtain much higher percentage exam results.

ii. School X (RESTUDY) will obtain moderately higher percentage exam results.

iii. School X (RESTUDY) will obtain slightly higher percentage exam results.

iv. Exam results for Schools X and Y will be about EQUAL

v. School Y (TEST) will obtain slightly higher percentage exam results.

vi. School Y (TEST) will obtain moderately higher percentage exam results.

vii. School Y (TEST) will obtain much higher percentage exam results.

Scenario 6 (‘spacing 2’)

Two secondary school depute headteachers are planning the S3-S4 curriculum. The depute in one school, Alpha High, plans the topics such that they are distributed across the year, with topics being partially covered and then returned to at a later date (learning is SPREAD OUT). The depute in another school Beta High, plans the topics such that each topic is covered in full within a couple of weeks, and pupil then move on to a different topic (learning is INTENSIVE). The same overall amount of lesson time is spent on the topics in both schools. Pupils are then given an end-of-year test which covers all of the topics.

Provide a prediction on the following 7-point scale in reference to typical pupils, assuming that the pupils are generally similar in every other respect.

i. Pupils at Alpha High (SPREAD OUT) will gain much higher end-of year test scores.

ii. Pupils at Alpha High (SPREAD OUT) will gain moderately higher end-of year test scores.

iii. Pupils at Alpha High (SPREAD OUT) will gain slightly higher end-of year test scores.

iv. End-of-year test scores for Alpha High and Beta High A will be about EQUAL

v. Pupils at Beta High (INTENSIVE) will gain slightly higher end-of year test scores.

vi. Pupils at Beta High (INTENSIVE) will gain moderately higher end-of year test scores.

vii. Pupils at Beta High (INTENSIVE) will gain much higher end-of year test scores.

Scenario 7 ('spacing 3')

Two computer science classes are learning coding skills. In one class, Class A, the teacher presents new coding processes, and these are then practiced several times within the same lesson (that is, the learning is INTENSIVE). In the other class, Class B, the same coding processes are practiced across several lessons, only once per lesson (that is, the learning is SPREAD OUT). The same overall time is spent on the terms by both classes. At the end of the topic, both classes are given the same test.

Please provide a prediction on the following 7-point scale in reference to typical pupils:

i. Class A (INTENSIVE) will gain much higher test scores.

ii. Class A (INTENSIVE) will gain moderately higher test scores.

iii. Class A (INTENSIVE) will gain slightly higher test scores.

iv. Test scores for classes A and B would be about EQUAL.

v. Class B (SPREAD OUT) will gain slightly higher test scores.

vi. Class B (SPREAD OUT) will gain moderately higher test scores.

vii. Class B (SPREAD OUT) will gain much higher test scores.

Scenario 8 ('retrieval 3')

Two similar classes of pupils are learning terminology for their latest topic. In one class, Mrs Smith shows the pupils the terminology one term per slide on a Powerpoint, and then shows the same Powerpoint two more times in follow-up lessons (that is, they RESTUDY). Mrs Jones shows the pupils the terminology one term per slide on a powerpoint, and then tests them on the items two times in follow-up lessons (that is, they take a TEST). A couple of weeks later, both classes are given a multiple-choice quiz on the terminology.

Please provide a prediction on the following 7-point scale in reference to typical secondary pupils:

i. Mrs Jones's class (TEST) will gain much better scores on the quiz.

ii. Mrs Jones's class (TEST) will gain moderately better scores on the quiz.

iii. Mrs Jones's class (TEST) will gain slightly better scores on the quiz.

iv. Test scores for both classes will be about EQUAL

v. Mrs Smith's class (RESTUDY) will gain slightly better scores on the quiz.

vi. Mrs Smith's class (RESTUDY) will gain moderately better scores on the quiz.

vii. Mrs Smith's class (RESTUDY) will gain much better scores on the quiz.

Scenario 9 ('interleaving 3')

A visiting biologist presents pictures of butterflies to 11-year-old pupils in two schools. She shows the children four examples each of 16 species of butterfly. In School A, she shows all four examples of a single species consecutively and then moves on to examples of the next species, and so on, until all pictures have been presented (the images are GROUPED by species). In School B she presents the various species in an intermingled fashion, such that an example of one species is followed by an example of a different species, until all pictures have been presented (the images are MIXED). After viewing all the pictures, children are given a test that requires them to correctly identify previously presented pictures of the butterflies.

Please provide a prediction on the following 7-point scale in reference to typical pupils.

i. Pupils at School A (GROUPED) will get much higher test scores.

ii. Pupils at School A (GROUPED) will get moderately higher test scores.

iii. Pupils at School A (GROUPED) will get slightly higher test scores.

iv. Test scores for Schools A and B will be about EQUAL

v. Pupils at School B (MIXED) will get slightly higher test scores

vi. Pupils at School B (MIXED) will get moderately higher test scores

vii. Pupils at School B (MIXED) will get much higher test scores

Appendix 14: Tackling myths, a teacher short course.

Appendix 14, part 1: A description of a course that aims to tackle myths and misconceptions about memory and learning among the teaching profession.

Module name

Memory, Belief and Misconception (20 credit points)

Module Description and Rationale

This module will combine psychology and education concepts in a way that is highly applicable to students across the Faculty, demonstrating that learning is underpinned by scientific processes, and how these processes are often highly counterintuitive to learners. The module will provide a sector-leading insight into contemporary research on human long-term memory combined with insights into the metacognitive processes which often lead learners and instructors astray.

The module will be enhanced by the application of Strathclyde research into memories and metacognition, including investigations of metacognitive processes among teachers and children. This work has identified key areas where beliefs about learning show importance flaws, and these insights apply to multiple professions which engage with training or other learning processes.

This module is of particular relevance to teachers, but will also be helpful for everyone who will be involved in learning or training in a vocational setting or who need to rapidly and accurately take in new information or understand these processes in others, such as students of architecture, business, journalism, or psychology.

This class is being offered as a BA2 IDL module. It will show participants how memory and belief interact, and illustrate areas where both professionals and school children typically experience misconceptions about memory and learning. A rich body of scientific research has identified common misconceptions about how the human memory works; for example, members of the general public tend to endorse the idea that memory works like a video camera. Many people endorse ‘neuromyths’ such as the idea that people use only 10% of their brain, or that everyone has their own ‘learning style’. Students, too, show biases in their learning behaviour that indicate false beliefs, such as failing to make allowances for forgetting, as well as underestimating the benefits of practice. Understanding these biases and developing a more scientific understanding of how memory works is an important area of professional knowledge for educators and for other professionals who are involved in training or staff development.

The study of misconceptions about thinking and memory has grown since the seminal work of Kahneman & Tversky in the 1970s, which indicated that human intuition is highly unreliable, and is currently an active area of research in the psychology of education.

As well as education students, this module has a clear relevance to law, given the importance of memory to the legal process - many of the misconceptions and memory flaws relate to judges and expert witnesses. It would also be of benefit to students of social policy or governance, and to anyone who has a professional interest in training, such those who wish in the future to develop in-house courses or online training programmes, as well as being useful in helping participants to better understand memory processes for their own studies.

Learning Outcomes

The intended outcome is that participants will be able to engage in an evidence-based reflection on beliefs about learning processes in real contexts. They will be able to:

1. Explain key aspects of how human memory functions;
2. Explain the role of beliefs and metacognition in educational contexts;
3. Identify examples of beliefs about learning and memory, including misconceptions;
4. Discuss the implications of these for learning.
5. Accurately use research-based concepts and terminology such as ‘metacognition’, ‘long-term memory’, ‘working memory; and ‘beliefs’.

Participants will be expected to share their ideas with classmates during tutorials, and to reflect on their own experiences as a student or in workplace settings. They will be expected to independently find and analyse examples of statements about learning and memory relevant to their own course/chosen profession.

Assessment

The assignment should include two examples of misconceptions which you have independently identified (not, for example, misconceptions listed/quoted in research paper). Finding and analysing misconceptions in their original context is part of the task.

Appendix 14, part 2: A short article based on the course. Note that this is a first draft, written by J. Firth alone. The final draft was co-authored and published in September 2020: Firth, J. & Zike, J. (2020). Tackling learning myths among trainee primary teachers: A case study. Impact, 10, 20–22.

Tackling learning myths among trainee primary teachers: A case study.

by J. Firth and J. Zike

Myths and misconceptions about learning are pervasive among the teaching profession. Myths such as learning styles or the idea that some pupils are 'left brained' or right brained' have been endorsed by over 90% of participants in some studies (Howard-Jones, 2014). This case study describes the development and evaluation of a short course on memory and misconceptions which was trialled on a group of second-year undergraduate trainee primary teachers in Scotland.

