Journal of Applied Research in Memory and Cognition 3 (2014) 153–160

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Mixing topics while studying does not enhance learning

Hannah Hausman, Nate Kornell *

Hannah Hausman and Nate Kornell, Department of Psychology, Williams College, USA

ARTICLE INFO

Article history:
Received 11 December 2013
Received in revised form 20 March 2014
Accepted 21 March 2014
Available online 29 March 2014

Keywords:
Learning
Memory
Metacognition
Spacing
Interleaving
Mixing

ABSTRACT

According to a recent survey, it is common for students to study two topics at the same time using flashcards, and students who do so virtually always keep the topics separate instead of mixing flashcards together (Wissman, Rawson, & Pyc, 2012). We predicted that mixing might be a relatively easy way to increase learning efficiency because mixing increases the spacing between repetitions of a given item, and spacing enhances long-term learning. We compared two conditions: in the mixed condition, participants alternated on each trial between studying anatomy terms and Indonesian translations. In the unmixed condition they studied one topic and then the other. Items were interleaved within item-type in both conditions. Mixing did not have reliable effects when participants studied flashcards in a single day (Experiments 1 and 2) or on two different days (Experiments 3 and 4). Thus, the results seem to disconfirm two sets of beliefs: students' universal belief that mixing flashcards is undesirable and cognitive psychologists' belief that doing so should be encouraged.

Students constantly make decisions about how, when, and how much to study. These decisions can have a meaningful effect on learning (for a review, see Bjork, Dunlosky, & Kornell, 2012). Choosing to use flashcards is one common decision. In a recent survey of undergraduates, 68% of students reported using flashcards to study (Wissman et al., 2012), a number consistent with previous surveys (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2008b). Given that there are over 10 million college students in the United States alone, it is evident that millions of students use flashcards. This fact alone makes it seem important to investigate whether students are getting the most from their flashcards, especially if students have mistaken beliefs about how best to use them.

One decision that can have a major impact on learning is whether students choose to mass or space items within and between study sessions. Numerous studies have demonstrated large positive effects of spacing, with many different materials, lag times between presentations of a given item, types of tests, and delays before the final test (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1988, 1996). It is effective to space learning events such that they occur in different study sessions (between session spacing) and to mix items together rather than studying one item repeatedly within a given session (within session spacing, also known as interleaving). Kornell (2009) demonstrated that learning benefits from both between session and within session spacing.

One way students could increase the spacing between flashcards would be to mix together flashcards from two different topics or courses. According to a recent survey, 59% of students at Kent State University said they had encountered a situation in which they were using flashcards to study for more than one course at the same time (Wissman et al., 2012). Suppose, for example, students were studying biology and history flashcards. Students could choose to mass their study—i.e. study all of their biology and then all of their history flashcards—or they could mix topics, alternating studying one biology and one history flashcard. Wissman et al. (2012) found, however, that 98% percent of students said they would study flashcards from one subject at a time, rather than mixing them, and of those 98%, 68% said they would not mix topics because it would be confusing. Cognitive psychologists, on the other hand, have considered that mixing topics could enhance learning. Roediger and Pyc (2012) suggested that students could easily capitalize on the positive effects of spacing and interleaving when they study by mixing topics within a particular subject, such as different concepts from biology. Roediger and Pyc then asked, "Might it be even more beneficial to intermix study on entirely different topics, such as biology and history?" but noted, "The evidence on this matter is not yet at hand" (p. 244).

* Corresponding author at: Department of Psychology, Williams College, Williamstown, MA 01267, USA. Tel.: +1 413 597 4486; fax: +1 413 597 2085.
E-mail address: nkornell@gmail.com (N. Kornell).

http://dx.doi.org/10.1016/j.jarmac.2014.03.003

1. The present experiments

The present experiments were inspired by a practical question: should students mix topics together while studying. Previous research on interleaving and spacing has not directly addressed this question. As far as we know, the research presented here is the first to manipulate whether two different topics are studied separately or mixed together.

In Experiments 1 and 2 we explored the effect of mixing topics in a single study session on test performance 48 h (Experiment 1) or one week (Experiment 2) later. Participants studied Indonesian translations and anatomical definitions. Each definition was studied multiple times but as with real flashcards, individual items were not restudied consecutively. In the unmixed condition, participants studied all word pairs from one subject before switching topics (as if participants had two sets of flashcards). In the mixed condition, study trials alternated between Indonesian and anatomy word pairs (as if they mixed two sets of flashcards into one larger set). In Experiment 3, there were two study sessions that were separated by 48 h. Participants in the unmixed condition studied one topic on each day. In the mixed condition, both topics were studied during each session. Finally, Experiment 4 replicated Experiment 3 and we introduced an unmixed + spaced condition in which participants studied both topics on both days, but did not mix flashcards from the two topics. This final experiment allowed us to compare the relative benefits of mixing topics within sessions and spacing study trials across sessions.

