Review of Education
An International Journal of Major Studies in Education
BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION
Check for updates

Review of Education
Vol. ●●, No. ●●, ●● 2021, pp. ●●–●●

DOI: 10.1002/rev3.3266

A systematic review of interleaving as a concept learning strategy

Jonathan Firth[1], Ian Rivers[1] and James Boyle[2]
[1]*The University of Strathclyde, UK,* [2]*The University of Strathclyde, UK*

A systematic review was conducted into the effect of interleaving the order of examples of concepts in terms of both memory of items and transfer to new items. This concept has important implications for how and when teachers present examples in the classroom. A total of 26 studies met the inclusion criteria; a subset of 17 studies (with 32 constituent datasets) formed the basis of a meta-analysis, and the remainder were analysed within a narrative review. Memory (as tested by presenting studied items from a learned category) showed an interleaving benefit with effect sizes (Hedges' *g*) of up to 0.65, and transfer (as tested by presenting novel items from a learned category) a benefit with effect sizes of up to 0.66. 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, with implication for practitioner decisions over how and when to apply the technique. It also extended to delayed tests. 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 outlined.

Keywords interleaving, concept learning, memory, transfer, education.

Introduction

Background to the review

In this paper we present the findings of a systematic review into interleaving as a learning strategy. Interleaving means varying the order of a set of 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 can 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 (for example, paintings 1, 2, and 3 by Smith, Jones and Rigg), learners could be provided with examples from each artist in sequence or these could be interleaved, as shown in Figure 1.

Early research in cognitive psychology assumed that viewing examples together in blocks would be beneficial to the process of forming new categories, while spacing

Corresponding author. Jonathan Firth, School of Education, The University of Strathclyde, 141 St James Road, Glasgow, Scotland, G4 0LT. Tel: +44 (0)141 444 8069.
Email: jonathan.firth@strath.ac.uk.

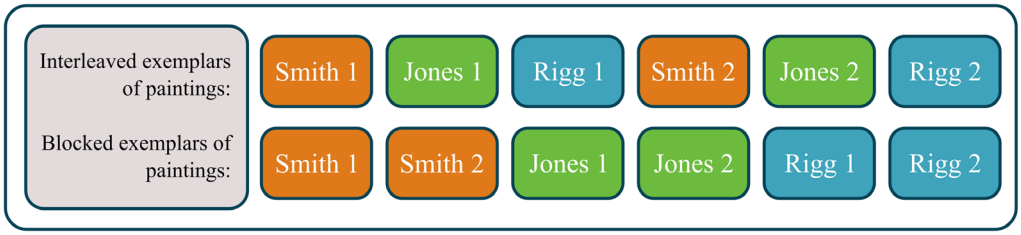

Figure 1. Example of interleaved versus blocked sequences

them out over time—with unrelated examples in-between—would be harmful (Elio and Anderson, 1981). However, when Kornell and Bjork (2008) put this to the test, the findings ran counter to this assumption. In their seminal study, Kornell and Bjork presented learners with artwork in blocked or interleaved formats, in a similar way to that suggested in Figure 1, above. In their study they presented the work of 12 obscure modern artists (6 paintings by each), and then tested participants' ability to identify the artist who had painted novel example paintings. The study had two key findings: first, that interleaving was superior to blocking in this test; and 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, Thomas and Jacoby, 2015; Eglington and Kang, 2017).

Presenting examples to learners facilitates inductive learning—a form of learning where concepts develop gradually through exposure and experience. Each target concept constitutes a category with boundaries, and learners come to be proficient at correctly categorising new examples. 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 mental category (or schema) for each tree species, allowing them to independently categorise 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.

Given that interleaving the order of items appears to improve learning of new concepts, the technique is potentially of great relevance to educators. An optimal sequence of examples during classroom tasks or private study has the potential to speed up learning and to make this learning more transferrable to situations beyond formal education; it also appears to be relevant to multiple curriculum areas. We will now briefly explore the underpinnings of the effect, and then consider its relevance to educators in more detail.

Explanations of the interleaving effect

At first, Kornell and Bjork and other researchers assumed that the benefits of interleaved schedules 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 practising sooner (see Cepeda, Pashler, Vul, Wixted, and Rohrer, 2006, for a review of the effect).

Interleaving examples of a concept with contrasting examples will inevitably lead to a degree of spacing. For example, in Figure 1, each example 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. Taylor and Rohrer, 2010; Kang and Pashler, 2012; Birnbaum *et al.,* 2013) have kept the level of spacing constant in interleaved versus blocked conditions by inserting filler materials such as trivia questions or cartoons.[1] 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 Birnbaum *et al.,* 2013; Zulkiply and Burt, 2013a).

If the benefit of interleaved presentations is not due to spacing, then why does it occur? The attention-attenuation hypothesis (Wahlheim, Dunlosky, and Jacoby, 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 (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). Increased mind wandering tends to be found when a task requires some effort but is undemanding (Baird, Smallwood, Mrazek, Kam, Franklin and Schooler, 2012), and it is plausible that it may be reduced due to a higher subjective sense of difficulty if category exemplars were interleaved.

The discriminative-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. This 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 discriminative contrast than by explanations of interleaving that focus on attention. It also helps to explain why studies that 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 recent studies of interleaving 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).

As discussed by Goldstone (1996), 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 categorise 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 someone 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 such cases, 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, with the key factor being 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. It is compatible with the other two candidate explanations described so far (attention attenuation and discriminative contrast), but has different implications for education—the similarity of examples may be at least as important as the order in which they are presented. As such, the classroom application of interleaving is highly context specific, drawing on teachers' professional knowledge of the material (Firth, 2021)

Applications

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 is 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; Agarwal and Bain, 2019). 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. We aim to provide a review of these issues.

A key concern in the present paper is to ask under what circumstances classroom teachers should interleave examples of new concepts. Guidance on how and when this should be done would be a valuable tool for both teachers and teacher educators, potentially impacting on everyday practice across the curriculum. Without such guidance, practice is likely to be guided more by intuition, but intuition can be flawed when it comes to learning strategies, with teachers preferring easier but less effective options, and avoiding spacing and interleaving (Bjork, 2018; Halamish, 2018).

The kind of classroom practice which might be relevant to interleaving as conceived here might include a teacher giving short, specific examples, either verbal or image-based, when introducing a new science or social science concept, a skill, a genre, or the work of a new artist. It should be noted that there also exists a compelling research literature on the interleaving of maths problems, with some studies having been carried out in classroom settings (e.g. Rohrer *et al.*, 2015, 2020a). It appears that interleaving different mathematics skills stands to provide similar benefits in terms of contrast and attention paid to these contrasts (Rohrer *et al.*, 2015). However, the role of interleaved practice after some initial learning has taken place will not be the focus of the present study, given that it is often unclear whether initial concept learning was blocked or interleaved. In addition, this practice often aims to provide consolidation rather than to foster new learning (e.g. Rohrer and Taylor, 2007). Nevertheless, for teachers whose subjects involve extensive practice of short items, the choice to interleave such items (or not) is again one that can be informed by the research literature.