Myths and misconceptions

By 'myths', we mean popular ideas about learning which are not in line with the consensus among researchers. Many learning myths represent a flawed attempt to explain individual differences in learner attainment. Probably the best known is the concept of VARK (visual, auditory, reading-writing and kinaesthetic) 'learning styles'; despite a lack of evidence that learners can be so simplistically categorised, teachers and learners alike may choose to refer to themselves as 'visual learners', and so forth. Reviews (e.g. Pashler et al., 2008) have shown that there is no reliable evidence that allocating particular individuals in this way improves attainment. Attempting to cater for learning styles via specific types of tasks is therefore at best pointless, at worst counterproductive.

More broadly, there are many misconceptions about how learning and memory function – misconceptions that are important if educators are to make choices that are in the best interests of their students. Teachers often lack a clear understanding of how new learning takes place, tending to assume that short-term performance equates to learning (and in doing so, neglecting the importance of varied, spaced practice). In fact, short-term performance often correlates negatively with learning; variations and delays reduce performance but improve learning over the long term (Soderstrom & Bjork, 2015). Learning is also often seen as a one-way input process, with little intuitive appreciation of the role of retrieval practice in consolidation.

How best to tackle these issue? In an optional short course offered to undergraduate student teachers, we took a two-pronged approach. Firstly, we tried to developed an improved understanding of the workings of metacognition and memory. This is important, because it's hard to make sense of misconceptions without knowing their context in terms of human cognition. It helps, for example, to better understand how learners form and store long-term memories; a failure to understand these issues may be affecting decision-making in the classroom, and therefore impacting on the progress and success of school pupils. However, the development of a metacognitive understanding of memory is not always considered a priority when preparing trainee practitioners. Long-term memory is complex and often counterintuitive in its functioning (Bjork, 2011), and this understanding does not appear to develop spontaneously through time in the classroom (Firth, 2018).

Secondly, we tackled the myths and misconceptions head on. Each weekly task focused on a specific issue and explained to course participants why it was a misconception. There is evidence that doing so is important; in one study, Will et al. (2019) found that directly refuting flawed beliefs (for example, learning styles) is more effective than simply providing contrary evidence. That is to say, it is important to state "x is wrong, and here is why". In our course, each weekly task focused on a specific issue, and we explained to course participants why it was a misconception, drawing on psychology and education research.

Findings

It can be difficult to transfer learning from training to real situations, and for this reason, we felt that it was important to give our participants the chance to identify myths in authentic educational contexts, rather than speaking about them in more abstract terms. Each participant was therefore asked to find example documents which included examples of misconceptions, and use their knowledge from the short course to explain why the ideas were flawed. Class time was allocated for discussion and peer feedback on this process, and the issues chosen were broad ranging. Many participants were keen to analyse contemporary education trends, including growth mindset and play-based learning.

Example documents which participants accessed for this purpose included school teaching & learning policies, popular blogs that provided homework advice to parents, and revision guides for students. Most commonly participants accessed sources on the web, although they had been told that printed sources were also acceptable. A likely reason for this is that it is simply so easy to find examples of myths and misconceptions using search engines. This search process therefore helped to raise participant awareness of how pervasive certain misconceptions are.

Analysis of sources often revealed a mixture of good and bad advice. For example, one text drawn from educationcity.com and discussed during our second session advised pupils to revise in short sessions with lots of breaks – an idea that fits well with the spacing effect – but also unhelpfully categorised them in terms of learning styles, advising some to voice record their material (auditory) and others to use post-it notes (visual). Overall, it was rare to find a source that didn't include at least one misconception – although most also included plenty of good advice too.

Participants became highly enthusiastic and aware of the issues, and by the end of the course they were able to demonstrate skill in analysing myths and misconceptions in the context of real-world documents. Some even felt embarrassed about their previous endorsement of the ideas. Among the ideas that participants took on board most completely were:

- The learning styles myth

- The left-brain right-brain myth

- The importance of spacing out one's practice

- The idea that learning is not just about passive input (retrieval being more valuable than repetition)

Participants did, however, also retain some misconceptions. It was notable that even when refuting ideas like VARK learning styles, many still expressed the view that 'everyone learns differently'. The idea that short-term performance does not always indicating learning was also hard for many to grasp; participants frequently expressed the view that it would be possible to see whether pupils had learned something after observing a single lesson. More broadly, even when endorsing the importance of relying on research evidence, many participants expressed their discomfort with evidence-based strategies in intuitive terms: 'it doesn't feel like that would be helpful', for example.

One interesting point of discussion was the way that the participants expressed concern about sharing their knowledge with in-school colleagues. The course participants were trainee teachers and most were young; there was a pervasive feeling that challenging myths and misconceptions in the workplace would be very hard for them to do, and that most of their colleagues would not welcome a discussion of such issues. In discussing why myths can be so hard to overcome, we considered the psychology of identity, and the way that a teacher's beliefs can become part of who we are as educators. This may cause any challenge to our beliefs to feel threatening.

As such, we feel there would be much to be gained from workplaces tackling myths and misconceptions directly, as we did during this course, but it is important to recognise that to do so effectively, some attention needs to be given to the value of a metacognitive understanding of learning as part of professionalism. If teachers see the understanding of learning and memory as part of their role, they may be more open to discussing and analysing the evidence, as well as the myths and misconceptions.

References

Bjork, R. A. (2011). On the symbiosis of remembering, forgetting, and learning. In A.S. Benjamin (Ed.) Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork (pp. 1–22). New York: Psychology Press.

Firth, J. (2018). Teachers' beliefs about memory: What are the implications for in-service teacher education? Psychology of Education Review, 42(2), 15–22.

Howard-Jones, P. A. (2014) Neuroscience and education: Myths and messages. Nature Reviews Neuroscience, 15(12), 817–824.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. Psychological Science in the Public Interest, 9(3), 105–119.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, 10(2), 176–199.

Will, K. K., Masad, A., Vlach, H. A., & Kendeou, P. (2019). The effects of refutation texts on generating explanations. Learning and Individual Differences, 69, 108–115.

Appendix 15: Practitioner enquiry conference session (Beck et al., 2020)

This is a submitted conference abstract for the EERA 2020 conference which was later cancelled due to the covid-19 crisis. Reference: Beck, A., Wall, K., Firth, J., Tonner, P., Arnott, L. (2020, August 24–28). Developing a practitioner enquiry approach to school-university research partnership [Conference paper]. EERA Conference, Glasgow, United Kingdom. https://eera-ecer.de/ecer-2020-glasgow/ (Conference cancelled).

Developing a practitioner enquiry approach to School University Research Partnership

by A. Beck, K. Wall, J. Firth, P. Tonner and L. Arnott

Abstract

In 1904, Dewey first discussed the importance of teachers engaging in pedagogic enquiry to fully engage with processes and outcomes in their classrooms. Since then the concept has been in and out of fashion and more or less tied up with the concept of the research engaged practitioner. Emerging practice in Scotland is therefore a useful case to explore as the new National Model of Professional Learning has ‘learning by enquiring’ as one of three main strands of professional learning (Education Scotland, 2019). The dominant approach from this policy draws on the work of Cochran-Smith and Lytle (2009) with ‘inquiry as stance’ being a common phrase, whereby inquiry becomes part of a teacher’s professional identity with every aspect of professional practice and the curriculum as a whole becoming potential subjects for inquiry and professional scrutiny. However, Wall (2018) noted that this epistemological tradition of practitioner enquiry is contrasted with practices that are often more ‘project based’ whereby teachers are focused on issues of method and data in relatively isolated enquiries. This means research engagement tends to be a one off and have more in common with a traditional research project than what Stenhouse (1981) proposed.

Underpinning these debates has often been an assumption that practitioner enquiry will naturally lead to an engagement with research as a means to generate answers to pertinent questions of practice (Nias and Groundwater-Smith, 1988). For many this position naturally involves the participation of university academics to facilitate this engagement (Baumfield & Butterworth 2007; McLaughlin & Black-Hawkins, 2004) and Timperley (2008) states an important role for the expert (although not necessarily university-based) in facilitating professional learning and providing critical support. This paper therefore looks to explore five different models of implementing a practitioner enquiry approach to research engagement in Scottish schools and nurseries when working in partnership with a university research team.

We will explore five different, locally developed, school university research partnerships (Thornley et al. 2004) all have their impetus in the Scottish education drive towards research engagement and enquiry, but each have found their own path to making this practicable within their own community of professional learners. We will look at the role of the university (and other organisations where appropriate) in the partnership, the structures, formal and informal, that facilitate professional learning through enquiry, and how supportive spaces for dialogue are created spanning what might be called a third space for learning (Reeves and Drew, 2013). The affordances and constraints of the school-university research partnership will be evaluated in regard the model of research encouraged, the balance between enquiry process and outcomes, as well as their inclusivity, sustainability and the extent to which their support teacher’s voice and metacognitive awareness (Wall and Hall 2016).

