Received: 29 May 2019 | Revised: 28 August 2020 | Accepted: 12 September 2020

DOI: 10.1002/acp.3740

WILEY

RESEARCH ARTICLE

The most efficient sequence of study depends on the type of test

Paulo F. Carvalho[1] | Robert L. Goldstone[2]

[1]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania

[2]Department of Psychological and Brain Sciences and Cognitive Science Program, Indiana University, Bloomington, Indiana

Correspondence
Paulo F. Carvalho, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA.
Email: pcarvalh@cs.cmu.edu

Funding information
National Science Foundation, Division of Behavioral and Cognitive Sciences, Grant/Award Number: 1824257; National Science Foundation, Division of Research on Learning in Formal and Informal Settings, Grant/Award Number: 0910218; Fundação para a Ciência e a Tecnologia, Grant/Award Number: SFRH/BD/78083/2011; Institute of Education Sciences, Grant/Award Number: R305A1100060

Summary

Across three experiments featuring naturalistic concepts (psychology concepts) and naïve learners, we extend previous research showing an effect of the sequence of study on learning outcomes, by demonstrating that the sequence of examples during study changes the representation the learner creates of the study materials. We compared participants' performance in test tasks requiring different representations and evaluated which sequence yields better learning in which type of tests. We found that interleaved study, in which examples from different concepts are mixed, leads to the creation of relatively interrelated concepts that are represented by contrast to each other and based on discriminating properties. Conversely, blocked study, in which several examples of the same concept are presented together, leads to the creation of relatively isolated concepts that are represented in terms of their central and characteristic properties. These results argue for the integrated investigation of the benefits of different sequences of study as depending on the characteristics of the study and testing situation.

KEYWORDS

example, learningconcept learning, interleaving, interrelated concepts, learning, study sequence

1 | INTRODUCTION

Understanding how students should organize their study to promote learning has emerged as a major area of active research interest in learning and cognitive sciences. Because so much of instruction is based on demonstrations or practice activities, changing how information is sequenced can be a simple, low-cost intervention.

Considerable previous research has identified interleaved study (alternating between examples of the to-be-learned concepts, e.g., alternating examples of concepts A and B: ABABABABAB), as improving learning. For example, Kornell and Bjork (2008) showed that when learning artist styles, test performance was better following interleaved rather than blocked study (seeing all examples of one concept before starting examples of another: AAAAABBBBB) . Later work has suggested that this benefit is likely related to interleaving examples of different concepts promoting discrimination between concepts during study (e.g., Birnbaum, Kornell, Bjork, & Bjork, 2013; Kang & Pashler, 2012). However, the fact that interleaved study does not always lead to improved testing performance–as a discrimination theory would imply–together with findings that blocked study improves learning of concepts that have high within-category diversity in their properties (e.g., the category "mammal" which includes bats, cows, and whales) has led to more refined theories emphasizing how the sequence of study shapes what is attended and therefore what is learned and remembered, as opposed to focusing solely on discrimination (Carvalho & Goldstone, 2014b).

Despite extensive research showing that the benefits of different sequences extend and interact with different types of materials (e.g., Carpenter & Mueller, 2013; Carvalho & Goldstone, 2014b), retention intervals (e.g., Carvalho & Goldstone, 2014a), study conditions (e.g., Carvalho & Goldstone, 2015a), individual differences in working memory (e.g., Sana, Yan, & Kim, 2016), and self-regulation

(e.g., Carvalho, Braithwaite, de Leeuw, Motz, & Goldstone, 2016), it is still not well-understood which differences–if any–exist between what is learned from different sequences. Do learners acquire different representations of studied materials from different sequences?

As described above, previous work has focused on changing the study and encoding conditions and investigating what effect–if any–they have on performance on the same test. In the current work, we investigate the other half of the equation, that is, the properties of the testing situation. What we learn is critically tied to our experience, and so, to the extent that different sequences create different representations, they might lead to different learning that can be used in different ways. Such understanding is critical for both theory and practice and has not been addressed before. If, instead of conceptualizing how one sequence improves learning (as discrimination theories do; e.g., Kornell & Bjork, 2008; Birnbaum et al., 2013; Kang & Pashler, 2012), one understands *how* learning is shaped by sequence and what are the learning consequences of each sequence and their strengths, we not only have a better model of learning as a whole but can use it to offer better, targeted, suggestions for practice. Instead of "always interleave study" we can suggest "interleave study if… but block it instead if…."