On the basis of what has been said so far, interleaving is likely to be beneficial when it allows learners to contrast new 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. However, the nature of what constitutes 'item' must be clarified—this could be one example from a concept category, such as one painting by an artist or one tree from a species, but it is less clear whether longer 'items'—such as texts that take many minutes to read—will elicit such strong effects, given that the boundaries and therefore contrasts between items would occur less frequently. It seems unlikely that it would be useful to interleave entire lessons or topics, as such a format would very much reduce the salience and frequency of contrasts.

We would also like to assess the role of the timing of learning in the research studies reviewed, whether interleaving can be gradually introduced, for example, and whether it is affected by a subsequent delay. Decisions over when and in what order to present examples is an important aspect of how interleaving can be implemented in the classroom, and such decisions tend to be under the control of teachers (Firth, 2021). To our knowledge at the outset of this review, 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; 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 independent study. These points will be returned to in the Discussion section.

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 that are previously unseen by the learner, and that they must categorise 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, we will refer to tasks that involve classifying novel items as 'transfer' and tasks that involve classifying older items as 'memory', although we recognise that both processes draw on long-term memory to an extent; transfer involves comparing new stimuli to existing memories (Morris, Bransford and Franks, 1977; Butler, 2010), and can benefit when the encoding conditions of those memories are varied (Gick and Holyoak, 1980). Use of the term transfer for novel items is in line with Carvalho and Goldstone's (2014, 2015a, 2015b, 2017) research in this review, as well as earlier work by J. R. Anderson and others (e.g. Elio and Anderson, 1981).

Transfer, with its inherent requirement to link incoming stimuli to existing categories, tends to be difficult for learners (Salomon and Perkins, 1989), although the level of difficulty depends upon multiple factors such as the time delay involved, or the similarity of the new setting to the prior learning context (see Barnett and Ceci, 2002). According to the transfer-appropriate processing principle, difficulty is reduced if there is greater overlap of the processes during initial learning and later demand. 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 neatly categorised, and variability during learning therefore contributes to transfer-appropriate processing. More broadly, they argue that many features that 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).

Beyond the general issue of memory and transfer, we also require a thorough understanding of how interleaving can be used in educational contexts. In particular, the materials used in the research and how they could generalise to the classroom is of great interest. As discussed, 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, Yan, Bjork and Maddox (2016, p. 24), we 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.

In the time since registering the present review (registered in PROSPERO in 2018), another meta-analysis was published by Brunmair and Richter (2019), reporting an overall pooled effect size for interleaving (versus blocking) of Hedges' $g = 0.42$ ($p = 0.001$; 95% CI [0.34, 0.50]). The present review does not attempt to combine the effect sizes of such diverse materials and activities as art items, word lists and mathematical practice, as was done in the Brunmair and Richter paper, because we intend to focus on concept learning rather than interleaved order per se. This is important because simply changing the order of examples will not always lead to effective classroom practice; as explored earlier, the major theoretical explanations of the interleaving effect imply that for interleaving to be beneficial, meaningful contrast needs to result from this scheduling of items, in such a way that learners' attention is captured, and modification of conceptual knowledge results.

We will also focus particularly on implications for teaching practice; we, like Brunmair and Richter (2019), will carry out a meta-analysis of studies to compare conceptual categories such as artworks, but will also explore what this means for different curriculum areas by comparing the effect with materials which lend themselves either to more inductive learning (e.g. artistic styles) or to more rule-based learning (e.g. science). Where relevant, we will also qualify our meta-analytic findings in terms of factors uncovered in a narrative review. This approach will allow us to focus particularly on the implications of the findings for educational practice (see the Discussion section).

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 and Holyoak, 1980). On the other hand, a higher number of items 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 informed the following research questions:

(i) What factors and boundary conditions impinge on the interleaving effect?. Factors and boundary conditions relating to interleaved learning will be explored in a narrative

synthesis in order to better understand the nature of the effect, such that recommendations for classroom practice can be made. Specifically, these include working memory span, type of stimulus used, similarity of items or categories, timing and schedule of learning tasks, and the availability of explicit rules to guide learning.

(ii) Does set size of categories and exemplars have an impact on performance?. We further predict that the number of categories and examples will affect performance. The number of categories interleaved could potentially be as low as two, and the upper bound may depend on practical factors; initial scoping suggested that the number used in the literature could be as high as twenty (MacKendrick, 2015) or more. Typically, a low number of exemplars from each category are presented to learners as training—six, in the Kornell and Bjork (2008) research—and more are used for testing. However, these numbers warrant further exploration, and this issue will be tackled in our narrative synthesis.

(iii) 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, and will help to confirm the findings of the Brunmair and Richter (2019) meta-analysis where these are relevant to concept learning in the classroom, while seeking to focus on classroom-relevant interleaving of conceptual examples rather than interleaved item order more generally.

(iv) Does interleaving have a larger effect on learning in the context of tasks that require transfer than with tasks that require identification of previously studied examples?. Given that interleaving differs from spacing and appears to be more relevant to categorisation than to rote memorisation, we 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. This question will also be addressed via meta-analysis.

(v) 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, we predict that science-type tasks will show a smaller interleaving benefit. Again, this question will be addressed via meta-analysis.

Method

Search strategy

The search strategy was pre-registered with PROSPERO and a summary has been previously published (see Firth *et al.*, 2019). We carried out database searching using the PsycINFO, Web of Science, BEI, AEI and ERIC databases. Database journal categories were used to exclude records 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 1 (each row represents an 'OR' function). Searches were constrained to records from 2008–June 2018.

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

This initial search was carried out by the lead author, and yielded an initial 683 studies, and these were then subject to reading of abstracts or full text (see Figure 2), in order to apply inclusion and external criteria.

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: participants were aged 13–65 and a typically (or assumed typically) developing sample; an experimental or quasi-experimental design was used; the study collected primary data; one or more independent research variables related to interleaved learning in an educationally relevant context. The chosen age range was based on the intention of the present study to provide an evidence base relevant to the pedagogy used in secondary schools and above. In addition, we recognised that the rapid neuro-psychological change that occurs across learners in early years and primary school settings would make it more difficult to ensure that the findings of any studies into those populations would be comparable to those conducted on older learners.