Methodology

A case study approach (Yin, 2005) was used to examine practice in four schools and one family centre. The cases were chosen based on their commitment to a school university research partnership with the University of Strathclyde, where the intent was for all staff or a smaller group to undertake collaborative practitioner enquiry using a model of practice loosely based on Baumfield et al. (2012) and Hall and Wall's (2018) approach to professional learning.

In each case a visual model of the partnership was developed and validated, based on participants' practice and experience over one academic year. These models attempt to show the partnership, who was involved and how this was maintained over time. This negotiated modelling was complemented by participatory observations and a presentation conducted by colleagues from each setting on 'their professional learning' through the process.

Secondary school 1:

This partnership targeted the whole school community through the concept of a research hub. Input from the university was through whole school development time and predominantly used coaching techniques to scaffold the teachers' practitioner enquiry projects. The interaction was regular and iteratively developed in dialogue.

Secondary school 2:

A university postgraduate certificate formed the structure of this partnership, with a small group of staff engaged in three 20 credit modules on supporting teacher learning. The modules were taught in school and teachers engaged in practitioner enquiry to support colleagues' learning therefore opening up the potential of expansion.

Secondary school 3:

A Teacher Learning Programme (TLP) run by Education Scotland formed the basis of the third case study with the university input providing coaching support for the participating teachers. The dialogue and enquiry was strongly scaffolded by TLP and so the university team acted as more of critical friend against this backdrop.

Secondary school 4:

This model was strongly influenced by the practice of school 3, but without the impetus of TLP. A group of staff looked to independently develop a practitioner enquiry group, coached by the university team, which the aim to eventually expand to a wider group of staff.

Nursery/family centre:

4 small teams of staff undertook collaborative practiotionr enquiry within the family centre. The impetus was previous engagement with the university, but the questions explored were generated around the family centre's development plan. This case was characterised by a much stronger engagement from the children that in the other 4.

Conclusion

Analysis is on-going, but we hope that the use of visual models will help us to establish the key structures and processes characterized by each of the school university research partnerships that form the five case studies. This will allow us to make comparisons and explore similarities and differences across the contexts and associated influences. Hall and Wall (2019) suggested four principles of a practitioner enquiry culture (autonomy, disturbance, dialogue and connectivity) and our discussion will use these principles to engage with the key characteristics of the practitioner enquiry community created within the partnership and the professional learning that resulted. Key to our conclusions will be the concept of partnership and a critical engagement with the roles played by practitioners in each setting and the

university team as they came together under the guise of practitioner enquiry. We will aim to explore dynamics of power, voice and inclusion as well as assessing the potential for sustainability over time in an attempt to draw out potential guidance for other settings embarking on a similar professional learning journey.

Key words: Practitioner enquiry, professional learning, school university partnership, research engaged

References

Baumfield, V., & Butterworth, M. (2007). Creating and translating knowledge about teaching and learning in collaborative school–university research partnerships: An analysis of what is exchanged across the partnerships, by whom and how. Teachers and Teaching: Theory and practice, 13(4), 411–427.

Baumfield, V., Hall, E., & Wall, K. (2012). Action research in education: Learning through practitioner enquiry. Sage.

Cochran-Smith, M., & Lytle, S. L. (2009). Inquiry as stance: Practitioner research for the next generation. New York: Teachers College Press.

Education Scotland (2019). The national model of professional learning. Education Scotland website. Retrieved 26 September 2019 from https://professionallearning.education.gov.scot/about/the-model-of-professional-learning/

Hall, E., & Wall, K. (2019). Research Methods for Understanding Professional Learning. Bloomsbury Publishing. Mclaughlin, C., & Black-Hawkins, K. (2004). A schools-university research partnership: Understandings, models and complexities. Journal of in-service education, 30(2), 265–284.

Nias, J., & Groundwater-Smith, S. (Eds.). (1988). The enquiring teacher: Supporting and sustaining teacher research. Routledge

Reeves, J., & Drew, V. (2013). A productive relationship? Testing the connections between professional learning and practitioner research. Scottish Educational Review, 45(2), 36-49.
Stenhouse, L. (1981). What counts as research?. British journal of educational studies, 29(2), 103–114

Thornley, C., Parker, R., Read, K., & Eason, V. (2004). Developing a research partnership: Teachers as researchers and teacher educators. Teachers and teaching, 10(1), 20–33.

Timperley, H. (2008). Teacher Professional Learning and Development. Educational Practices Series-18. UNESCO International Bureau of Education.

Wall, K., & Hall, E. (2016). Teachers as metacognitive role models. European Journal of Teacher Education, 39(4), 403–418.

Wall, K. (2018). Building a bridge between pedagogy and methodology: emergent thinking on notions of quality in practitioner enquiry. Scottish Education Review Journal.

Yin, R. K. (Ed.). (2005). Introducing the world of education: A case study reader. Sage.

Appendix 16: The teacher's guide to research (Introductory Chapter to a book on teacher research engagement) (Firth, 2019)

The following text is the introduction chapter of my solo-authored book, 'The Teacher's Guide to Research'. The aim in writing the book was to empower and encourage other teachers to engage with research evidence in their classroom work. Reference: Firth, J. (2019). The teacher's guide to research: Engaging with, applying and conducting research in the classroom. Routledge.

The Teacher's Guide to Research:

Engaging with, Applying and Conducting Research in the Classroom.

by J. Firth

Who is this book for?

This book is for and about teachers. It's primarily aimed at the classroom practitioner rather than managerial staff or other higher authorities. Why? Because in my view, engaging with research is a key part of teacher professionalism. It is the teacher who enacts changes in the classroom, and by informing themselves about research, teachers can select and investigate ideas that make sense in their context.

Taking a research-informed approach to teaching is a matter of professional knowledge and skill, and potentially empowering. However, there's no doubt that some people have mixed feelings about this. 'Why do I need research', they might ask, 'when I already know how to teach?'. This question has led me to a careful consideration of how and why research engagement can be useful to the classroom teacher.

If education is not informed by research, what is it informed by? All too often the answer will be traditions and assumptions, some of which will be flawed (and we won't even know which ones).

Can a classroom teacher benefit from engaging with and taking part in research activities? In my view, the answer is yes—a research-engaged teacher has a deeper understanding of how learning works, and is in a better position to make changes to the benefit of pupils.

This book works through the process of informing yourself about research, using research, contributing to the research community, and sharing the outcomes of your research and conclusions with others as part of a broader movement.

How to engage with research

Engaging with research as a teacher does not necessarily mean that you have to conduct and publish your own research studies. It may be best, especially in the early stages, to focus more on understanding the work that has been done by others, as such evidence (especially that which is recent and high quality) will provide both a model of how to conduct research, and an introduction to current theories and debates. However, one of the best ways of engaging with any field of study is to participate in it, and there comes a point where you may wish to try things out for yourself, or attempt to answer questions that have not been satisfactorily dealt with in the research literature to date.

Answering questions and contributing to knowledge is not the only reason for conducting research in your classroom. It can also constitute an effective form of CPD, being highly motivating and informative, and prompting professional reading and practical engagement with current issues.

How can teachers take control of their development

To say that teachers can and should be evidence-informed is a challenging statement, and one which suggests a shift in the nature of the role. Research can be an empowering process, allowing a teacher to change aspects of their practice on the basis of evidence. However, in order to be empowering, the teacher must remain the decision maker. A fundamental dilemma in the education system as a whole is that while governments and other external bodies want improve teacher professionalism, doing so in a top-down—where teachers are told what to do and how to do it—is fundamentally disempowering, and therefore harmful to the processes as a whole. So one initial principle is that in order for evidence-based practice to work, the teacher needs to at least have a role in setting the research agenda, and needs to be able to act on the outcome of the research. If neither of these principles hold, then teacher research engagement is reduced to a hollow, tick-box exercise.

What about research that does not study learning?

Although the emphasis in this book is on research which can directly affect your teaching practice (with such work often termed action research or practitioner enquiry), there is no reason that teachers need to be limited to researching issues of teaching and learning. There is an entire chapter (Chapter 13) on research within your subject discipline, which also includes guidance on encouraging and supporting pupil-led research projects.

How this book is structured

Each chapter covers a key aspect of research engagement, in an order which moves from information (for example, how to find and read journal articles), via increasingly elaborate practical engagement, and finally on to sharing research and establishing research systems in your school.