2. Theoretical considerations

In addition to their practical importance, these studies have theoretical implications because they contrast the benefits of spacing and interleaving. The primary difference between interleaving and spacing is the activity that occurs in between repetitions of a given item: With interleaving, repetitions of an item are separated by other similar items; with spacing, they are separated by unrelated activities. The mixed and unmixed conditions both involve interleaving, because, for example, in between repetitions of a specific Indonesian pair there are always other Indonesian pairs. The difference between the conditions is a difference in spacing: the unmixed condition involves pure interleaving, whereas the mixed condition involves interleaving with additional spacing as well. The spacing comes from the unrelated trials that occur between repetitions of a pair (e.g., anatomy items, which are unrelated to Indonesian, create spacing between Indonesian trials). In the experiments reported here, if students study 16 anatomy and 16 Indonesian flashcards, mixing topics increases the number of items that intervene before participants restudy any given definition (31 versus 15 intervening flashcards). Therefore, the comparison of the mixed and unmixed conditions is actually a comparison of larger versus smaller amounts of spacing (which is sometimes known as lag). Previous research has demonstrated the benefits of increased spacing using word pairs and lags similar those of our mixed and unmixed conditions (Karpicke & Bauernschmidt, 2011; Pyc & Rawson, 2009, 2012). Thus, based on the increased spacing, we predicted a benefit of mixing topics.

At first glance, recent research might seem to suggest reasons why mixing could also have negative effects. The last few years have seen a considerable amount of research demonstrating benefits of interleaving in category learning (Birnbaum, Kornell, Bjork & Bjork, 2013; Kang & Pashler, 2012; Kornell & Bjork, 2008a; Wahlheim, Dunlosky, & Jacoby, 2011) and math learning (Mayfield & Chase, 2002; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010; or for reviews see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rohrer, 2012). Some of this research points to a specific benefit of seeing related materials appear in consecutive trials. These results have been explained by the discriminative contrast hypothesis, which states that juxtaposing exemplars from similar concepts or categories helps people learn by highlighting the differences that distinguish among the concepts or categories (Birnbaum et al., 2013; Kang & Pashler, 2012; Wahlheim et al., 2011). For example, Kang and Pashler (2012) had participants study 12 paintings by three different artists. In the interleaved condition, paintings by all of the artists were mixed together. In the temporal spaced condition, participants studied paintings blocked by artist. The amount of spacing was equated in the two conditions by using unrelated filler material in the temporal spaced condition between presentations of paintings by the same artist. On a final test, participants more accurately classified novel paintings by the three artists in the interleaved condition than the temporal spaced condition, even though spacing was held constant. Interleaving helped participants notice stylistic differences that separated the work of one artist from another.

Mixing, in the present research, interrupts the juxtaposition of items (e.g., anatomy) by interposing unrelated items (e.g., Indonesian). Thus, one might predict that mixing could have a negative effect on learning of related flashcards. This prediction rests on the assumption that discriminative contrast applies when learning word pairs, however, when in fact there are important and relevant differences between learning word pairs and learning categories. In induction tasks, such as classifying paintings by similar artists, participants have to abstract general classification rules and learn to tell the difference between two categories. Discriminative contrast is crucial because the main challenge of the test is telling one category apart from the other (especially because many of the categories were very similar). When learning word pairs, telling the stimuli apart is trivial—the cue is a direct and unambiguous signal of which item the participant is meant to retrieve. Thus, discriminative contrast does not seem relevant when participants learn word pair associations.

If discriminative contrast does not affect learning word pairs, we would expect a positive effect of mixing topics, because of increased spacing, without any negative effect to balance it out. If discriminative contrast does affect vocabulary learning, however, we would expect the benefit of mixing to diminish or disappear.

3. Experiment 1

3.1. Method

3.1.1. Participants

Fifty-five participants (31 female, 24 male; median age = 26 years, range = 18–70 years) were recruited online using Amazon's Mechanical Turk and were paid $1.00 for completing the fist session and another $1.00 for completing the second session. All participants reported being fluent English speakers living in the United States, except for one who did not provide a country of residence. There were 27 participants in the unmixed condition and 28 in the mixed condition.

In addition to the 55 participants whose data were analyzed from Experiment 1, five more participants completed the experiment but were excluded. One of these participants was excluded for having a short median response time on the final test of 0.48 s; the next shortest median response time was 1.50 s. Another participant was excluded for not being a fluent English speaker. The remaining three participants were excluded for answering yes to a question that asked about previous knowledge of any of the word pairs being tested in the experiment.