Studies were excluded if: they were neurological/fMRI-based studies; or outcome variables did not directly relate to concept learning. The main reason for excluding studies from the original search was their focus on issues unrelated to the learning of novel, meaningful concepts, in particular studies of verbal learning (for example, modern language vocabulary), perceptual learning (for example, learning of tones), or motor learning. This was considered parsimonious, as concept learning draws on semantic long-term memory, while excluded domains of learning draw on different long-term memory stores (Schacter, 1990; Squire, 2004) and may therefore respond to timing-based manipulations differently. 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.

Table 1. Database search terms

interleav*	AND	learning
shuffl*		"conceptual knowledge"
"contextual interference"		inducti*
intermix*		

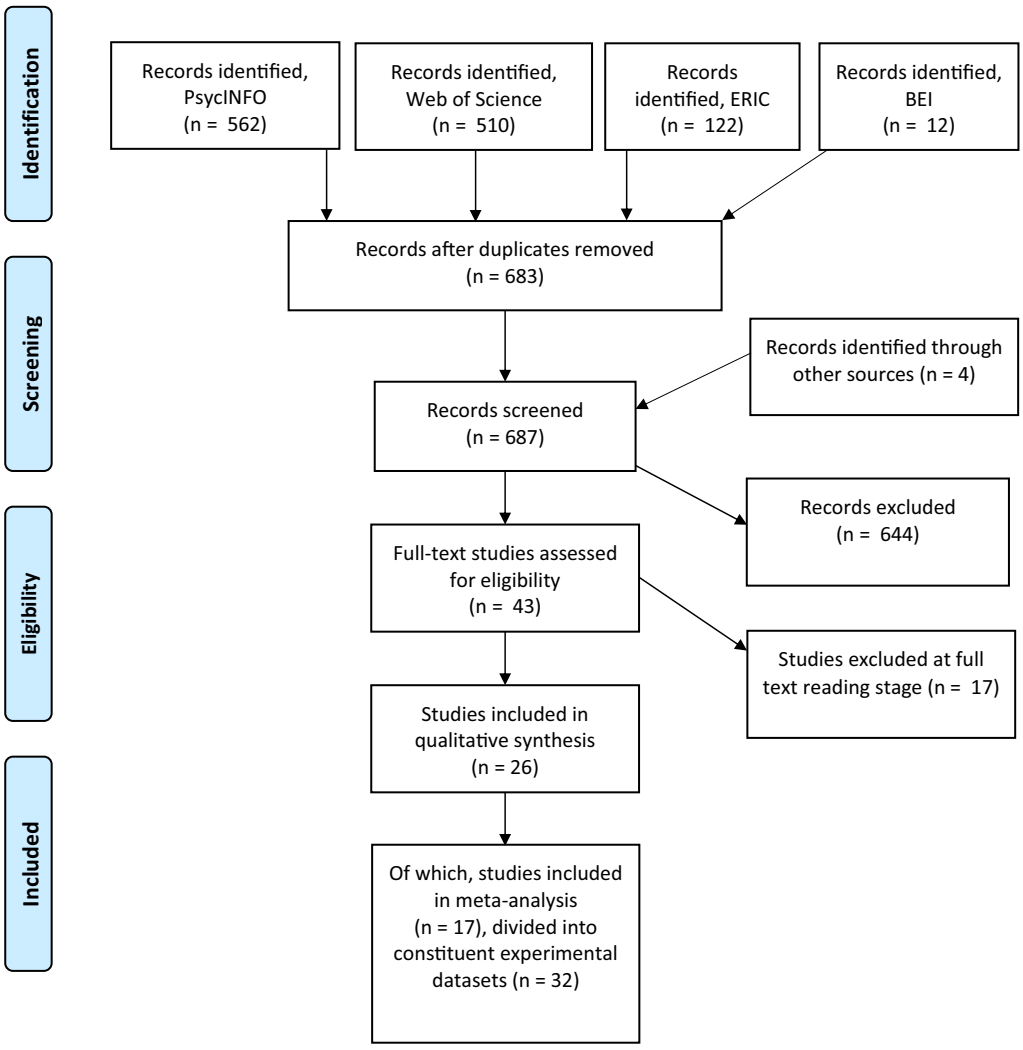

Figure 2. PRISMA diagram summarising search and screening process

We 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. Decisions were made by the lead author based on abstracts, or, in some cases, from full text reading. This search strategy yielded a total of 43 studies suitable for a full text search.

Full text search

The full texts of all of these inclusions were read and discussed within the research team, all of whom are experienced educators with a background in psychology and statistics. The same inclusion and exclusion criteria also applied here, together with

the requirement that the full text was available. 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 reviewed subsequent to initial learning), and we excluded the latter group because they did not compare interleaving and blocking during the initial learning phase.

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 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 1; column 1 details which component experiments were included.

At times the relevance of experimental tasks to classroom learning caused some difficulty which was resolved by discussion among all three co-authors. For example, we 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.

Overall, this phase resulted in 17 further exclusions, with 26 studies being retained.

Review strategy

At this point it was important to identify which of the 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 that had used a very different methodology from the norm (such as Linderholm, Dobson and Yarbrough (2016), who interleaved written passages texts and tested learning via themes in essay writing) were included in the narrative synthesis only. Likewise, studies that 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 that featured a clear interleaved versus blocked comparison as two of their conditions. This element of our 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.

It should be noted that most of the 26 studies comprise more than one experiment; the full list of 56 constituent experiments is given in Appendix 1. Even within the 17 studies that were appropriate for inclusion in the meta-analysis, not every constituent experiment within a particular study could be included, and some needed to be subdivided to separate memory and transfer conditions, yielding 32 experimental datasets overall for the meta-analysis. These are listed in Appendix 2.

Weight of evidence

We also reviewed the extent to which the studies and their constituent experiments provided strong evidence that could be used to answer the review questions. Following Gough (2007), we 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 testing memory and transfer (for example, images), and the mundane realism of the task (for example, categorisation, quizzing) with respect to common classroom activities.

Consistency of rating was established by training and a reliability check on a 20% sample of the included studies by all three co-authors. In this activity, there was complete agreement as to relevance, and agreement over methodology was reached through discussion, with authors coming into line with one another after the first two examples rated. After this process, ratings for the remainder of the sample were completed by the lead author. 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 1, column 4.

Narrative synthesis

As noted already, a list of all experiments included is provided in Appendix 1, and all 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 questions (i) and (ii).

Working memory span

Categorisation draws on existing knowledge of categories/concepts 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 (WM) 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 et al. (2011) 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. However, in terms of practical implications, it can at least be said that student working memory capacity does not appear to be a major barrier to the use of interleaving as a classroom technique.

Type of stimulus

One task in particular dominates the literature on the interleaving of concept learning —categorisation 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.