Each chapter also explores an educational case study, includes a summary of a key issue from the research literature, discussion points for staff reading groups, and a suggestion for a practical project.

Later in the book, you will also find a glossary of terms.

There are three main parts in the book:

Part 1: The place of research in schools

This part looks at how and why teachers access research, overcoming barriers including the issue of finding enough time, and explores how to find, apply and evaluate research-informed changes to practice. It answers some of the big questions about engaging with research: why, when, what, who and how:

Why: Chapter 1, why teachers should engage with research

When: Chapter 2, finding the time for research

What: Chapter 3, accessing and using research evidence

Who: Chapter 4, will this work for my learners?

How: Chapter 5, using a research-based intervention in your classroom, and Chapter 6, evaluating your intervention.

Part 2: The teacher as researcher

This part describes how to ensure that teacher-led research projects are ethical and valid, and sets out the various methodology options available for running your own full-scale research study.

Ethical practice and an understanding of ethics is relevant at any point, but it seemed to make sense to place the ethics chapter before the parts on data gathering and running research studies. Having said this, as a research-engaged teacher you would benefit from reading this chapter at any point in the process.

Part 3: The networked teacher

This part focuses on the systems and networks that can support a teacher who is becoming research-engaged, including the establishment of research centres in schools, local groups and issues around sharing and disseminating research.

Appendix 17: Short piece for psychology teachers

This was an invited piece for the magazine of the Association for the Teaching of Psychology, a UK-wide network of psychology teachers. It focused on how to apply spacing and interleaving to the teaching of that subject in pre-university settings. Reference: Firth, J. (2019). Spacing and interleaving in the psychology classroom. ATP Today Magazine, February, 10.

How spacing and interleaving can be applied to psychology topics

by J. Firth

We psychology teachers are in a great position to access psychological research and to apply it to practice. Over recent years, I've been working on applications of long-term memory to the classroom, and two key areas that I have focused on are spacing and interleaving — techniques which are evidence-based and which are becoming more popular throughout education nowadays. So what exactly would spacing and interleaving look like in the psychology classroom?

The **spacing effect** means that when study and re-study are separated by a delay, learning is more effective. A lot of the research into this effect has used language vocabulary, and so an obvious link to psychology would be to use spacing for terminology, for example when teaching research methods. This could involve learning a list of terminology in one session, and then practicing it or doing a quiz during sessions that follow days or even weeks later.

It's really important to point out that for spacing to work, the information needs to be well learned in the first session. The learners need to have mastered it. As Rawson and Dunlosky (2011) put it, "*our prescriptive conclusion for students is to practice recalling concepts to an* *initial criterion of 3 correct recalls and then to relearn them 3 times at widely spaced intervals*" (p. 283).

Although it's counterintuitive to think that waiting longer before restudy would be helpful, an element of forgetting actually seems to help, in comparison to following up on something more quickly (contrary to what might be predicted from the multi-store model of memory!). A useful analogy is to imagine painting a wall — it's better to wait, because there's no point in applying the second coat until the first one has dried.

Other areas where spacing might be useful include remembering methodological detail from research studies, or evaluating these studies. Again, a degree of forgetting is likely to make the later study sessions more effective — they will serve the purpose of reminding students, which appears to be more effective than consolidating learning which is still fresh in their minds. It's worth noting, though, that learners and teachers alike tend to assume that a delay is a bad thing, and therefore opt to revise material too soon (Firth, 2018; Zechmeister & Shaughnessy, 1980).

The term interleaving refers to varying the order of a set of tasks or examples, whereby each item is immediately followed and preceded by an example of a different concept rather than appearing in 'blocks' of the same type (which is termed a 'blocked' arrangement). For example, when studying factors that affect conformity, a teacher could present multiple real-world examples of the same factor, or could instead mix up show examples of multiple different factors. The latter might sound confusing, but according the discrimination-contrast hypothesis, it's best if easily-confused examples are seen side by side so that learners are able to notice key difference between them (Birnbaum et al., 2013).

There are two main things you can interleave — initial learning or practice, and I will focus on practice here. The benefits of interleaved practice have best been demonstrated for high school maths and are therefore likely to apply to statistics in psychology, though a similar idea could be applied to any set of short questions. It is most useful when items are similar in some way; this allows learners to make conceptual links, and helps them to notice differences between items that are easily confused in a way that they may fail to do if the examples were presented across different lessons. Mixing in questions from topics studied earlier also helps to provide consolidation, and improves learners' ability to correctly interpret later exam questions (Rohrer et al., 2015).

Clearly some of the interleaved practice described above will also increase spacing — if you mix in questions based earlier lessons, there is also a time delay! These two concepts can therefore be used in combination. Group discussions, essays or individual projects which build upon previous learning could prompt spacing and interleaving, too, especially with tasks that combine more than one topic (for example, analysing the role of sleep and stress in human memory). Low-stakes review quizzes which are conducted after a long break (e.g. after the summer holidays) will also benefit from both effects, especially if questions are in a mixed order rather than being categorised into sections.

References

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.

Firth, J. (2018). Teachers' beliefs about memory: What are the implications for in-service teacher education? *Psychology of Education Review, 42*(2), 15–22.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough?. *Journal of Experimental Psychology: General, 140*(3), 283–302.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*, 900–908.

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*, 41–44.

Appendix 18: Supporting beginning teachers to link learning, memory and inquiry (chapter).

An important way of applying an understanding of memory and metacognition to the teaching profession is via mentoring. In part, this is necessary because human memory is counterintuitive, and – as shown from the data in this thesis – staff are unlikely to come to a full understanding of its workings via experience alone. I addressed this issue through this invited chapter, which is aimed at mentors to beginning science teachers, and guides the application of techniques such as the spacing effect, as well as advising on common misconceptions. Reference: Firth, J. (in press). Supporting beginning teachers to link learning, memory and inquiry. In Salahjee (Ed.), Mentoring science teachers in the secondary school: A practical guide. Routledge.

Supporting beginning teachers to link learning, memory and inquiry

by J. Firth

Introduction

The situation facing a beginning science teacher is challenging. Teacher preparation is a rapid process, with teachers taking responsibility for groups of learners as soon as they complete their Initial Teacher Education (ITE), or even before. A beginning teacher is not exactly put in at the deep end, but they do need to learn quickly if they are not going to find themselves out of their depth. During ITE and the first few years of their careers, beginning teachers need to be guided to develop professional skills and knowledge.

However, such professional learning can be a haphazard or flawed process without suitable guidance. Mentoring in ITE and the early years will help to ensure these beginning teachers develop into effective practitioners.

For a beginning teacher to make effective decisions when planning lessons requires more than just developing a repertoire of suitable classroom tasks. Mentors have a responsibility to support their beginning teacher in planning, and a beginning teacher needs to know which actions to take and when to carry them out. These decisions must be tailored to particular learners or groups. Such aspects of teaching skill are diagnostic and require an accurate analysis of evidence about learners' progress (see Chapter 10 on planning for pupils' learning).

A teacher's professional knowledge underlies their ability to make rapid and appropriate decisions on an ongoing basis, both when planning lessons and activities, and during the moment-to-moment running of a class. One area of expertise that is particularly relevant to these decisions is an understanding of human memory – how learners process incoming information and form it into a coherent, interconnected representation that they can later recall and use.

This chapter will explain some of these fundamental principles of memory and learning, and look at how they play a role in the classrooms. It will explore some misconceptions that a beginning science teacher may have about memory. Finally, it will discuss how mentoring can help to set a beginning teacher upon a path of inquiry to make their practice more effective.

Objectives

At the end of this chapter, you should be able to:

- Understand the importance of placing a focus on effective teaching and successful learning during mentoring sessions with a beginning teacher;
- Guide your beginning teacher to discuss their own beliefs about memory and learning;
- Support a beginning teacher to incorporate learning and memory-related processes of desirable difficulties in their practice;
- Guide your beginning teacher to inquire into memory and learning processes using classroom evidence;

1. Effective teaching and successful learning in science classrooms

Professional knowledge underlies effective, evidence-based practice. In the same way that doctors need technical knowledge of principles such as infection and the immune response, so teachers have a knowledge base that underpins their classroom actions. For science teachers – indeed, for any educator – a vital aspect of this knowledge is an understanding of *how learning works*, such that they can make optimal decisions on issues such as what information to present, and when and how to present it.

Placing the focus on learning processes during your mentoring sessions makes it easier for a beginning teacher to think systematically about effective teaching. Any classroom action or decision taken by a beginning teacher could have either a positive effect or a negative effect; it could be beneficial over the short term, the long term, or not at all. For example, a task that aims to review ionic bonding could be carried out too soon or too late to have the maximum benefit for pupils as learners, while a maths practice exercise could be either too short or too long in terms of time efficiency (Rohrer & Taylor, 2006). Some classroom activities might not affect learner understanding at all.