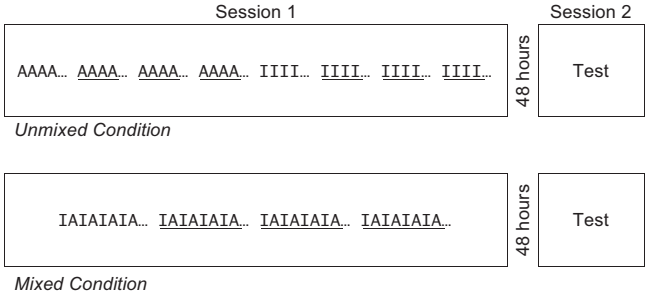

Fig. 1. In Experiments 1 and 2, participants studied 16 Indonesian translations (each represented by an I) and 16 anatomical definitions (each represented by an A) four times each during the study phase. The first study trial was a presentation (not underlined). The next three study trials were tests with feedback (underlined). In the unmixed condition (top), participants studied one topic at a time. In the mixed condition (bottom), participants studied alternating word pairs from two topics. A cued recall test of all 32 word pairs followed in a second session, that occurred 48 h (Experiment 1) or one week (Experiment 2) after session 1.

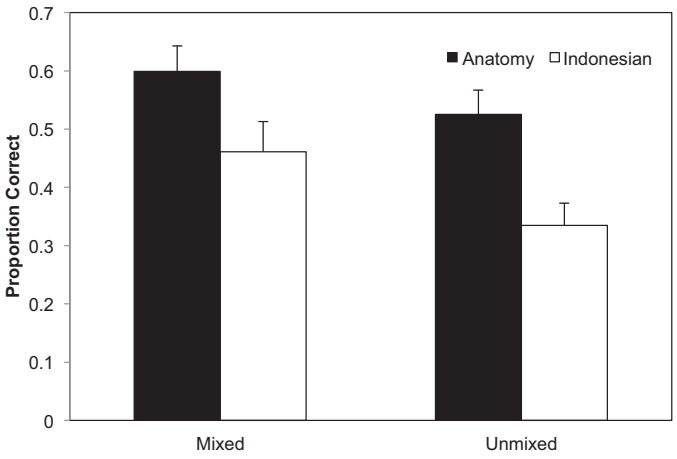

Fig. 2. Proportion correct on the final test in Experiment 1. Error bars represent one standard error of the mean.

3.1.2. Materials

The materials were 16 Indonesian terms and their English translations (e.g., sabun–soap) and 16 anatomical terms (e.g., of the arm–brachial). The full list of pairs appears in the Appendix.

3.1.3. Design and procedure

As Fig. 1 shows, there were two between-participant conditions: mixed and unmixed. All participants studied anatomy and Indonesian and were randomly assigned to study one of these topics first. In the mixed condition, topic order simply amounted to whether the first word pair participants studied was anatomy or Indonesian. In the unmixed condition, topic order determined which topic they would study in its entirety first. Within each topic, the order the word pairs appeared in was randomized between participants.

A study phase in session 1 was followed by a cued recall test 48 h later in session 2 (see Fig. 1). In the study phase there were two kinds of trials. During study trials, the cue and target were presented together (e.g. sabun–soap). During test trials, participants were shown the cue (e.g. sabun) and asked to type the target (e.g. soap); after responding they were then given feedback with the cue and target presented together. In both types of trials, timing was under the participant's control.

In all experiments, the final cued recall test was blocked by topic, with the participants being tested first on the topic they studied first. Again, participants were given as much time as needed to answer. They were also given feedback on the final test. All of the experiments took place online.

3.2. Results and discussion

As Fig. 2 shows, participants recalled more items on the final test in the mixed condition than the unmixed condition, but a 2 (mixed vs. unmixed) × 2 (Indonesian vs. anatomy) mixed-design analysis of variance indicated that the difference was not significant, $F(1,53) = 3.03$, $p = .09$, $\eta_p^2 = .05$. Recall of anatomical definitions was significantly higher than recall of Indonesian translations, $F(1,53) = 43.82$, $p < .001$, $\eta_p^2 = .45$. The interaction between word type and study condition was not significant, $F(1,53) = 1.12$, $p = .29$, $\eta_p^2 = .02$.

Studying topics separately led to numerically higher accuracy than mixing topics during the three test trials that occurred during the study phase (see Table 1). However, a planned comparison showed the decrease in performance that occurred between the last test of the study phase and the final test 48 h later was greater in the unmixed than mixed condition, $t(53) = 3.83$, $p < .001$, $d = 1.05$. It is tempting to interpret this difference as suggesting that mixing was an effective way to prevent forgetting. There is a simpler explanation, though: during the study phase, the task was easier in the unmixed condition than in the mixed condition (because there was less time for forgetting), but on the final test, the task was equivalent. Thus, performance in the study phase probably provides an inflated measure of actual knowledge. For this reason, we refrain from drawing strong conclusions based on study phase performance in this and subsequent studies.

In summary, contrary to our prediction, mixing topics did not significantly enhance learning.

Table 1
Proportion correct on test trials during the study phase for Experiments 1–4.