As can be seen from Table 2, many of the materials used are directly relevant to education, ranging from images to the verbal statistics descriptions used by Sana *et al.* (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.* (2016). 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

Table 2. Stimuli used across all studies

Stimuli	Research study
'Alien' cells (blob shapes)	Carvalho and Goldstone (2014a)
Abstract digital pictures	Zulkiply and 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 and Bath (2012)
Chemistry molecules	Eglington and Kang (2017)
Fribble objects (Williams, 1997)	Carvalho and Goldstone (2015b)
Images of 'alien' creatures	Carvalho and Goldstone (2017)
Medical/physiology passages	Dobson (2011)
Modern art painters' styles	Kornell and Bjork (2008)
Psychology concept definitions	Rawson *et al.* (2015)
Statistics examples	Sana, Yan and Kim (2017)
Ziggerin objects (Wong *et al.*, 2009).	Carvalho and Goldstone (2015b)

studies used lengthy tasks or whole lessons. The classroom implications of these findings are that teachers may be able to use interleaving across many areas of the curriculum, but that the interleaving of lengthy tasks, or entire lesson or topics, cannot (at least at present) be described as evidence based.

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 versus 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 cf. 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 versus 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 (e.g. 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. This may appear to be confusing for the practitioner, with interleaving only sometimes helpful to learners. However, a more general rule—drawing on the points in the Introduction—would be to say that teachers should juxtapose items that cause confusion to learners, in order to draw attention to similarities or contrasts that might otherwise be missed.

Task timing and schedule

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 3 times, and therefore seeing each one 12 times in total. Blocking was superior during four learning phases, but in a test phase that included transfer, interleaving was superior for high-similarity items, and neutral for low-similarity items.

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 that are more in line with educational applications (see Table 3).

Out of the studies in Table 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 the Introduction), and suggests that its advantages can persist across educationally relevant timescales.

In authentic educational settings, of course, study is not a one-off experience with novel items, but typically builds on previous learning and is followed up by further practice. 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, this review did not find much evidence to either support or repudiate the idea of combining interleaving and blocking, and nor was it much studied.

As noted earlier, it is often within the power of practitioners to determine how and when they present examples. While recommendations regarding task schedule must

Table 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

remain tentative given the extent of the available evidence, this part of the review raises three main implications: that the benefit of interleaving may be cumulative over several learning sessions, that it will tend to endure over at least a few days, and that while a blocked-to-interleaved schedule is not strongly supported by the research, it is preferable to blocking alone.

Rule-based (non-inductive) learning

Categorisation of novel objects based on a deductive verbal rule may rely on explicit cognitive processes which differ from the implicit learning referred to in the introduction (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?

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

However, 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; but this study also found that there was an advantage when learning was top down, though it was reduced.

The work of Eglington and Kang (2017) is also relevant to rule-based learning; their work involved interleaving of images of molecules—categories where there is an objective rule governing category membership, and where specific component features can allow learners to make accurate categorisation. Nevertheless, there was again evidence that interleaving can be beneficial with this material.

Further relevant examples, beyond the scope of the present review, might include the interleaved learning of language skills with evidence, for example, that grammar rules could benefit from interleaved practice (see the Discussion section).

A key factor may be the salience and ease of using a rule. When an explicit rule is readily available, learners may be more inclined to focus on applying it than on the examples themselves. This idea was supported by the work of Rawson *et al.* (2015, Experiment 1b), who 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. Another factor may be the working memory demands involved in applying a rule to a specific example (Ashby and Gott, 1988; Noh *et al.,* 2016).

Overall, then, interleaving appears to be especially useful for subtle pattern learning where there is no clear verbal rule, perhaps because this prompts learners to pay attention to the examples, and specifically to the salient contrasts. Nevertheless, for the teacher, these findings suggest that interleaving will often be equivalent to or superior to blocking in more clear-cut domains too, particularly in situations where sets of examples are presented as a starter task, or are encountered incidentally followed by a rule-based explanation.

Categories and exemplars

There was a level of variation in terms of how many exemplars and how many categories were used in the experiments selected. The variation in number of categories used is summarised in Table 4; as can be seen, the modal number of categories was 12 (due in large part to the many replications of the Kornell and 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 and Bjork).

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.

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

Some studies varied category number to increase difficulty. For example, Eglington and Kang (2017) raised this from four to eight between Experiments 3 and 4; however, in the 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 and 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 being the other principal 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. These numbers are shown in Appendix 1. However, it is important to

Table 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

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, and also suggest that when it comes to classroom application, the benefits of interleaving are not confined to large sets of contrasting examples.

Meta-analysis

Eligible studies

Datasets suitable for inclusion in the meta-analysis came from 17 out of the 26 studies overall. 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 our inclusion criteria) only data from the younger participants were analysed here. Some experiments were also separated into memory and transfer conditions (see Transfer versus memory).

Table 5 summarises the materials from those datasets used in the meta-analysis only (see also Appendix 2). It is notable that with the exception of the modern art images, all of the materials could easily be used within a typical high school science classroom. For simplicity of terminology, we decided to categorise modern art as 'art', and the remaining types of material as 'science'. This difference was founded on ecological validity (i.e. would these things be found in a real art classroom or a real science classroom, and be used by teachers of these subjects?) rather than on an analysis of features of the materials themselves.

Transfer versus memory

As noted earlier, 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, we felt that it was necessary to separate out the transfer and memory conditions of some experiments. Splitting the experiments in this way resulted in a total of 32 datasets for further statistical analysis (transfer, within-participants design: 13 datasets; memory, within-participants design: 3 datasets; transfer, between-participants: 13 datasets; memory, between-participants: 3 datasets). These are shown in Appendix 2. As can be seen, some of the rows in Appendix 2 relate to the same study, divided into two parts, memory and transfer, each with effect sizes (Hedges' g) calculated. All of the effect sizes were in a positive direction (interleaving superior to blocking). These effect sizes were calculated using the Comprehensive Meta-Analysis software package (Borenstein, Hedges, Higgins and Rothstein, 2014), with mean scores and standard deviations requested directly from the researchers where these were absent from the published articles.

However, as the participants in these parts of the experiment were the same (none of the included studies compared memory versus transfer between groups), the

Table 5. Stimuli used from datasets (n = 32) included in the meta-analysis

Materials used	Number of experiments
Modern art	14
Animal images	3
Chemistry/math images	6
Verbal—biology/psychology	6
Verbal—statistics	3

findings lacked independence, and we therefore compared them separately. Similarly, we compared within-groups and between-groups designs separately, resulting in four main analyses. These are reported in the next section.