Teachers' thinking and decisions about their pupils' learning is a form of *metacognition,* i.e. thinking about thinking. More specifically, it draws on beliefs about the cognitive processes that underpin learning. Metacognitive beliefs have consequences for a beginning teacher because they guide minute-by-minute decisions both during planning and when in the classroom. An understanding of learning processes can, therefore, be seen as a key part of the beginning teacher's professional toolkit (Firth, 2017). As a mentor, you can share some classroom examples with your beginning teacher to help them recognise that immediate performance is not the same thing as learning, and that learning cannot be judged on a pupil's performance in a single lesson (Soderstrom & Bjork, 2015). For example, a pupil may leave the room full of confidence after a lesson on circuits, but two weeks later be unable to explain the difference between series and parallel circuits.

It is safe to say that one aim of a beginning teacher is to ensure that pupils remember the material they have been taught, and that they will already have some awareness of the risk of forgetting. However, this professional understanding is still at an early stage and will be characterised by misconceptions. Case study 5.1 presents one example of this issue, and the broader state of a beginning teacher's understanding of learning and memory will be discussed in the next section.

Case study 5.1 Paul's mentoring sessions

Paul is a deputy head teacher of a large school and has responsibility for running weekly sessions with beginning science teachers. These beginning teachers are given a short academic article about the psychology of memory to read in advance of each week's session, which is then discussed at the beginning of the session. The focus of each session is an aspect of learning that is commonly neglected or misunderstood.

This week, beginning science teachers are looking at the timing of written practice tasks which are used to consolidate laboratory work. After discussing the research article, Paul asks the beginning teachers to briefly explain how and when they would carry out written consolidation of practical experiments in Chemistry.

Paul has noticed that even confident and capable beginning teachers have misconceptions about how and when pupils should consolidate learning, and about how quickly forgetting takes place. In their planning processes, most of the beginning teachers suggest that reviewing at the end of a lesson and via an immediate homework task would be optimal, and that further review of the taught topic could then be left until exam time.

Instead, Paul recommends that reviews are delayed beyond the individual lesson and that consolidation is built in regularly during the academic year, with older material being explicitly linked to newly-learned concepts. Paul says, ‘It is difficult to judge how quickly pupils will forget information they appear to have mastered in a single class’.

Now complete Task 5.1 which asks you to discuss Case study 5.1.

Task 5.1 Mentor reflection: Discussions about effective learning

Having read Case study 5.1, consider the following:

1. Do you agree with the points made by Paul?

2. What other advice would you give to your beginning teacher?

3. What misconceptions about learning have you noticed among your beginning teacher?

4. As an experienced teacher and mentor, how can you be sure that you do not have your own misconceptions about how memory works?

2. What do beginning science teachers believe about memory and learning?

From a psychological perspective, learning involves long-term changes in what people think or what they can do. Such changes rely on the formation or strengthening of long-lasting memory traces which can then be *transferred*, i.e. flexibly applied to novel situations (Bjork, 1994). An event that leads to a pupil improving their skills, understanding or knowledge can be linked to changes which are both neural and psychological, i.e. affecting the structure of brain cells in a way that is reflected in what the pupil can think, do and understand. Of course, such learning cannot be reduced to memorising separate pieces of information; memory is based around a set of meaningful, interconnected schemas (Bransford et al., 2000).

Unfortunately, many features of memory can be counterintuitive, and are not easily understood through life experience alone. This was demonstrated in a survey of the general public by Simons and Chabris (2011), who found that most people endorsed a set of false statements about memory (for example, that memory works like a video camera in that it accurately records the events we see and hear for later inspection) none of which were accepted by memory researchers. This finding shows that many common assumptions about memory conflict with the scientific consensus.

To support your beginning teacher, complete Task 5.2, which asks them to analyse their own understanding of memory and asks you to facilitate their thinking about these issues.

Task 5.2 A beginning teacher's beliefs about memory and learning

With your beginning teacher, discuss the idea raised by Simons and Chabris (2011) that the general public often have flawed beliefs about how memory works.

Next, ask your beginning teacher to take the following quiz:

Which of these things are true?

1. Human memory works like a computer hard-drive.
2. New terminology does not enter memory unless you pay attention to it.
3. Revision for a science exam mainly involves short-term memory.
4. Memorisation of new ideas primarily depends on the rapid repetition of information.
5. Learning a new science fact involves creating a file in long-term memory which is new and separate from existing knowledge.

After they have completed the test, discuss their answers along with the suggested answers at the end (Appendix 5.1) in a mentor-mentee meeting.

You may discover from the above discussions that your beginning teacher has considerable flaws in their understanding of memory, leading to misjudgements of how best to promote learning among pupils. If this is the case, you should help your beginning teacher to recognise that several aspects of learning are not obvious.

Such misconceptions start early; pupils' self-directed learning tends to be characterised by a preference for flawed study strategies such as re-reading and highlighting texts. In addition, therefore, it would be beneficial for the beginning teacher if you could have a conversation with them regarding effective study guidance for their classes, with an awareness that pupils do not spontaneously come to perceive the limitations of flawed study strategies, or begin to adopt better strategies without guidance (Rohrer & Pashler, 2010).

You could consider asking your beginning teacher to list what they think are advantageous strategies for learning among pupils and what they think are not, and why. You can then extend the discussion in relation to what pupils believe are advantageous learning strategies.

This kind of professional dialogue can be supported by relevant literature that demonstrates metacognitive flaws in how we understand learning. For instance, in an experiment which presented sets of images to learn, Kornell and Bjork (2008) found that 85% of learners did as well or even better when different categories of images were mixed (or 'interleaved'; see below) rather than separated into blocks of the same type, but 83% of the participants incorrectly believed they had done at least as well in the blocked condition. Discussing such research with your beginning teacher will help to make them aware of a striking mismatch between beliefs about learning and the reality of how well learners actually do in tasks, as well as showing the limited effects of direct experience.

It would be helpful for a beginning teacher if such flawed ideas about effective pupil learning strategies were corrected during the process of Initial Teacher Education (ITE). However, given the considerable variability in what is taught to beginning teachers (Carter, 2015), it is unlikely that the profession can rely on this. There is no guarantee that what beginning teachers have previously been told about learning, for example by university tutors, will concur with the scientific consensus, because even experienced professionals seem to be prone to errors. For example, Morehead et al. (2016) found that while university instructors held slightly more accurate views of learning and memory than did their students, the difference was small and the overall pattern was broadly similar.

Therefore, through dialogue, you will need to encourage your beginning teacher to analyse memory and learning processes in their teaching and learning practices, with an awareness that metacognitive errors are common and widespread. The following section looks in more detail at some of the classroom-relevant memory processes that you might choose to focus on.

3. Supporting beginning teachers to optimize learning by embedding the memory phenomenon of desirable difficulties in their classrooms

One group of important and particularly counterintuitive learning phenomena are known as *desirable difficulties*. These are a range of factors which increase the challenge level of a study task – often slowing down performance during practice and causing more errors – but which improve learning over the longer term. Contrary to the typical subjective perception that study tasks which feel easier are desirable (Rhodes & Castel, 2008), interventions which increase the learner's sense of difficulty can actually lead to more effective and durable learning. As Bjork and Bjork (2011) put it, 'Conditions that create challenges and slow the rate of apparent learning often optimize long-term retention and transfer' (p. 57).

My intention here is to present three major desirable difficulties, as well as some mentoring strategies which you can use to promote their use by your beginning teacher.

1. Spacing effect

As suggested by Paul in Case study 5.1, a delay between initial learning and later practice work results in a boost to retention (Rohrer, 2015), a phenomenon known as the *spacing effect.* This can apply to any type of learning, although the benefit may not be apparently immediate – it is best judged at the point of a later test. For example, if a teacher spaced out the teaching of dialysis by reviewing this concept a fortnight after the initial class, pupils may find this more difficult than if the review were to happen straight away, but the payoff is that their eventual test scores would be better. In your coaching sessions, you can:

- Emphasise to your beginning teacher that the spacing effect is counterintuitive; because it implies that an element of forgetting is helpful, in comparison to practising when items are still well remembered by a class.
- Advise them that when spacing out practice, a longer gap is better than a short one (Cepeda et al., 2008), though it is important to emphasise that the information or skills need to be well learned in the first session. As Rawson and Dunlosky (2011) put it, ‘our prescriptive conclusion for students is to practice recalling concepts to an initial criterion of 3 correct recalls and then to relearn them 3 times at widely spaced intervals’ (p. 283).
- Raise their awareness that many popular learning strategies fail to take account of spacing. Most notably, any benefits of *overlearning* – i.e. continuing to practice beyond the point of mastery – seem to be short-lived, and this technique can therefore be seen as an inefficient use of learning time (Rohrer & Taylor, 2006). Such additional practice would be of more benefit if delayed, in keeping with the spacing effect.