	Test 1	Test 2	Test 3
Experiment 1			
Mixed			
Anatomy	0.45 (0.26)	0.59 (0.26)	0.68 (0.26)
Indonesian	0.32 (0.24)	0.52 (0.3)	0.63 (0.32)
Unmixed			
Anatomy	0.53 (0.2)	0.64 (0.21)	0.73 (0.21)
Indonesian	0.33 (0.21)	0.51 (0.24)	0.66 (0.24)
Experiment 2			
Mixed			
Anatomy	0.41 (0.21)	0.51 (0.24)	0.61 (0.25)
Indonesian	0.27 (0.22)	0.43 (0.26)	0.55 (0.31)
Unmixed			
Anatomy	0.66 (0.24)	0.77 (0.23)	0.86 (0.15)
Indonesian	0.51 (0.28)	0.66 (0.26)	0.79 (0.24)
Experiment 3			
Mixed			
Anatomy	0.42 (0.25)	0.37 (0.26)	0.51 (0.26)
Indonesian	0.26 (0.25)	0.22 (0.25)	0.39 (0.31)
Unmixed			
Anatomy	0.50 (0.24)	0.65 (0.25)	0.78 (0.24)
Indonesian	0.34 (0.31)	0.50 (0.27)	0.63 (0.25)
Experiment 4			
Mixed	0.27 (0.23)	0.26 (0.27)	0.42 (0.31)
Unmixed + Spaced	0.44 (0.25)	0.24 (0.21)	0.50 (0.28)
Unmixed + Massed	0.35 (0.23)	0.52 (0.27)	0.62 (0.26)

Note: To control test delay across the mixed and massed conditions, analysis of final test performance data was restricted to the topic participants studied second in Experiments 3 and 4. Therefore, only study-phase recall for that topic is included in this table for Experiments 3 and 4. In Experiment 3, some participants studied anatomy second and some studied Indonesian second. In Experiment 4, all participants studied Indonesian second. Standard deviations are given in parentheses.

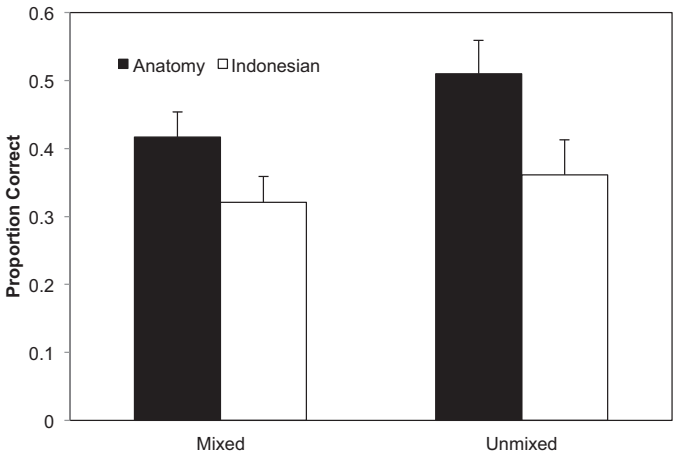

Fig. 3. Proportion correct on the final test in Experiment 2. Error bars represent one standard error of the mean.

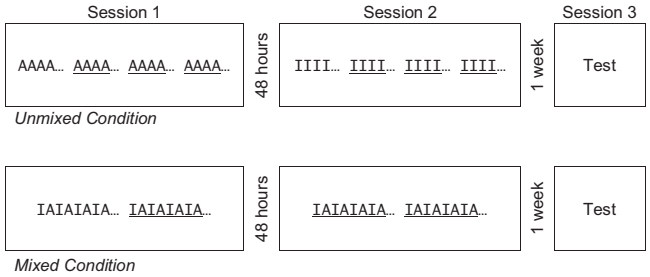

Fig. 4. In Experiment 3, participants studied the 32 word pairs four times each, once in a study trial (not underlined) and three times in test trials with feedback (underlined). The study trials spanned two sessions that occurred two days apart. In the unmixed condition, participants studied every anatomical definition four times in one study session and every Indonesian translation four times in the other study session. In the mixed condition, participants studied every word pair twice in the first session and twice in the second session. A cued recall test of all 32 word-pairs was given in the third session, one week after the completion of session 2.

4. Experiment 2

There was a small, non-significant benefit of mixing topics in Experiment 1. Experiment 2 tested the hypothesis that this benefit would be larger if the delay between study and test was increased from two days to one week.

4.1. Method

4.1.1. Participants

Seventy-nine participants (35 female, 43 male, 1 did not report; median age = 28 years, range = 18–62 years) were recruited online using Amazon's Mechanical Turk and were paid $1.00 for completing the fist session and another $1.00 for completing the second session. All participants were fluent English speakers living in the United States, except for one participant who did not report a country of residence. An additional 21 participants completed the experiment but were excluded for answering yes to a question that asked about previous knowledge of any of the word pairs being tested in the experiment.

There were 44 participants in the mixed condition and 35 in the unmixed condition, with the differing numbers of participants due to random assignment.

4.1.2. Materials, design and procedure

Experiment 2 was nearly identical to Experiment 1. The only difference was the delay between the study phase and the final test, which increased from 48 h in Experiment 1 to one week.