The Sequential Attention Theory (Carvalho & Goldstone, 2015b, 2017), proposes a mechanism and computational process (Carvalho & Goldstone, 2019) for how the sequence of study affects learning by affecting what is attended and encoded during learning. According to SAT, during blocked study attention and encoding are progressively directed toward the similarities among successive items belonging to the same category, whereas during interleaved study attention and encoding are progressively directed toward the differences between successive items belonging to different categories. Because of this influence on cognitive processing, Carvalho and Goldstone (2017) propose that the sequence of study can accelerate or delay learning, depending on whether the constraints created by the sequence of study match those of the encoding situation (e.g., interleaved study in situations critically hinging on the encoding of differences between concepts, such as the study of highly similar concepts), or mismatch it (e.g., blocked study in the same situations).

In the current work, we aim to extend Carvalho and Goldstone's proposal to demonstrate that different encoding experiences will result in creating different representations of the concepts based on the same examples. We hypothesize that, if Carvalho and Goldstone's theory is correct and different information is encoded with different sequences of examples, then different sequences of examples potentiate different representations of what was studied. This general hypothesis is consistent with contemporary views of the transfer-appropriate encoding theory suggesting that best test performance is achieved when encoding is similar to the demands of the testing situation (e.g., Lockhart, 2002), and extends it beyond memory tasks.

More specifically, encoding the differences between concepts through interleaved study of examples will tend to lead to the creation of *interrelated concepts* whose representations are *contrasted away from each other* by emphasizing or exaggerating their distinctive characteristic relative to each other (Corneille, Goldstone, Queller, & Potter, 2006; Goldstone, 1996). For example, when studying two mathematical operations interleaved the learners' representation of the operations would be what differentiates one from the other. Conversely, blocked study will tend to lead to encodings of the similarities within each concept that will, in turn, create relatively *isolated, stand-alone, representations* that do not depend on the characteristics of the other concepts studied in close proximity (Goldstone, 1996). Using the same example, blocked study of the same two mathematical operations would yield representations that include non-discriminative features of the concepts (shared properties) but might nonetheless be important.

One way to test this proposal, the one used across the four experiments in this paper, is to test not whether interleaved or blocked study are more beneficial for learning by keeping the test constant, as in previous work, but instead vary the type of test used, investigating which tests yield differences between interleaved and blocked study. If interleaved study leads to the creation of interrelated representations of two concepts, then it should improve learning in situations where appreciating differences between concepts is critical for good performance (as in multiple-choice classification). Similarly, if blocked study leads to the creation of relatively isolated representations of each concept, then it should improve learning in tests that require access to self-contained information within each concept rather than fine differences across trained concepts, such as when a concept must be differentiated from other new concepts possessing new distinctive features.

For this purpose, we developed four experiments in which learners studied concepts of psychology (e.g., "Hindsight bias") developed by Rawson and collaborators (Rawson, Thomas, & Jacoby, 2015; see e.g., Appendix), in one of the sequences and were then tested in different situations, similar to common study practices by students. Importantly, some of the tests required discrimination between different concepts (e.g., multiple-choice test), whereas others required an independent representation of each concept (e.g., writing a definition). We consider writing a definition to require an independent representation because these definitions can be expressed without referring to other learned concepts. For example, a student could write a definition for "availability heuristic" without having learned or remembered any of the other studied concepts (Goldstone, 1996).

To foreshadow, in Experiment 1 (and a follow-up study Experiment 1b) we show that learners write better definitions following blocked study compared to interleaved study, in the absence of differences between the two sequences for other tests. In Experiment 2 we replicate these results and extend them to show that learners write more correct definitions following interleaved study compared to blocked study in a situation where discriminating between the two concepts was challenging. Finally, in Experiment 3, we show using a 2-alternative forced-choice classification task test that interleaved study results in better discrimination between the concepts studied, but not among non-similar concepts. Overall, these results are consistent with the hypothesis that interleaved study promotes interrelated representations whereas blocked study promotes isolated, stand-alone, representations of the materials studied.