You could then ask your beginning teacher to attempt Task 5.3, which asks them to reflect on their teaching and invites you to support them in embedding spacing into their professional practice.

Task 5.3 Supporting beginning teachers to use the spacing effect in the school context

Consider following the steps below:

1. Ask your beginning teacher to identify one area of topic content that pupils often find difficult in science
2. Encourage the beginning teacher to think of two or more practice tasks relating to this content, such as lab demonstrations, workbook exercises, homework, and so forth.
3. Discuss with the beginning teacher how these are commonly scheduled in your school, and think of ways that the practice could be spaced out more. For example, could review/consolidation tasks be delayed by a week or so, and homework on a new concept completed a couple of weeks or a month later?

2. Interleaved practice

The term *interleaving* refers to varying the order of a set of tasks or examples such that each item is immediately followed and preceded by an example of a different category or concept; item types A, B and C would be presented in the order: ABCABCABC. For example, teaching types of plant cells to a class might involve alternating images of one type (e.g. a root hair cell) with images of a different type (e.g. xylem tissue), rather than showing the class several examples or illustrations of the same type of cell at the one time (which is termed a ‘blocked’ arrangement).

Eglington and Kang (2017) found significant benefits of using interleaving when learning about the molecules of different hydrocarbon categories. They found that learners demonstrated better recall of these molecules, and their ability to categorise previously unseen examples was also improved.

You could consider the following mentoring steps to help introduce your beginning teacher to the benefits of interleaving:

- Ask the beginning teacher to think of concepts in the topics they teach which could be presented either blocked by category or interleaved (different categories mixed together).
- Then encourage the beginning teacher to plan interleaved questions or problems for practice tasks. These help pupils to notice and understand the differences between contrasting items.
- Complete Task 5.4 which focuses on the ways you can support your beginning teacher to reflect and incorporate interleaving in their classrooms.

Task 5.4 Supporting beginning teachers to use interleaving

Ask your beginning teacher to review a recent lesson's PowerPoint slides and/or a worksheet prepared for the lesson. This self-review could focus on questions such as:

- Are multiple scientific concepts included in the materials?
- If so, are the concepts presented together (allowing for contrast) or within different sections/tasks?
- Are all of the questions in a worksheet on similar concepts?
- In future, how could the beginning teacher increase the extent to which different types of question are mixed together?

Discuss the answers to the above questions with your beginning teacher and support them to embed interleaving in their planning. For example, you could consider supporting your beginning teacher in planning contrasting and contingent cases to teach *energy* using differential concepts, such as the stringy cheese (Task 4.2) and sticky lolly (Task 4.3) examples.

3. *Retrieval practice*

The technique known as *retrieval practice* involves asking learners to effortfully retrieve things from memory, for example using a short quiz. This is more likely to boost and consolidate learning than the provision of information via a lecture or reading, even though retrieval is more difficult and can lead to more errors than simply listening or copying (Agarwal et al., 2012).

This means, for example, that it would be preferable for learners to listen to a teacher explaining acids and alkalis and then subsequently summarise these concepts into their notes (rather than taking down notes during the talk).

As with other desirable difficulties, the benefits are only obvious after a number of days, not after an immediate test (Roediger & Karpicke, 2006), and therefore performance during the class itself may mislead teachers and learners alike into thinking that retrieval is ineffective, and that easier strategies such as copying or re-reading are preferable. Task 5.5 requires you to invite one of your mentees to reflect on lesson planning and to analyse whether (or not) retrieval practice was adequately included. You could first encourage your beginning teacher to complete Table 5.1, which asks your beginning teacher to reflect on their current understanding.

Table 5.1 A beginning teacher’s views on retrieval practice

Teaching example	*My view*
Pupils are instructed by the teacher to listen to a brief explanation of a new concept, and then to summarise this new concept in their jotters using a textbook to help.	This approach does/does not make use of retrieval practice because... I could increase the level of retrieval required by...
Pupils are allowed to take down notes while the teacher is reviewing a recent sub-topic using verbal explanations and a PowerPoint.	This approach does/does not make use of retrieval practice because... I could increase the level of retrieval required by...

Next, using the answers from Table 5.1, discuss your beginning teacher's understanding of retrieval practice. During this discussion and/or from their recent observed classroom practices, you might realise that although your beginning teacher acknowledges the importance of retrieval practice they exhibit some reluctance to use it in their classrooms, and prefer to plan and implement easier strategies such as asking pupils to copy texts from the board or repeating verbal explanations of scientific concepts. Therefore, Task 5.5 invites your beginning teacher to self-reflect on one of their lesson plans and to analyse whether (or not) retrieval practice was adequately included.

Task 5.5 Supporting beginning teachers to use retrieval practice

Follow the steps below:

1. With your beginning teacher, find one example from one of their recent lesson plans where pupils were asked to take notes, copy from a screen or books, or listen to an explanation.
2. Now ask your beginning teacher to think of an alternative method of completing this section of the lesson, such that the lesson aims would still be met but the pupils would be required to actively retrieve information from memory.
3. Prompt them to consider not just quizzes but also strategies such as summarising, oral questioning or class discussion (all of which require retrieval).
4. Ask them to try out this approach in a similar situation in a future lesson and discuss their evaluation of its effectiveness with you.

All the strategies mentioned above have potential benefits if applied in the classroom, but as highlighted in the previous section, it is very likely that beginning teachers will have misconceptions about them. Therefore, the next section encourages you to take an approach to mentoring which will help them to overcome misconceptions via professional inquiry.

4. Beginning teachers' inquiry into misconceptions about learning and memory

From the points made so far, it may seem to be an obvious conclusion that mentors should correct erroneous misconceptions and beliefs about learning held by beginning science teachers in their schools. However, there are problems with such a simplistic solution. Researchers who study metacognition have increasingly come to recognise that it is difficult to tackle misconceptions about learning and memory directly, as merely being told about techniques such as spacing and interleaving does not necessarily work (Yan et al., 2016). However, a theoretical explanation combined with practical experience may be more beneficial (Yan et al., op cit.). You can see an example of a theoretical explanation along with a practical solution in Case study 5.2.

Case study 5.2 Dorothy and Viktoria

Dorothy is a science head of department who is mentoring Viktoria, a beginning teacher, during a 5-week school placement. Viktoria has frequently expressed the view that pupils need to be told things four times during a single lesson in order to remember them.

Dorothy chooses to take a coaching approach, asking questions such as "How can you be sure that this works?" and "What evidence is there for that idea?"

Dorothy writes down all the conversations with Viktoria. They then access a research study by Soderstrom and Bjork (2015) on the difference between performance and learning, before returning in subsequent weeks to discuss the notes on their earlier conversations on learning and memory.

Through these discussions, Viktoria has recognised that while learners may appear to learn something well by repeating it within a single class, it is possible that they may remember the facts and skills better if repetitions are spaced out over separate weeks and are combined with quizzes.

In order to put this to the test, with the support from Dorothy, she agrees to try this approach with half of her classes, while continuing her current approach with the others.

As can be seen in Dorothy and Viktoria's case, a non-directive approach through probing questions empowered Viktoria to analyse and interrogate her previous assumptions.

While there may sometimes be advantages for you to provide direct instructions to your beginning teacher, this can serve to discourage the self-directed development of beginning teachers (Ehrich et al., 2004).

Case study 5.2 also demonstrates an empowering way to correct misconceptions about learning through engagement in research activities. In such an arrangement, mentors can guide the process of inquiry without mandating particular changes to practice. This moves the mentor-mentee relationship towards a collaborative inquiry, facilitates autonomy on the part of the beginning teacher, and is more effective in promoting professional learning over the long term (Sachs, 2016).

Moreover, such research-led classroom projects in schools can also take the form of shared inquiry carried out by a group of teachers including your beginning teacher. You could discuss with your beginning teacher the potential benefits of such inquiry-based group projects, for instance:

- Doing such work collectively is *less challenging* for a beginning teacher and allows for the pooling of both research skills and findings
- A group of inquiry-focused teachers at any stage of their careers could be *established in a school or cluster of school settings*
- This group may constitute what researchers call a ‘community of practice’: there is a genuine *real-world context, shared goals*, and a framework where beginning teachers can *learn from more experienced peers* without the need for formal instruction (Lave, 1991; see chapters 2 and 13).
- Such projects will allow beginning teachers to *test out evidence-based* techniques in a *collaborative environment*

Having a research mentor to guide a group of research-engaged teachers can provide a useful stimulus. As exemplified in Case studies 5.1 and 5.2, sharing relevant literature can be a good starting point when encouraging a beginning teacher to become research-engaged. The “Teacher ready research review” format found in the *Scholarship of Teaching and Learning in Psychology* is a good example of research papers written to be accessible, while many teacher organisations provide access to research journals (see the further resources section).