Statistical findings: meta-analysis

Research questions (iii) to (v) were addressed using meta-analysis. Random effects models were used throughout (see Borenstein, Hedges, Higgins, and Rothstein, 2009).

Within-participants designs, transfer. We first ran an analysis on the 13 within-participants datasets that had tested transfer. Using a random model in the Comprehensive Meta-Analysis software we calculated a pooled effect size (Hedges' g) of interleaving at + 0.59, $p < 0.001$, 95% CI [0.45, 0.72]. Egger's test (Egger, Smith, Schneider, and Minder, 1997) was used to test for publication bias for this group of datasets, 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 0.142).

We noticed high levels of heterogeneity among this group: a Q value (12 df, p = 0.00) of 45.4; I-squared = 73.6; Tau squared = 0.04. Analysis revealed that Metcalfe and Xu (2016) (effect size + 1.08), Zulkiply (2013) (effect size 1.00), Guzman-Munoz (2017, experiment 2) (effect size + 0.84) and Kornell and Bjork (2008) (effect size + 0.80) were outliers when viewed on a funnel plot (see Figure 3). Their removal yielded improved heterogeneity (Q-value = 8.64 [8 df; p = 0.37]; I-squared = 7.38; Tau squared = 0.001). The effect size from the remaining nine datasets (Hedges' g) was + 0.43, $p < 0.001$, 95% CI [0.35, 0.52].

We also wanted to know if the learning materials used contributed to the variability found among this group of studies. We therefore conducted a further analysis on the original 13 datasets in this category, subdividing them by materials (art or science). In terms of the datasets from studies that had used art images, the pooled Hedges' g = 0.64 ($p < 0.001$, 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 datasets (n = 5) also showed one outlier: Zulkiply (2013). A sensitivity analysis with the outlier removed resulted in the following findings: g = 0.38 ($p < 0.001$, 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 our fifth prediction, namely that art-based materials will have a larger interleaving effect.

Within-participants designs, memory. We next ran an analysis on the three datasets from within-participants designs that had tested memory, and calculated a pooled effect size Hedges' $g = 0.65$ ($p < 0.001$, 95% CI [0.50, 0.80]). Egger's test was again used to test for publication bias, 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 0.516). This time, there was a low level of heterogeneity across the datasets ($Q = 1.02$), and we did not detect any outliers. However, we decided that this group was too small to make a meaningful comparison of different types of stimuli/materials.

Between-participants designs, transfer. We next analysed the datasets from studies of transfer, this time focusing on the 13 with between-participants designs. Again using a random model, we calculated a pooled effect size Hedges' $g = 0.66$ ($p < 0.001$, 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, we found $Q = 18.4$. There was one outlier apparent, MacKendrick (2015) (effect size + 1.65). Removing this dataset and re-analysing the data led to a pooled effect size of $g = 0.61$, and to a 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 datasets by stimuli led to pooled effect sizes of Hedges' $g = 0.56$ ($p < 0.001$, 95% CI [0.41, 0.71]) for the experiments that had used science-

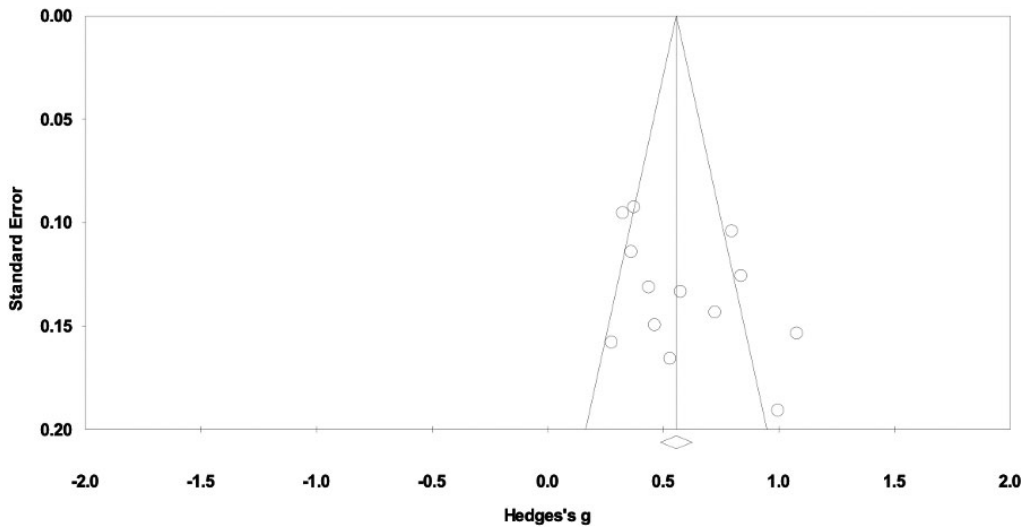

Figure 3. Funnel plot showing four outliers (within-participants; transfer)

based materials ($n = 8$), and $g = 0.82$ ($p < 0.001$, 95% CI [0.46, 1.17]) for the experiments that had used art-based materials ($n = 4$), providing further support for the prediction that there would be a larger interleaving effect with art-based materials.

Between-participants designs, memory. Finally, we analysed the three datasets relating to memory from experiments with between-participants designs. Their pooled effect size Hedges' $g = 0.39$ ($p = 0.011$, 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 0.72). Again, there was a low level of heterogeneity ($Q = 2.74$). However, as with the other set of memory findings, this group was too small to make a meaningful comparison of different types of stimuli/materials.

Sensitivity analysis. We were aware that there were multiple datasets from the same researchers or labs among the data, presenting a possible source of bias. Therefore, as a sensitivity analysis, we further analysed each set of 13 datasets focusing on transfer as follows (each of the following effect sizes is again based on Hedges' g).

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

For between-groups designs, we compared the work of Kang and colleagues ($n = 6$) versus the others ($n = 7$), and found effect sizes of $g = 0.63$ and $g = 0.73$, respectively; in the latter case the effect size was $g = 0.62$ 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 groups of datasets when analysed without the outlier. Finally, for Sana and colleagues ($n = 3$) the effect size was $g = 0.53$ compared to $g = 0.66$ 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 both groups of datasets, with comparable effect sizes across different labs.

Discussion

Principal findings

In the previous sections we outlined the review protocol and described how 26 studies of interleaving were identified. The full set of studies were taken into account in the narrative synthesis, providing an exploration of variables that interact with or limit the interleaving effect. Seventeen of the studies were then investigated via a meta-

analysis. Their most homogenous constituent experiments, with some subdivided in line with their focus on either memory or transfer, resulted in 32 datasets for analysis.

With respect to research questions (iii)–(v), the findings have revealed that presenting examples of concepts 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 materials relevant to science education, and therefore shows potential for broad educational application. The methodology of the studies reviewed was 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.