4.2. Results and discussion

As Fig. 3 shows, unmixed study led to higher accuracy on the final test than mixed study, but a 2 (mixed vs. unmixed) × 2 (Indonesian vs. anatomy) mixed-design analysis of variance revealed that this difference was not significant, $F(1,77) = 1.33$, $p = .25$, $\eta_p^2 = .02$. Again, anatomy definitions were again recalled more accurately than Indonesian translations, $F(1,77) = 31.30$, $p < .001$, $\eta_p^2 = .29$ and there was no significant interaction between word type and study condition, $F(1,77) = 1.55$, $p = .22$, $\eta_p^2 = .02$.

As in Experiment 1, unmixed study led to better recall throughout the study phase (see Table 1). A planned comparison revealed that there was significantly more forgetting from the end of the study phase to the final test in the unmixed study condition than the mixed condition, $t(77) = 3.99$, $p = .002$, $d = 0.91$.

To summarize, the non-significant benefit of mixed study in Experiment 1 did not grow larger in Experiment 2. Instead, contrary to our predictions, unmixed study resulted in higher recall than mixed study, though the difference was not significant. Taken together, Experiments 1 and 2 provide no support for the hypothesis that mixing flashcards from different topics together would help promote learning, despite the fact that the average spacing between repetitions of a given item was 31 other items during the mixed condition and only 15 other items in the unmixed condition.

5. Experiment 3

In Experiments 1 and 2, participants studied all of the word pairs from both topics four times on a single day. In reality though, students frequently study one subject one day and another subject on a different day. If students mix flashcards from both topics though, using the same amount of time to study each day, repetitions of a given flashcard naturally occur during both study sessions. Therefore, it is possible that mixing topics enhances learning in a way that Experiments 1 and 2 did not capture: It might cause students to study a given flashcard on multiple days (even if mixing topics does not benefit recall *per se*). Experiment 3 tested this hypothesis.

Participants in the mixed condition alternated Indonesian and anatomy word pairs, studying every word pair on day one and again on day two. Participants in the unmixed condition studied Indonesian one day and anatomy another day, massing all of their studying on a given item into one day (see Fig. 4). We hypothesized that recall would be higher in the mixed condition because of the demonstrated benefits of between-session spacing (including with respect to flashcards; Kornell, 2009).

5.1. Method

5.1.1. Participants

Seventy-seven participants (40 female, 37 male; median age = 31 years, range = 18–59 years) were recruited online using Amazon's Mechanical Turk and were paid $1.00 for each of the first and second sessions and $2.00 for completing the third session. All participants were fluent English speakers, except for one who did not provide information on English fluency; all participants were living in the United States. There were 42 participants in the mixed condition and 35 in the unmixed condition, with differences arising from random assignment.

An additional 15 participants completed the experiment but were excluded for answering yes to a question that asked about previous knowledge of any of the word pairs being tested in the experiment.

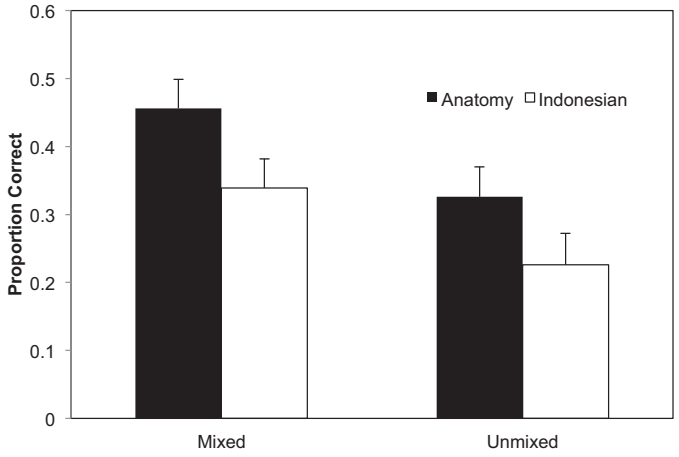

Fig. 5. Proportion correct on the final test in Experiment 3. To control test delay across the mixed and unmixed conditions, only data from the second topic is included in the figure. Error bars represent one standard error of the mean.

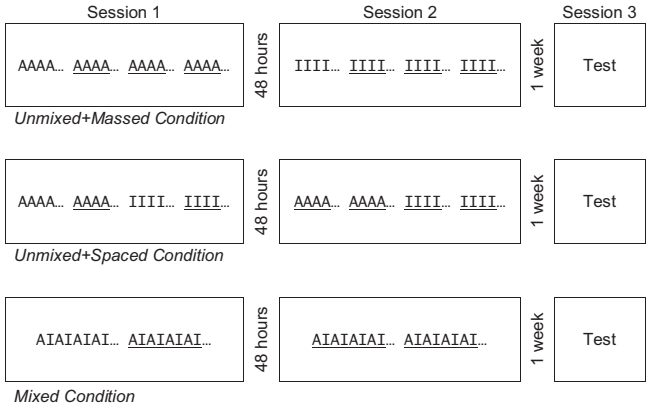

Fig. 6. In Experiment 4, participants studied the 32 word pairs four times each, once in a study trial (not underlined) and three times in test trials with feedback (underlined). The study trials occurred in two sessions separated by two days. In the unmixed + massed condition, participants studied only one topic during each session. In the unmixed + spaced condition, participants studied both topics in both sessions, but studied all anatomy items consecutively before switching topics. In the mixed condition, participants alternated between studying individual anatomy and Indonesian word pairs during both sessions. A cued recall test of all 32 word-pairs was given in the third session, one week after the completion of session 2.