As mentioned in previous sections, the mentoring process can highlight flaws in a beginning teacher’s professional knowledge and expertise, and can push a beginning teacher beyond their comfort zone and away from early assumptions that they may have developed, as exemplified in Case study 5.2. Accurate feedback on the teacher’s classroom performance – something that is difficult for them to achieve alone due to the metacognitive issues discussed in the earlier sections – can be obtained from the mentor, the peer group, and indeed from the research data itself, reducing the chance that teachers confuse immediate classroom performance for long-term learning. Such feedback can help them to become a more effective professional.

Finally, Task 5.6 asks you to reflect on your understanding of memory, learning, misconceptions, and inquiry practices in association with your beginning teacher.

Task 5.6 Mentor’s reflection on learning, memory and inquiry

Reflect and review the ideas covered in this chapter by doing the following:

1. Summarise the misconceptions about learning and memory which you are likely to encounter when working with beginning teachers
2. Identify one ‘desirable difficulty’ that you feel would be especially useful for

your beginning teacher to focus on in their planning and classroom work

3. Plan a process of mentored inquiry, either for an individual beginning teacher or for a group of teachers. This mentored inquiry could be focused on one of the desirable difficulties mentioned in the last point.
4. Identify one or more sources of information that your beginning teacher could draw on for background reading about learning and memory.
5. Follow up the classroom inquiry with collaborative reflection on a recent lesson. This could be informed by your beginning teacher's reflective account and pupils' or mentor's feedback, and will help in identifying areas to be further investigated and/or changed.

Summary and key points

This chapter has highlighted ways to support a beginning science teacher in understanding and applying memory processes to promote effective learning among pupils. Key points from the chapter are:

- An understanding of learning is vital for successful teaching
- Pupils' learning of science is based on memory, and the counterintuitive nature of memory – in particular desirable difficulties – means that a beginning teacher is likely to have misconceptions about its functioning
- A mentored research and inquiry approach can motivate a beginning teacher to engage with evidence and provide opportunities to build their professional expertise.

Further resources

Gilchrist, G. (2018) *Practitioner Enquiry: Professional Development with Impact for Teachers, Schools and Systems*, Abingdon, Oxon: David Fulton Books.
This is a highly readable guide to mentoring teachers through the process of conducting their own research and inquiry.

Horvath, F.J., Lodge, J. and Hattie, J. (2016) *From the Laboratory to the Classroom: Translating Science of Learning for Teachers*, Abingdon, Oxon: Routledge.
A comprehensive edited volume with chapters on topics ranging from memory to dyslexia, providing a gateway into the research background on a beginning teacher's educational area(s) of interest.

Smith, M. and Firth, J. (2018) *Psychology in the Classroom: A Teacher's Guide to What Works*, Abingdon, Oxon: Routledge.
An accessible guide to applying psychological concepts in teaching, including memory, understanding, creativity and emotion. This can be used to find out more about desirable difficulties or as recommended reading ahead of mentoring sessions.

The Chartered College for Teaching (CCT), available at:
https://chartered.college
The CCT's easy to follow website focuses on applying learning science in the classroom, with access to their journal *Impact* which has numerous articles on memory, evidence-based professional practice, and ways for beginning teachers to engage with research.

References

Agarwal, P.K., Bain, P.M. and 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), p437–448.

Bjork R.A. (1994) ‘Memory and metamemory considerations in the training of human beings’ in Metcalfe, J. and Shimamura, A. (eds.) *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press, pp. 185–205.

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

Bransford, J.D., Brown, A.L. and Cocking, R.R. (2000) *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy Press.

Carter, A. (2015) *Carter review of initial teacher training (ITT)*. London: Department for Education.

Cepeda, N.J., Vul, E., Rohrer, D., Wixted, J.T. and Pashler, H. (2008) ‘Spacing effects in learning: A temporal ridgeline of optimal retention’, *Psychological Science, 19*(11), p1095–1102.

Eglington, L.G. and Kang, S.H. (2017) ‘Interleaved presentation benefits science category learning’, *Journal of Applied Research in Memory and Cognition, 6*(4), p475–485.

Ehrich, L.C., Hansford, B. and Tennent, L. (2004) ‘Formal mentoring programs in education and other professions: A review of the literature’, *Educational Administration Quarterly, 40*(4), p518–540.

Firth J. (2017) 'Experts in learning' in Rycroft-Smith, L. and Dutaut, J.L. (eds.) *Flip the system UK: A teachers' manifesto*. Abingdon, Oxon: Routledge, *pp. 20–28.*

Firth, J. (2018) *How to learn: Effective study and revision methods for any course*. Glasgow: Arboretum Books.

Gilchrist, G. (2018) *Practitioner enquiry: Professional development with impact for teachers, schools and systems.* Abingdon, Oxon: Routledge.

Horvath, J., Lodge, J. and Hattie, J. (2016) *From the laboratory to the classroom: Translating science of learning for teachers. Abingdon, Oxon:* Routledge.

Kornell, N. and Bjork, R.A. (2008) 'Learning concepts and categories: Is spacing the "enemy of induction"?', *Psychological Science, 19*, p585–592.

Lave J. (1991) 'Situating learning in communities of practice' in Resnick, L.B., Levine, J.M. and Teasley, S.D. (eds.) *Perspectives on socially shared cognition*. Washington, DC: American Psychological Association, pp. 63–82.

Morehead, K., Rhodes, M.G. and DeLozier, S. (2016) 'Instructor and student knowledge of study strategies', *Memory, 24*(2), p257–271.

Rawson, K.A. and Dunlosky, J. (2011) 'Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough?', *Journal of Experimental Psychology: General, 140*(3), p283–302.

Rhodes, M.G. and Castel, A.D. (2008) 'Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions', *Journal of Experimental Psychology: General, 137*, p615–625.

Rohrer, D. (2015) 'Student instruction should be distributed over long time periods', *Educational Psychology Review, 27*, p635–643.

Rohrer, D. and Taylor, K. (2006) ‘The effects of overlearning and distributed practice on the retention of mathematics knowledge’, *Applied Cognitive Psychology, 20*, p1209–1224.

Rohrer, D. and Pashler, H. (2010) ‘Recent research on human learning challenges conventional instructional strategies’, *Educational Researcher, 39*, p406–412.

Sachs, J. (2016) ‘Teacher professionalism: why are we still talking about it?’, *Teachers and Teaching, 22*(4), p413–425.

Simons, D.J. and Chabris, C.F. (2011) ‘What people believe about how memory works: A representative survey of the US population’, *PloS one, 6*(8), p.e22757.

Smith, M. and Firth, J. (2018) *Psychology in the classroom: A teacher's guide to what works.* Abingdon, Oxon: Routledge.

Soderstrom, N.C. and Bjork, R.A. (2015) ‘Learning versus performance: An integrative review’, *Perspectives on Psychological Science, 10*(2), p176–199.

Yan, V.X., Bjork, E.L. and Bjork, R.A. (2016) ‘On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit’, *Journal of Experimental Psychology: General, 145*(7), p918–33.

Appendix 19: Psychology in the classroom (Introductory chapter to a book on evidence-based practice) (Smith & Firth, 2018)

The following text is the introduction chapter of my co-authored education book on evidence-based teaching practice. The book represents a sharing of my own professional learning at this point (2017–18 at the time of writing), and includes chapters on memory, cognition and self-directed learning. I wrote the introduction, which sets the aims out in terms of teacher professional learning. Reference: Smith, M., & Firth, J. (2018). Psychology in the classroom: A teacher's guide to what works. Routledge.

Psychology in the Classroom: A Teacher's Guide to What Works

by M. Smith and J. Firth

Introduction

Psychology is the study of the human mind and human behaviour. Its areas of interest include both the workings on an individual's mind and the way that people behave in pairs and groups. Its theories and research touch on processing abilities such as memory and thinking, as well as more emotional issues such as motivation and curiosity.

In any field of human action and interaction, therefore, the study of psychological principles can potentially be applied to problems that people face. The world of education is just such a field – how people behave and learn in the classroom can be analysed and explained in terms of the psychological processes involved. This book aims to help both new and experienced teachers to understand these processes, helping them to apply them in the classroom in practical ways that will make a positive difference to learners.