Our 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 and using different materials—strengthen the idea that it is the interleaved order of examples, 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 and Pashler, 2010; Yan *et al.*, 2017). In short, interleaving is a desirable difficulty (Bjork and 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. 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, Sandhofer and Kornell (2008) found that interleaving was advantageous for preschool children, suggesting that the benefits found in this review can potentially generalise to learners at any educational stage.

Of course, prior knowledge can affect how stimuli are categorised, and young children are likely to have less prior knowledge, meaning that they may on occasion miss subtle conceptual differences that an older learner would focus on. It is also long established within developmental psychology that children often categorise stimuli in ways that can differ from those of adults; pre-schoolers, for example, often categorise moving objects such as clouds as being animate (Piaget, 1930). Further research into the way that the effect emerges developmentally and interacts with the prior knowledge and assumptions of younger groups of learners, including those with special educational needs, is desirable.

The reviewed studies have used 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 we note that the statistical power of this test is low in cases where there are fewer than 10 studies (Higgins and Thomas, 2019), which here applies to the studies of previously learned examples (memory rather than

transfer). However, the reviewed studies primarily feature relatively artificial and short-term tasks, and most use visual stimuli. Future studies should investigate school learners in real classroom contexts, ideally using materials that are drawn directly from their school courses and that build on prior learning.

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 —an issue that relates to a question of ecological validity of the research reviewed, given that most included large sets of practice examples and lacked a broader learning context. The likelihood is, of course, that teachers do not present 50+ examples at the start of a lesson, and instead integrate a smaller number of examples into other activities over time. However, as revealed by some of the studies reviewed (Kang and Pashler, 2012; Birnbaum *et al.*, 2013), a gradual approach to presenting examples can be beneficial as long as contrast is maintained, allowing spacing and interleaving to work in concert (Birnbaum, 2013). It may reasonably be speculated—subject to confirmation via field research—that it would be helpful when teaching a new concept to spontaneously raise examples of prior concepts—especially ones that could easily be confused with the new material. Such a juxtaposition would make the contrasts more salient than having them appear only in separate lessons. Rohrer *et al.* (2020a) can provide a useful methodological model here; in their study, high school students were provided with blocked or interleaved practice tasks over the course of five months of the school year.

We predicted that interleaving would have a greater impact on transfer-based tasks than memory-based tasks, 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 molecules 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.

Furthermore, we predicted that the set size and number of categories would affect interleaving, 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, and to understand the circumstances under which a greater number of specific examples will be needed. One finding that may be immediately useful to teachers is that smaller categories sizes can be effective, with some studies reviewed employing as few as two categories. This may be helpful where contrasting three or more categories is impractical due to time constraints, or goes beyond the requirements of the syllabus, or for sub-topics where only two meaningfully related concepts/categories exist.

Theoretical implications

On a theoretical level, a number of the studies provide support for the discriminative-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 understandings.

It is interesting to consider why, as found by Eglington and Kang (2017), visually highlighting key differences between items did not modulate the effect. These findings support the idea that category learning processes happen on an automatic level, with learners' 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.

The current study was more focused in its approach than the meta-analysis conducted by Brunmair and Richter, and this is worth addressing. Those authors arrived at similar conclusions regarding art items and the importance of item similarity as we did; 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. Brunmair and Richter (2019) found a benefit of *blocking* for word-only lists (interleaving Hedges' $g = -0.39$), a finding which could suggest that such materials draw on internally diverse categories (see the Introduction). On the other hand, there is good reason to suppose that interleaving of word lists raises different issues to the learning of meaningful concepts, and the negative effect size for word lists/foreign language vocabulary found by Brunmair and Richter may support this contention. We would argue that a single word cannot necessarily be considered to be an example of a meaningful category in the same way as the materials which were the focus of the current review (for example, styles of artwork). However, we would argue that the effect sizes which we found for science items ($g = 0.38$–0.54) suggests that interleaving these items tends to elicit similar cognitive processes as does the interleaving of art materials: inductive, and meaning-based. This supports the idea presented here that initial learning of *concepts* recruits different cognitive processes from other interleaved tasks, and deserves a specific focus. The present findings therefore apply to science and social science activities, but educators should be wary of generalising them to such activities as spelling practice or the learning of vocabulary.

Indeed, it is important to emphasise to teachers that interleaving is not just about the order of items, and cannot be mechanically applied to all types of items with the assumption that it will help. In combination with the points made about similarity, there is an emerging picture that interleaving is beneficial particularly to those tasks and areas where conceptual confusion is likely—for example, mixing up the work of two artists, or confusing an alkyne for an alcohol. In this, we concur with the view of Kang and Pashler (2012), who say in their study of paintings: 'We think that our

findings have implications not just for the learning of art styles but also for the learning of complex categories found in the real world' (p. 102).

Practical implications

This point implies a professional decision over what examples to contrast in the classroom—a decision that may best be made by a teacher who is familiar with their course content and has observed mistakes and misunderstandings among their previous students. However, it is also worth highlighting that such confusion depends on the learner. The level of prior knowledge affects whether a difference is clear-cut or not, and these distinctions cannot meaningfully be made out of context. Item similarity, for example, is not merely a property of the items themselves but rather an interaction between the items and a group of learners. Distinctions that are obvious to experts may be easily missed by beginners (Chi *et al.,* 1981).

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-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; among older learners there could be obvious links with medical or dental education.

When using verbal materials, the length of examples is potentially quite variable, and we found some support for the idea that verbal examples should be brief rather than extended, and for the related idea that they should be juxtaposed without an intervening delay. These aspects of the findings fit the discriminative-contrast hypothesis (Wahlheim *et al.,* 2011; Kang and Pashler, 2012; Birnbaum *et al.,* 2013), and suggest that there is unlikely to be a major benefit if teachers were to interleave entire activities, lessons or topics—practices which were absent from the literature reviewed. An important implication is the need for teachers to understand the importance of comparing brief examples. This is not always made clear to educators. For example, the Sutton Trust report entitled 'What makes great teaching' (Coe *et al.,* 2014) recommends 'interleaving with other tasks or topics' (p. 17), perhaps implying that longer tasks should be interleaved rather than focusing on the contrast between specific examples.