5.1.2. Materials, design and procedure

Experiment 3 differed from the previous experiments by having two study sessions that occurred two days apart. The procedure is summarized in Fig. 4. Participants in both the mixed topics and unmixed conditions studied each of the 32 word pairs a total of four times, in one study trial followed by three test trials. In the unmixed condition, participants studied one topic (Indonesian or anatomy) in the first study session and the other topic in the second study session. Participants in the mixed topics condition studied all 32 items twice in session one and twice in session two. A final cued recall test occurred one week after the second study phase.

5.2. Results

An initial 2 (mixed vs. unmixed) × 2 (Indonesian vs. anatomy) mixed-design analysis of variance showed that recall was in fact significantly higher for participants in the mixed condition ($M = .40$, SD = .24) than the unmixed condition ($M = .28$, SD = .22), $F(1,75) = 5.27$, $p = .02$, $\eta_p^2 = .07$, supporting our hypothesis. As in the previous experiments, recall was higher for anatomical definitions ($M = .40$, SD = .28) than Indonesian translations ($M = .29$, SD = .28), $F(1,75) = 11.13$, $p = .001$, $\eta_p^2 = .13$. The interaction was not significant, $F < 1$.

The initial analysis is flawed, however, because the retention interval between the last study trial and the test on a given item differed across conditions. For participants in the unmixed condition, the retention interval was nine days for the topic studied first and seven days for the topic studied second. The retention interval was seven days for both topics in the interleaved condition (see Fig. 4). To equalize retention interval, we did a follow-up analysis examining only items from the second topic. Therefore, this follow-up analysis led to a comparison of recall across four distinct groups of participants: mixed + anatomy second ($N = 21$), mixed + Indonesian second ($N = 21$), umixed + anatomy second ($N = 21$), and umixed + Indonesian second ($N = 14$). (Since only one topic is analyzed per participant, word type becomes a between-participants variable.)

The results of this analysis are displayed in Fig. 5. A 2 (mixed vs. unmixed) × 2 (Indonesian vs. anatomy) factorial analysis of variance indicated that, unlike in the first analysis, the effect of study method was not significant, $F(1,73) = 0.87$, $p = .35$, $\eta_p^2 = .01$, though mixed study led to slightly higher average recall than did unmixed study. The effect of word type remained significant, $F(1,73) = 5.61$, $p = .02$, $\eta_p^2 = .07$ and the interaction of study method and word type remained non-significant, $F < 1$. In short, when retention interval was equalized, the effect of mixing topics was not significant.

Similar to Experiments 1 and 2, unmixed study led to higher recall on the last test of the study phase in comparison to mixed study (see Table 1). The amount of forgetting from the end of the study phase to the final test was significantly greater in the unmixed study condition than in the mixed study condition, $t(75) = 7.10$, $p < .001$, $d = 1.64$.

5.3. Discussion

Experiment 3 showed no benefit of mixing topics even though items that were mixed were repeated on two different days instead of being studied on just one day. This finding seems to go against a wealth of research on the spacing effect (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), but it should be interpreted with caution because similar conditions produced somewhat different results in Experiment 4.

Nevertheless, we speculated about why mixing (and spacing) did not enhance learning. It seemed possible that two manipulations, mixing and spacing, were having opposite effects that counteracted each other. If introducing between session spacing actually enhanced learning, as the prior literature would predict, then perhaps mixing impaired learning, and the two effects canceled each other out.

We could not test this conjecture in Experiment 3 because spacing and mixing were confounded: Items were not mixed within sessions or spaced across sessions in the unmixed condition, and they were both mixed and spaced in the mixed condition. In Experiment 4 we de-confounded these variables to better understand the joint effects of mixing and spacing.

6. Experiment 4

Experiment 4 had two conditions that were nearly identical to the mixed and unmixed conditions of Experiment 3. For clarity, we named the condition that was equivalent to the unmixed condition of Experiment 3 the unmixed + massed condition in Experiment 4. A third condition was included in Experiment 4—the unmixed + spaced condition—in which participants studied individual word pairs spaced across two sessions, but the