TABLE 1 Participant characteristics for Experiments 1, 1b, 2, and 3

	Exp. 1	Exp. 1b	Exp. 2	Exp. 3
Mean age (SD)	33 (10)	35 (12)	36 (11)	31 (9.88)
Gender (% females)	54.5	N/A[a]	68	N/A[a]
Education (% Bachelor's degree or higher)	50	70	64	58
% native English speakers	95	100	93	93

Note: Standard deviations in parenthesis.
[a]Due to a coding error, this information was not registered for this experiment.

2 | EXPERIMENT 1

This study explored whether different sequences of study can affect test performance differently when different measures of learning are used. We contrasted three types of tests: Multiple-Choice classification of novel examples, Multiple-Choice classification of correct definitions, and Writing definitions. In the first two tests, participants must use the knowledge acquired during study to classify new information, either a new example or a definition they never saw before. These tests, because they present a new instance and all the possible concepts for classification, are likely to require discrimination between the concepts studied. Writing a new definition, on the other hand, only requires knowledge about the concept the participant needs to describe. Therefore, we predict that in tests that emphasize isolated and independent knowledge of the properties of each concept–such as writing a definition–learners will perform better following blocked study. Conversely, for tests that require discriminating different concepts, that is, those that involve choosing between several categories in which to place an example, learners will perform better with interleaved study.

2.1 | METHOD

2.1.1 | Participants

A group of 28 people was recruited through Amazon's Mechanical Turk (https://www.mturk.com/). Data from 6 participants were excluded from analyses because of possible compliance issues (see below for details). The demographic characteristics of participants in the overall sample are presented in Table 1. Participants were asked to indicate in years their age when they learned English (with 0 if native speakers).

2.2 | Stimuli

We used a stimulus set of introductory concepts and examples created by Rawson et al. (2015). The stimuli included 10 concepts taught in Introductory Psychology and 10 example situations for each concept, collected from textbooks of Introductory Psychology. The concepts were divided into two groups by relatedness. Each group contained unrelated concepts only, whereas across groups pairs of related concepts existed (see Table 2). Relatedness of the concepts was judged by the authors by comparing the definitions of the concepts and confirmed by analyzing the pattern of errors in a confusion matrix using data from multiple-choice questions without feedback in a pilot study. Previous research looking at the sequence of study using these materials used this concept grouping as well (Rawson et al., 2015). Examples of concept definitions and example-situation are represented in Appendix.

TABLE 2 Groups of concepts used in Experiment 1

Concept pair	Group A	Group B
1	Availability heuristic	Representativeness heuristic
2	Door-in-the-face technique	Foot-in-the-door technique
3	Hindsight bias	Counterfactual thinking
4	Fundamental attribution error	Deindividuation
5	Mere exposure effect	Social facilitation

Note: Each row includes a pair of related concepts. Columns contain only unrelated concepts.

2.3 | Design and procedure

This Experiment had two conditions manipulated within-subjects: Study Sequence (Blocked vs. Interleaved) and Type of Test (Multiple-Choice Test vs. Definition Match Test vs. Write Definitions Test).

A schematic depiction of the study design is presented in Figure 1. The experiment had three phases: pretest, study, and test. Participants completed one pretest, two study phases, and two test phases in the following order: Pretest–Study 1–Test 1–Study 2–Test 2.

The first and second study phases were the same in every aspect except for the sequence of study and the concepts studied. One study phase was interleaved and the other blocked (order counterbalanced across participants). A different group of to-be-learned concepts was used in each study phase (order counterbalanced as well). In the interleaved condition learners studied an example of each concept before studying the same concept again. The same sequence of concepts was repeated until all examples of each concept were studied (e.g., ABCABC…). Conversely, in the blocked condition learners studied all examples of each concept before starting a new concept (e.g., AABBCC…). Moreover, the test phase only tested the concepts learned in the immediately preceding study phase. Between each study and test phase, participants completed a distractor task. In this