The positive findings for science-based materials in the current review in addition to the evidence on rule-based learning led to a cautious conclusion that such domains can benefit from interleaving provided that the focus on a rule or definition does not distract from the examples given. Several studies found at least some benefit of interleaving over blocking in such domains, and outside of the scope of this review, Pan *et al.* (2019) found a benefit from interleaving grammar principles in a foreign language. In terms of classroom practice, there is clearly a greater efficiency of telling learners information directly where possible, so that they could (for example) classify chemical molecules on the basis of facts about their structure rather than through inductive learning of multiple images. However, as explored in a classic study by Schwartz and Bransford (1998), the 'time for telling' (p. 475) of direct rules is more helpful when it follows a period in which learners are prepared for the factual

information by analysing contrasting cases. This proposal is consistent with the research reviewed here, though fuller investigation in field settings is desirable. Perhaps most importantly, there appears to be no strong reason to *avoid* the interleaving of examples in science and other rule-based domains provided that learners are induced to focus on and interact with the examples given, though again the focus should be on easily-confused concepts in particular.

Limitations

The present review has a number of limitations worth highlighting. First, 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 that have been overlooked due to the terminology used (for example, synonyms that we did not search for), or because they were not included in the databases searched. The research body is also still quite limited. As new studies are published, future reviews could address some of the issues from the section 'Narrative Synthesis' in more depth; in particular, we feel that the relationship between working memory capacity and interleaving would be an appropriate matter for a future meta-analysis.

Secondly, the time period is a limitation. At the outset, we 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.

Thirdly, there was a degree of subjectivity in applying the inclusion/exclusion criteria. As noted earlier (see the Method section), there could be some doubt about what constitutes relevance to science or social science learning. Similarly, the division of materials into art and science categories is somewhat imprecise. Other researchers may interpret this differently, and the views of practising teachers as judges could be used in future work. What's more, the criteria led to the exclusion of useful classroom-based studies of interleaving (such as Rau *et al.,* 2013) and a focus on lab-based studies. Future research could instead choose just to review studies which took place in high-school classrooms, but the inclusion criteria would need to be broader.

Finally, as discussed earlier, 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, as discussed in the Introduction. Interleaving has been described as a useful practice and consolidation technique in other work (e.g. Rohrer *et al.,* 2015, 2020a), and Brunmair and Richter (2019) found a pooled effect size (Hedges' g) of 0.34 for maths items, with many of the constituent studies having been based on practice tasks. The decision to exclude this compelling body of work was in part to ensure that all studies reviewed here had a comparable focus on initial concept learning, and in part because guides to evidence-based teaching practice that include interleaving (e.g. Agarwal and Bain, 2019) tend to focus more on practice tasks than on initial learning, even if the strategy does remain underused by teachers and under-represented in teaching materials (Rohrer *et al.*, 2020b). The classroom application of the technique to initial learning, in contrast, receives less attention. However, other educators may view this choice as overly cautious, and we would accept that a fuller picture of the research will only emerge if it is clear how

and when interleaving should be used both for learning new concepts and for the review and consolidation of those concepts.

Conclusion

In conclusion, the findings of this systematic review show that interleaving, when applied appropriately, can be a powerful tool for learning new concepts. It can boost both memory and transfer, and applies across different subject domains. Its benefits stem from helping learners to contrast similar items that would otherwise be easily confused, and as such it should be applied where teachers believe common misconceptions to lie. It can be applied even to science areas where objective rule-based principles exist. However, the benefits found were restricted to the presentation of visual examples or very short texts, and should not, on the basis of the available evidence, be generalised to a recommendation to interleave lengthy educational tasks.

Acknowledgements

The authors wish to acknowledge the assistance of several researchers in responding to our queries and requests for data: Jennifer Burt, Paulo Carvalho, Francisco Guzman-Munoz, Sean Kang, Nate Kornell, Janet Metcalfe, Katherine Rawson.

Conflict of interest

The authors declare no conflicts of interest with regard to this study.

Data availability statement

The data sets analysed during the current study are available from the corresponding author on reasonable request.

NOTES

[1] Note that in experiments on spacing, the comparison condition is usually referred to as 'massing', implying that items are close together in time. This can be differentiated from 'blocking', which refers to sets of items in a consecutive, non-interleaved order, although many learning experiences will be both blocked and massed.
[2] Birnbaum (2013) Expts 6a and 6b did not meet inclusion criteria in that they did not test interleaving as an IV.
[3] Rawson *et al.*, Expts 1a, 1b and 3 did not meet inclusion criteria in that they did not test interleaving as an IV.
[4] Yan *et al.*, Expts 3 and 4 did not meet inclusion criteria in that they did not test learning/retention as a DV.

References

Agarwal, P.K. & Bain, P.M. (2019) *Powerful teaching: Unleash the science of learning* (San Francisco, CA, Jossey-Bass).

Ashby, F.G. & Gott, R.E. (1988) Decision rules in the perception and categorisation of multidimensional stimuli, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 33–53. https://doi.org/10.1037/0278-7393.14.1.33.

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.

Benassi, V.A., Overson, C.E. & Hakala, C.M. (Eds., 2014). Applying science of learning in education: Infusing psychological science into the curriculum. Retrieved 4 March 2019 from http://teachpsych.org/ebooks/asle2014/index.php

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.

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

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.

Borenstein, M.H., Hedges, L.V., Higgins, J.T. & Rothstein, H.R. (2009) *Introduction to meta-analysis*. Chichester, UK, John Wiley & Sons Ltd.

Borenstein, M.H., Hedges, L.V., Higgins, J.T. & Rothstein, H.R. (2014) *Comprehensive meta analysis version 3* (Englewood, NJ, Biostat).

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

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.

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 generalisation, *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.

Chi, M.T.H., Feltovich, P.J. & Glaser, R. (1981) Categorization and representation of physics problems by experts and novices, *Cognitive Science*, 5, 121–152. https://doi.org/10.1207/s15516709cog0502_2.

Coe, R., Aloisi, C., Higgins, S. & Major, L. E. (2014) What makes great teaching? Review of the underpinning research. https://www.suttontrust.com/wp-content/uploads/2014/10/What-Makes-Great-Teaching-REPORT.pdf.

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.

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.

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.

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 generalisations 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.

Firth J. (2021) Boosting learning by changing the order and timing of classroom tasks: implications for professional practice. *Journal of Education for Teaching*, 47(1), 32–46. http://dx.doi.org/10.1080/02607476.2020.1829965.

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

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.

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.

Halamish, V. (2018) Pre-service and in-service teachers' metacognitive knowledge of learning strategies, *Frontiers in Psychology*, 9, 2152.

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.

Higgins, J.P.T. & J. Thomas (Eds) (2019) *Cochrane handbook for systematic reviews of interventions (V6)*. The Cochrane Collaboration. Available from https://training.cochrane.org/handbook.

Kang, S.H. (2016) The benefits of interleaved practice for learning, in: J.C. Horvath, J.M. Lodge & J. Hattie (Eds) *From the laboratory to the classroom: Translating the science of learning for teachers* (London, UK, Routledge), 79–93.

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.

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., 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.

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.