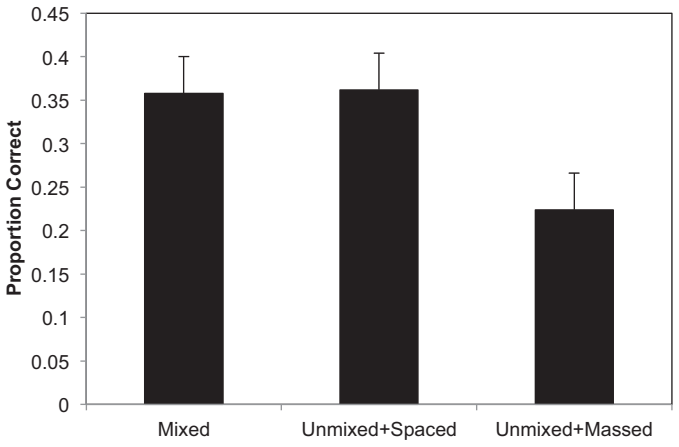

Fig. 7. Proportion correct on the final test in Experiment 4. To control test delay across the three conditions, only data from the second topic (Indonesian translations) is included in the figure. Error bars represent one standard error of the mean.

two topics were not mixed (see Fig. 6). The introduction of the unmixed + spaced condition allowed us to separate the effects of mixing topics and spacing study trials across multiple days. The effect of spacing items across sessions, without mixing topics, can be gleaned by comparing the unmixed + massed and unmixed + spaced conditions. The effect of mixing, when all pairs are studied in multiple sessions, is apparent in the comparison of the unmixed + spaced and mixed conditions.

6.1. Method

6.1.1. Participants

One hundred and thirty-three participants (75 female, 58 male; median age = 34 years, range = 18–67 years) were recruited online using Amazon's Mechanical Turk and were paid $1.00 for each of the first and second sessions and $2.00 for completing the third session. All participants were fluent English speakers, except for one who did not provide information on English fluency; all participants were living in the United States. The number of participants in the mixed, spaced, and unmixed conditions were 46, 46, and 41, respectively. These sample sizes varied slightly because of random assignment.

No participants answered yes to a question that asked about previous knowledge of the Indonesian translations being tested in the experiment (as discussed below, we only analyzed performance on Indonesian translations).

6.1.2. Materials, design and procedure

Experiment 4 was nearly identical to Experiment 3 with two exceptions. First, because lag to test confounds required us to only analyze the second topic participants studied, we decided to have all participants study the same topic second to make comparisons more equivalent across participants. Thus, all participants studied anatomy first and Indonesian second and we only analyzed performance on Indonesian pairs. Second, we added an unmixed + spaced condition in which participants studied all anatomy word pairs consecutively before studying Indonesian word pairs, but every word pair from both topics was studied on both days (see Fig. 6).

6.2. Results

To equalize retention interval across the three conditions, we restricted our analyses to recall of the last set of items studied, the Indonesian translations (see Fig. 7). A one-way ANOVA revealed that the proportion of translations correctly recalled on the final test was significantly affected by study condition, $F(2,130) = 3.76$, $p = .03$, $\eta_p^2 = .05$. Performance was similar for the mixed and unmixed + spaced conditions ($M = .36$, SD = .29 and $M = .36$, SD = .25, respectively) and lower in the unmixed + massed condition ($M = .22$, SD = .25). A Tukey–Kramer post hoc test showed that recall was significantly higher in the unmixed + spaced than the unmixed + massed condition. The difference between the mixed and unmixed + massed conditions was not significant. (As Fig. 7 shows, however, the magnitude of the non-significant difference between the mixed and unmixed + massed conditions was almost identical to the magnitude of the significant difference between the unmixed + spaced and unmixed + massed conditions.)

On the test at the end of the study phase, the pattern of means was inverted (see Table 1). Overall, study condition significantly affected the amount of forgetting between the end of the study phase and the final test one week later, $F(2,130) = 45.51$, $p < .001$, $\eta_p^2 = .41$. A Tukey–Kramer post hoc test revealed that the amount of forgetting was significantly larger in the unmixed + massed condition than the unmixed + spaced and mixed conditions, which did not differ significantly.

6.3. Discussion

The main question in the current research was whether mixing two topics together enhances learning. Experiment 4 compared two spaced conditions, one where topics were mixed and one where they were not. As Fig. 7 shows, recall rates were almost identical in these two conditions on the final test. Thus, again, mixing topics seemed to have no effect on learning, positive or negative. Consistent with prior research on the spacing effect though, the difference in performance between the unmixed + massed and unmixed + spaced conditions was significant.

Finally, we compared recall in the mixed and unmixed + massed conditions, which used the same procedure as the mixed and unmixed conditions of Experiment 3, respectively. Experiments 3 and 4 gave slightly different results, however, which is difficult to explain. The difference in recall on the final test between the mixed and unmixed conditions in Experiment 3 was 5%, which was not significant. The difference was also not significant in Experiment 4, although it increased to 14% (and was almost identical in magnitude to the significant difference between the unmixed + spaced and unmixed + massed conditions). However, we could not isolate the unique effects of spacing and mixing in this 14% advantage because their effects were confounded. Nevertheless, we speculated that the difference between the mixed and unmixed + massed conditions in Experiment 4 was driven by the benefits of spacing rather than mixing topics *per se*. In the current experiment, no benefit of mixing topics was found when spacing was held constant, but there was a benefit of increased spacing when topics were not mixed.