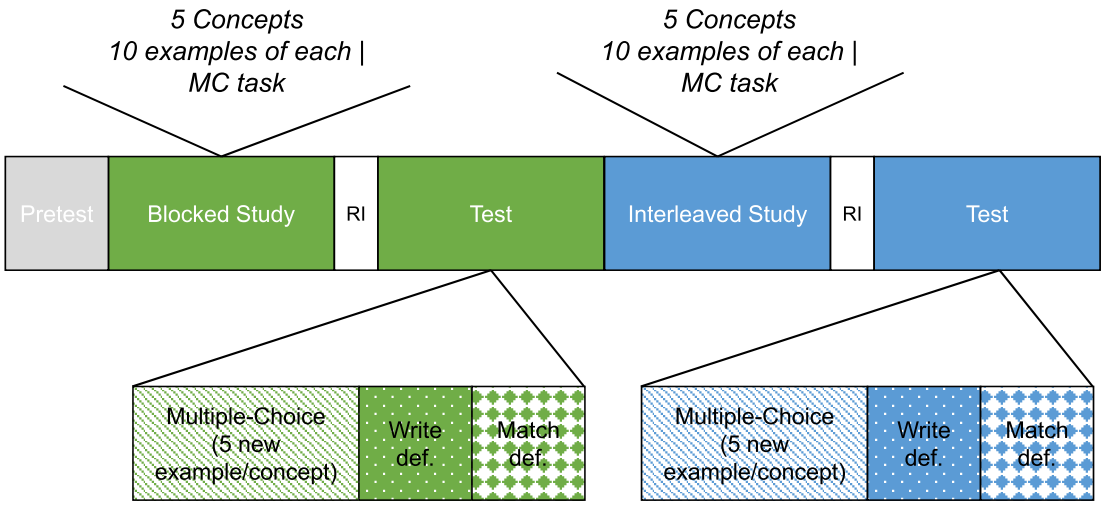

FIGURE 1 Schematic representation of the procedure used in Experiments 1 and 1b. During study, participants studied 5 concepts by reading examples and classifying them in a multiple-choice task with feedback. RI indicates the 4-min retention interval between study and test phases. Each test phase included three types of tasks, always presented in the same order: Multiple-choice, writing definitions, and classifying definitions with concept names. The order in which participants studied concepts in each sequence was counterbalanced across participants [Colour figure can be viewed at wileyonlinelibrary.com]

task, participants watched a 4-min video on an unrelated topic and answered a question about the video.

During the pretest phase participants were told that they would be presented with several psychology concepts that they were asked to rate regarding their familiarity/knowledge. Participants were told that not knowing the concepts was not an issue for the study, would not impact their eligibility or payment and that they should be honest in their responses. On each trial, the name of a concept was presented and participants had to rate on a scale from 1 ("Not familiar at all") to 7 ("Very familiar") how familiar they were with that concept. Following each rating, participants were asked to provide an example of that concept or enter "I don't know" if they did not know any. Participants completed the pretest for the 10 to-be-studied concepts across both study phases.

Following the completion of the pretest, participants completed the first study phase. During the study phase, participants read examples of situations depicting each of five concepts, one at a time and were asked to choose the name of the concept they thought the example instantiated. Participants were given feedback after each response (see Figure 2 for an example of a study trial). Participants studied five examples of each of the five concepts.

During the test phase, participants completed three types of tests: Multiple-Choice, Writing Definitions, and Match Definitions, always in that order. The Multiple-Choice test used the same procedure as the study phase with new examples and without feedback. In the Writing Definitions, test participants were shown the name of each of the concepts studied one at a time and asked to write the best definition possible for that concept, based on what they had learned in the previous study phase. In the Match Definitions test participants were presented with the textbook definition of each concept, one a time, and asked to identify what concept that definition belonged to by pressing the corresponding button on the screen. The order of the tests was kept constant because some tests provide an answer to the other tests. For example, classifying a correct definition as the concept it defines renders a subsequent test of writing a definition potential trivial. The order of trials within each of the tests was randomized across participants. None of the test phase tasks had any time limit.