MacKendrick, A. (2015) *Interleaved effects in inductive category learning: The role of memory retention*. Unpublished doctoral dissertation, University of South Florida.

Metcalfe, J. & Xu, J. (2016) People mind wander more during massed than spaced inductive learning, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 978–984. https://doi.apa.org/doiLanding?doi=10.1037%2Fxlm0000216.

Morris, C.D., Bransford, J.D. & Franks, J.J. (1977) Levels of processing versus transfer appropriate processing, *Journal of Verbal Learning & Verbal Behavior*, 16(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9.

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.

Pan, S.C., Lovelett, J.T., Phun, V. & Rickard, T.C. (2019) The synergistic benefits of systematic and random interleaving for second language grammar learning, *Journal of Applied Research in Memory and Cognition*, 8(4), 450–462. https://doi.org/10.1016/j.jarmac.2019.07.004.

Piaget, J. (1930) *The child's conception of causality*. Kegan Paul.

Rau, M.A., Aleven, V. & Rummel, N. (2013) Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave?, *Learning and Instruction*, 23, 98–114. https://doi.org/10.1016/j.learninstruc.2012.07.003.

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.

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., Dedrick, R.F. & Stershic, S. (2015) Interleaved practice improves mathematics learning, *Journal of Educational Psychology*, 107, 900–908. https://doi.org/10.1037/edu0000001.

Rohrer, D., Dedrick, R.F. & Hartwig, M.K. (2020) The scarcity of interleaved practice in mathematics textbooks, *Educational Psychology Review*, 32, 873–883.

Rohrer, D., Dedrick, R.F., Hartwig, M.K. & Cheung, C.N. (2020) A randomized controlled trial of interleaved mathematics practice, *Journal of Educational Psychology*, 112(1), 40–52. https://doi.org/10.1037/edu0000367.

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. (2007) The shuffling of mathematics problems improves learning, *Instructional Science*, 35(6), 481–498. https://doi.org/10.1007/s11251-007-9015-8.

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.

Schacter, D.L. (1990) Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate, *Annals of the New York Academy of Sciences*, 608, 543–571. https://doi.org/10.1111/j.1749-6632.1990.tb48909.x.

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.

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.

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.

Sweller, J., Ayres, P. & Kalyuga, S. (2011) *Cognitive load theory* (New York, NY, Springer).

Taylor, K. & Rohrer, D. (2010) The effects of interleaved practice, *Applied Cognitive Psychology*, 24, 837–848. https://doi.org/10.1002/acp.1598.

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). New York, NY: Lawrence Erlbaum.

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.

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.

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

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.

Appendix 1. (Continued)

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 5[2]	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 did not interrupt contrast. Large spacing was superior to small spacing.
Carvalho and 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 and 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.

Appendix 1. (Continued)

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
Carvalho and Goldstone (2014a) Expt 3	High	High	Medium	High	'Blob' figures	3 (8)	Compares simultaneous and successive presentation for low-similarity categories only.
Carvalho and 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 did not interact with delay.
Carvalho and Goldstone (2014b) Expt 1b	High	High	Medium	High	'Blob' figures	3 (8)	Removed 'refresher' study session and found very low recall rates.
Carvalho and Goldstone (2015b) Expt 1	High	High	Medium	High	'Fribble' objects	3 (4)	Interleaving beneficial for active learning but blocking superior for passive.
Carvalho and Goldstone (2015b) Expt 2	High	High	Medium	High	'Ziggerins'	3 (4)	Replicated findings of Expt 1 with different materials.
Carvalho and 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.

Appendix 1. (Continued)

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
Carvalho and 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 and 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 and Kang (2017) Expt 1	High	High	High	High	Science images (chemical molecules)	5 (12)	Interleaving benefit extends to STEM categories (chemistry molecules).
Eglington and 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 and 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 and 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.

Appendix 1. (Continued)

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
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 and Pashler (2012) 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 and Pashler (2012) 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 and 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 and Bjork (2008) Expt 1b	High	High	Medium	High	Art images (paintings)	12 (6)	Replication of 1a with between-participants design.
Kornell and 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.

Appendix 1. (Continued)

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
Kornell *et al.* (2010)	High	High	Medium	High	Art images (paintings)	12 (6)	Used same methodology as Kornell and 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 and Xu (2016)	High	High	Medium	High	Art images (paintings)	12 (12, 15 or 18)	Novel selection of paintings; also looked at mind wandering via a probe, with more mind wandering reported in the blocked condition.
Noh *et al.* (2016) Expt 1	High	High	Medium	High	Math 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	Math images (patterns of lines and shapes)	4 (16)	Added another dimension of complexity. The findings followed the same pattern as Experiment 1.

Appendix 1. (Continued)

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
Rawson *et al.* (2015) Expt 2^3	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 and Bouwmeester (2014)	High	High	Medium	High	Art images (paintings)	12 (6)	Replicated findings of Kornell and 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.

Appendix 1. (Continued)

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
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 2^4	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 and Bjork (2008). Compared rule-based learning (via prior factual information about each artist) with inductive. Interleaving superior over both conditions.
Zulkiply and Burt (2013a) Expt 1	High	High	Medium	High	Art images (paintings)	12 (6)	Same images as Kornell and Bjork (2008). Insert 30 s delay, and found that interleaving was better immediate than temporally spaced, though the difference disappeared by the 4^{th} test block.

Appendix 1. (Continued)

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
Zulkiply and Burt (2013a) Expt 2	High	High	Medium	High	Abstract digital images	12 (6)	Designed high-discriminability (HD) vs. low-discriminability (LD) materials. Interleaving best with LD, but effect reversed with HD.
Zulkiply and Burt (2013b) Expt 1	High	High	Medium	High	Art images (paintings)	12 (6)	Same images as Kornell and Bjork (2008). Interleaving advantage persists over a long-term retention condition (7 days).
Zulkiply and 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 2

Table of 32 datasets included in meta-analysis, drawn from 17 studies.

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

Appendix 2. (Continued)

No.	Study/experiment	Design (B = between, W = within)	Transfer/ memory?	Materials – category	Sample	WoE D	Effect size
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 (experiment 1)	B	Transfer	Art (paintings)	120 undergraduates	Medium	1.65
19	Metcalfe & Xu	W	Transfer	Art (paintings)	66 undergraduates	High	1.08
20	Noh, Yan, Bjork & Maddox (experiment 1)	B	Memory	Science (math images)	132 adults (MTurk)	High	0.35
21	Noh, Yan, Bjork & Maddox (experiment 2)	B	Transfer	Science (math 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

Appendix 2. (Continued)

No.	Study/experiment	Design (B = between, W = within)	Transfer/ memory?	Materials – category	Sample	WoE D	Effect size
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