As a scientific subject, psychology relies on practical research and tests hypotheses via experiments, trying to establish an objective picture of how mental processes work. Just as in other sciences, this means that there will be situations where common-sense assumptions eventually have to give way to a body of evidence. The drive for research findings to underlie teaching practices – to make education more ‘evidence-based’ – reflects an increasing desire to scrutinise the educational status quo, and to question whether common teaching practices are actually the best way to impart knowledge and understanding, to motivate young people for lifelong learning, and to foster positive learning behaviours.

This means that popular ideas about how people learn and behave – including some older scientific theories – can be scrutinised and compared to current research knowledge. Your own professional knowledge of learning and thinking can be developed and informed by this process. Through reading this and other books on the psychology of education, your planning and thinking can become increasingly evidence-based. This shouldn’t in any way threaten your professional autonomy to make those decisions that you think are in the best interests of your learners. Indeed, an informed professional is in a better position to both make and justify educational decisions, rather than having practices imposed on them.

Some teachers may feel that they are too busy to learn about psychology, and that feeling is understandable. The profession is facing a crisis of workload and conditions while wages stagnate, with thousands of good teachers leaving the profession. However, there are two main counter-considerations to the argument that we are too busy to learn about the evidence:

- *Many of the ideas in this book will make teaching easier and reduce stress*. Learning better means that more gets done in the same amount of time, and is therefore more efficient. Better results cut the stress level for teachers, too. Higher-achieving, more motivated and resilient learners are easier and more pleasant to teach. When learners understand their own learning processes, are highly motivated and work well independently, this again reduces the workload.

- *A deep understanding of classroom processes is important to the profession*. Deep, research-based knowledge about learning is one of the key things that sets a professional teacher apart from a student helper or classroom volunteer, and is what puts the teacher in the position to make high-stakes choices about their learners and about the curriculum. Concurrently, there is personal satisfaction and a confidence boost to be gained from improving one's professional skill level and delivering classes more effectively. An empowered, skilled professional is also in a better position to negotiate working terms and conditions.

Who is this book for?

This book is for teachers in all sectors and at all stages of their career. While a teacher would ideally learn many of the psychological principles discussed in this book towards the start of their career, there is also a great deal to be said for developing a reflective and evidence-based approach as one progresses through the years of a teaching career. This book will give you a primer on many key considerations, and make you aware of a great many theories and concepts, but the best way to understand the issues in depth is to regularly engage with new findings and debates in the psychology of education. The knowledge gained from reading this book will set you up to do so by making it much easier to read and act on relevant new research, and to be sceptical about educational fads.

It should be clear that this book places an emphasis on understanding the theories and evidence behind human behaviour, but we are not for a minute suggesting that classroom experience is unimportant. As one of the first major psychology researchers, William James, wrote:

> *I say moreover that you make a great, a very great, mistake if you think that psychology, being the science of the mind's laws, is something from which you can deduce definite programs and schemes and methods of instruction for immediate schoolroom use. Psychology is a science, and teaching is an art; and sciences never generate arts directly out of themselves. An intermediary inventive mind must make the application, by using its originality.*
>
> *(James, 1899, p. 23)*

In other words, it is important to understand the science of how the mind works, but this is only a start – this knowledge must be applied to a particular educational context by you, the teacher. A solid understanding of the mind is therefore a part of our professionalism, but only a part. Informing ourselves about psychological and educational research can be an empowering force, allowing us to make judgements confidently and in full knowledge of both the facts and the uncertainties highlighted by current psychological research. This book is therefore not going to tell you how to teach your classes, but to provide a grounding in the psychological background that underpins what happens in the classroom. We therefore see a sound knowledge of psychological theory as necessary but not sufficient for the teacher's 'art'.

How to use this book

The book is divided into eight relatively discrete topics, and the chapters can therefore be read in any order. Each can function as a stand-alone guide. It is presented in this way to make it easier to tackle the issues one at a time, perhaps prioritising the areas that are of most relevance to your teaching context. However, they are ordered in what we feel is a logical progression – beginning with the fundamentals of how individuals think and reason, and moving on to areas more concerned with social interaction and long-term outcomes. If you do intend to read all of the sections, it may therefore be best to read them in the order they are presented.

Also, as with any area of psychology, there are considerable overlaps between different topics. For example, there are separate chapters on memory and creativity, but it is acknowledged that these (and other) areas are intermeshed. It would be beneficial for the reader to check back to the other chapters periodically, and to consider how the various topics could affect each other in your teaching context. For example, what are the implications of mindset on independent study? How could the limitations of student cognition play a role in classroom behaviour? Such connections are productive areas for a teacher to reflect upon, and perhaps to research independently.

As well as using the book for your own professional development, it could be used to structure staff development within a department or team. One option would be to use a 'book group' format – setting each chapter in turn as a piece of reading, and following it up with a group discussion. If there is sufficient time, further sessions could be allocated to the books suggested for further reading, or colleagues could read and report back on one or more relevant research articles. The suggested further reading titles, it should be noted, have been chosen because they are accessible and well-researched – they don't necessarily reflect our viewpoints! We have also tried to ensure that the bulk of the choices are primarily education-focused, and that they are up to date.

The structure of each chapter in this volume comprises an introduction, a section that tackles several theoretical concepts and how they apply to the classroom, an explanation of a relevant psychological theory or model and its implications, and finally a discussion of how practical tasks such as lesson planning and materials design can be informed by the psychological ideas presented.

The chapters cover the following issues:

1 *Memory and understanding.* This chapter presents the fundamentals of how people learn and retain information in the memory. It explains the main types of long-term memory, as well as many of the factors that affect whether information and concepts are retained in memory or not.

2 *Cognition*. This chapter looks at how learners take in and use information in the here and now – issues concerning thinking, processing and working memory. It looks at the key role of attention and what psychologists call 'executive function'. Applications to the classroom are founded on Baddeley's working memory model, which sees working memory as dynamic but limited.

3 *Self-theories in teaching and learning*. This chapter describes nonacademic components within the individual that impact learning outcomes. Self-esteem, self-concept and beliefs about intelligence can both help and hinder students, and identifying these can foster a more appropriate view of learning.

4 *Creativity*. This chapter looks at what psychology can tell us about the processes of creative thinking, including association, incubation and divergent thinking, and presents a view of creativity that is based on transferring knowledge and skills to new contexts.

5 *Emotions*. This chapter looks at the often-overlooked influence emotions have on academic outcomes. Emotions, both positive and negative, can impact learning in a number of unexpected ways, and helping students to nurture the most adaptive emotions can improve both well-being and learning outcomes.

6 *Resilience, buoyancy and grit*. Often misunderstood, resilience is a complex construct with many variations. This chapter helps teachers to identify those aspects of resilience that are more useful to academic outcomes and those that can help promote general well-being.

7 *Motivation*. This chapter investigates different types of motivation and why some motivational interventions fail. While the emphasis is on intrinsic motivation, the chapter also discusses the best way to motivate using extrinsic factors.

8 *Independent learning*. This chapter identifies concepts from psychological research that suggest more effective ways to structure independent learning, including revision and homework. It also looks at the debates about active learning and discovery learning, and proposes evidence-based ways of structuring and supporting project work to take account of student metacognition.

Reference

James, W. (1899). Talks to teachers on psychology: And to students on some of life's ideals. New York: Henry Holt.

Appendix 20: Abstract to an article on teacher identity and professional learning

A short article was prepared with colleague Anna Beck and former colleague Philip Tonner based on a research project carried out in my former school, investigating teachers' engagement with research. It will be submitted for publication later this year. The abstract is reproduced here.

I don't have that power': teacher identity, agency and the development of a school-based research centre.

by A. Beck, J. Firth and P. Tonner.

Abstract

This study investigated the development of a unique teacher 'research centre' at a Scottish school, and the impact of this on teacher research engagement. We analyse this development in the context of the changing conceptions of teacher professionalism within Scottish education system, which in turn link to international debates about teachers' research-engagement and their professional agency. In particular we considered the role of social identity theory as a potential mechanism for understanding why some teachers engage with research and others do not. To assess this, we conducted semi-structured interviews with a self-selecting sample of teaching staff. Our thematic analysis includes four themes, each of which ties into this theoretical debate: the conceptualisation of research by members of the teaching profession, barriers to research engagement, the role of teacher agency, and the role of social identity. Each of these areas raised separate issues which build on previous work, but more importantly, we argue that they interact; the conceptualisation of teaching as a research-informed profession, and the affordance of agency to teachers in their professional learning and classroom choices which help them to identify as research engaged.

Keywords: teacher research engagement, practitioner enquiry, social identity, teacher agency, school research centre, professional learning.