3 | RESULTS AND DISCUSSION

Because the study was conducted online without experimenter supervision, we first inspected the data to identify potential compliance issues. For each participant, we calculated the median response time during both study phases. The sample's median response time to complete the study phase was 10.5 s per problem (max: 22.9 s/problem; min: 0.73 s/problem). We calculated the 10th and the 90th percentiles for the distribution of median response times, 3.3 s/problem and 16 s/problem respectively and used these values as a measure of non-compliance in the task. Responding too fast (faster than the 10th percentile) is likely due to participants who are not reading the problems and just advancing through the experiment quickly; similarly, longer response times (above that of the 90th percentile) are likely due to potentially distracted participants. Six participants were identified based on this analysis and their data were excluded from further analyses.

All the analyses below are ANCOVA analyses including average pretest score and counterbalancing condition as covariates. Data and learners' responses to open-ended questions for this and subsequent studies are available through OSF (https://osf.io/y4a2r/).

3.1 | Pretest

To analyze the data from the pretest we calculated 25th, 50th, and 75th percentiles of the familiarity ratings (see Table 3). As can be

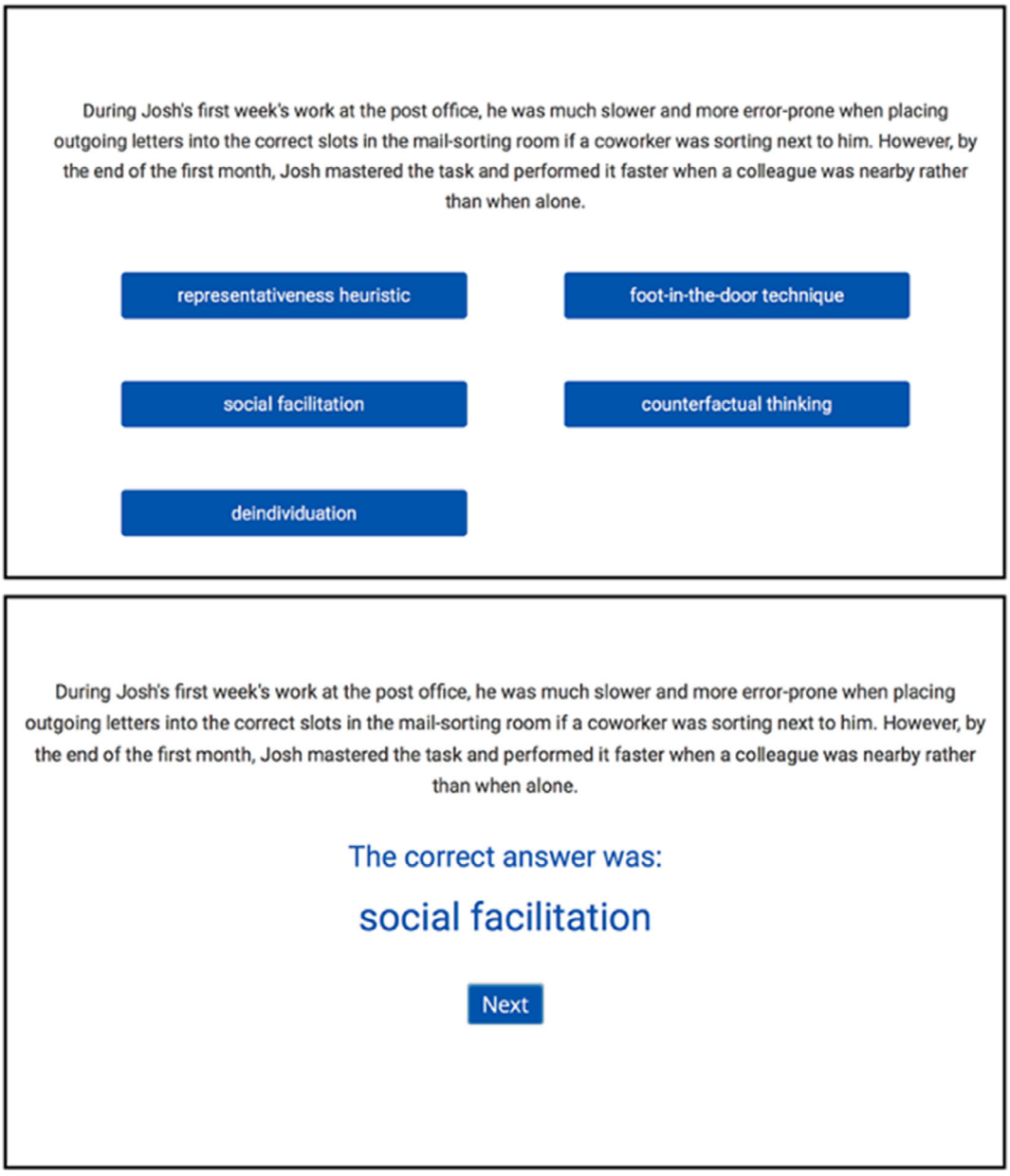

FIGURE 2 Example of the study interface of Experiments 1 and 2. The top image illustrates the interface during study and test: A situation is described on top and the name of 5 concepts is presented below. The participating was asked to read the description carefully and decide which concept it illustrated by pressing the corresponding button. The bottom image illustrates the feedback screen (study phase only). The description is presented again along with the correct response [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Overall results of familiarity ratings for Experiments 1, 1b, 2, and 3

	25th percentile	50th percentile	75% percentile	Mean	SEM
Experiment 1	1.000	1.550	2.025	1.707	0.155
Experiment 1b	1.150	1.600	2.400	1.988	0.133
Experiment 2	1.343	1.792	2.388	1.97	0.160
Experiment 3	1.500	2.300	3.300	2.37	0.200

seen, most learners showed little or no knowledge of the to-be-studied concepts (mean of ~2 in a 1–7 scale) at pretest. Inspection of the provided examples further confirmed this interpretation (because most participants did not provide any example or wrong examples we did not further analyze their examples).