7. General discussion

Millions of students study with flashcards every day. Although students often study more than one topic at a time, they almost unanimously reject the idea of mixing together flashcards from different topics while studying (Wissman et al., 2012). We hypothesized that mixing would actually enhance learning by increasing the amount of spacing between repetitions of a given flashcard.

There was a benefit of mixing topics in Experiment 1, but it was not significant and it was not replicated in Experiment 2 with a longer test delay (48 h vs. 1 week). There was no benefit of mixing topics in Experiment 3 or 4 either, with a one-week test delay and study trials happening during two sessions 48 h apart.

It is unclear why mixing topics did not enhance learning despite the fact that it increased spacing. It could be that the spacing that resulted from mixing simply did not enhance learning. Given the consistent and robust spacing effects in prior research (e.g., Cepeda et al., 2006), it is also possible that spacing did have a positive effect on learning. If so, it must have been counteracted by a negative effect of a similar magnitude. Thus, it remains possible that mixing topics impaired learning. One explanation of this (possible) impairment, which we discussed in the introduction, is that learning benefits when similar items are juxtaposed (i.e., studied on consecutive trials) through a process called discriminative contrast. We argued in the introduction that discriminative contrast does not necessarily apply to tasks, like learning word pairs, in which discriminating between different items is not difficult. Despite this argument, the data suggest that it is possible that juxtaposing word pairs of the same type is beneficial (whether or not the mechanism behind this benefit is discriminative contrast). Research that holds total spacing constant while comparing interleaved and blocked studying would be needed to further explore this issue (cf. Birnbaum et al., 2013; Kang & Pashler, 2012).

7.1. Practical applications

The results presented here suggest that mixing flashcards from two topics does not enhance student learning. This conclusion seems to contradict the predictions of both researchers and students. Because mixing topics increased the spacing between repetitions of any particular item, we believed that it would enhance learning. Students' survey responses make it clear that they believed it would impair learning. Neither set of beliefs was supported.

It is unclear why students consider mixing to be such a bad strategy. It seems likely that the students were right when they said mixing would be confusing. Indeed, performance during the study phase showed that mixing made it harder to remember previously studied items in comparison to studying topics separately (see Table 1). Unfortunately, short-term performance during the study phase can be a misleading predictor of long-term learning (Bjork, 1994). Thus, mixing topics seems to be yet another example of a general principle: It is a mistake to interpret difficulty while studying as a sign that one is not learning (e.g., Benjamin, Bjork, Schwartz, 1998; Bjork et al., 2012). When students choose a study strategy, it is important they choose one that is based on long-term learning and not performance while studying.

When making practical recommendations based on research, it is often tempting to assume that effects generalize to new situations. In this case, it was tempting to assume that mixing topics would enhance learning because mixing increases spacing and spacing enhances learning. This assumption was not supported. Thus an ironclad law of research asserted itself once again: the only real way to find out whether or not an intervention works is to try the intervention and see if it works.

Conflict of interest statement

The authors declare that they have no conflict of interest.

Acknowledgements

A Scholar Award granted to the second author by the James S. McDonnell foundation supported this research. Doug Rohrer provided valuable feedback on this project.

Appendix.

Indonesian		Anatomy	
Cue	Target	Cue	Target
Terlambat	Late	Of the elbow	Cubital
Tinggal	Live	Of the head	Cephalic
Perhiasan	Jewelry	Of the ear	Otic
Keberangka	Departure	Of the cheek	Buccal
Sandiwara	Theater	Of the belly	Ventral
Angin	Wind	Of the hip	Coxal
Sungai	River	Of the armpit	Axillary
Sabun	Soap	Of the ankle	Tarsal
Telur	Egg	Of the foot	Pedal
Baru	New	Of the heel	Calcaneal
Jelek	Bad	Of the arm	Brachial
Basah	Wet	Of the hand	Manual
Gendang	Drum	Of the wrist	Carpal
Makan	Eat	Of the thigh	Femoral
Makanan	Food	Of the eye	Orbital
Kacamata	Eyeglasses	Of the back	Dorsal

References

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68.

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*, 392–402.

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

Bjork, R. A., Dunlosky, J., & Kornell, N. (2012). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In R. Bjork, & E. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1250–1257.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317.

Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories: Is spacing the enemy of induction? *Psychological Science*, *19*(6), 585–592.

Kornell, N., & Bjork, R. A. (2008b). Optimising self-regulated study: The benefits – and costs – of dropping flashcards. *Memory*, *16*(2), 125–136.

Mayfield, K. H., & Chase, P. N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis*, *35*, 105–123.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

Pyc, M. A., & Rawson, K. A. (2012). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition*, *40*(6), 976–988.

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 242–248.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, http://dx.doi.org/10.1007/s10648-012-9201-3
Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481–498.
Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848.
Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750–763.
Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568–579.