3.2 | Study phase

Mean performance during the blocked study phase was 72% (SEM = 5%), whereas during interleaved study it was 67% (SEM = 5%). This difference was not statistically significant, $F(1, 20) = 1.95$, $p = .169$, $\eta^2_G = .012$.

3.3 | Test phase

Two trained coders, blind to condition assignment, rated as correct or incorrect each of the written definitions. These two coders agreed 87% of the time and inter-coder reliability was high, Cohen's Kappa = .725, $p < .0001$. Disagreements were resolved by a third coder, also blind to condition assignment of the responses.

Performance for the test phase is depicted in Figure 3. As can be seen in the figure, the type of tests varied in their level of difficulty, with participants performing better in the Definitions Match test and worse in the Write Definitions test, $F(2, 42) = 17.62$, $p = < .0001$, $\eta^2_G = 0.151$. Although there was no overall main effect of study sequence, $F(2, 42) = 1.36$, $p = .256$, $\eta^2_G = .006$, there was a significant interaction between type of test and study sequence, $F(2, 42) = 5.26$, $p = .022$, $\eta^2_G = .022$.

To further investigate this interaction, we compared the effect of type of study sequence on each of the tests by calculating the difference in performance following blocked and interleaved study for each type of test (interleaved–blocked). The difference in performance between the two conditions varied across type of test, $F(2, 42) = 5.10$, $p = .011$, $\eta^2_G = .110$. Planned contrasts using FDR correction (Benjamini & Hochberg, 1995) indicate that the effect of study sequence was significantly different when comparing the Write Definitions test ($M = -.15$, SEM = .06) with the Multiple-Choice test ($M = .009$, SEM = .03), $p = .033$, $d = .593$, and the Match Definitions test ($M = .02$, SEM = .05), $p = .040$, $d = 0.508$, but not when comparing the Multiple Choice and the Match Definitions tests, $p = .844$, $d = 0.042$.

These results are consistent with our proposal that blocked study encourages learners to develop independent, stand-alone representations rather than highlighting diagnostic features (i.e., those that discriminate between the concepts). Interleaved study emphasizes features that discriminate between concepts, which would be more helpful for a subsequent categorization task than a task that requires generation of a stand-alone definition of the concept.

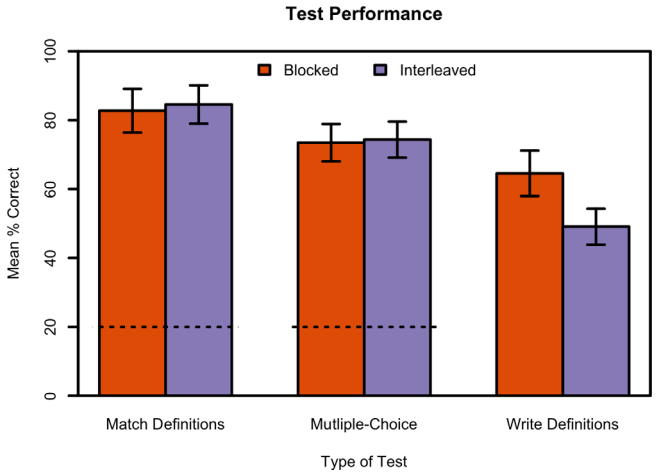

FIGURE 3 Results for the Test Phase of Experiment 1. Dotted lines represent chance level. Error bars represent standard errors of the mean [Colour figure can be viewed at wileyonlinelibrary.com]

EXPERIMENT 1B

We had two main goals for this experiment: (a) extend the results of the previous experiment, and (b) investigate whether the content of the test (definitions vs. examples) influences learners' performance differently following different sequences.

In the previous experiment, we saw that learners provided more correct definitions when concepts had been studied blocked compared to interleaved, along with no differences between the two study sequences when completing multiple-choice classification of new examples or classifying definitions tests. However, the writing definitions test differs from the other types of tests used in Experiment 1 in at least a couple of aspects. As we argued before, writing definitions involves stand-alone, possibly independent representations of the concepts, contrary to multiple-choice tests which benefit from identifying and highlighting diagnostic properties of the concepts relative to one another. However, they also differ in terms of how difficult they are and the nature of the task; tests in which learners are required to choose an option are typically easier than open-ended questions in which learners must write their answer (e.g., Anbar, 1991; albeit with important drawbacks, cf. Cantor, Eslick, Marsh, Bjork, & Bjork, 2015; Roediger & Marsh, 2005). This difference in difficulty was found for our tasks too. Moreover, a multiple-choice test closely matches the study procedure, whereas writing a definition is a substantially different transfer test. Any of these aspects could be the basis of the differences found in the previous experiment.

For these reasons, in the current experiment we included a new type of test–writing examples. Consequently, the design of Experiment 1b includes two factors manipulated at test: (a) whether the test required choosing an option on the screen or writing a response (modality of response) and (b) whether the test involved providing a response related to examples or concept definitions (content of the test).

The combination of modality of the test and the content being probed results in different types of knowledge being required for good performance, and by hypothesis, the type of knowledge prioritized determines which sequence of study is more beneficial. Any writing test might benefit from blocked study either because of its added difficulty or because of the mismatch between test and study. Alternatively, as discussed before, writing the definition of a concept is likely to prioritize knowledge about the characteristic features of the concept–the properties that are frequent among most instances of that concept, even if not discriminative of that concept compared to others. How are these two types of features different? Imagine that you need to indicate from a set of images which ones are bats and which ones are birds. The feature “wings” is not discriminative (both species tend to have “wings”), but it is characteristic of bats (and birds) because bats tend to have this feature.

Conversely, providing an example does not necessarily involve these same processes. It is possible, for example, that learners generate examples that emphasize how that concept is different from others or generate an example similar to one they studied without activating specific characteristic or discriminative features. Put another way, whereas writing a definition might frequently emphasize

characteristic features of the stimuli, writing an example might be approached in several ways that require different representations/memories.

Thus, how the sequence of study affects performance on tests about definitions compared to tests about examples will allow us to determine whether the results seen in Experiment 1 were due to different representations promoted by different sequences, as we proposed, or instead by the difficulty of the writing definitions test or the mismatch between study and transfer tasks. If our proposal is correct, then we expect to see no difference between interleaved and blocked study for classification tests (regardless of whether they require classifying an example or a correct definition to the concept they instantiate). Moreover, we expect an effect of sequence of study for tests that require learners to write a definition but not for tests that require learners to generate an example.

Finally, to extend the results to new situations we introduced two other main changes: study sequence was manipulated between participants and, after completing the pretest, learners were given the textbook definition for each of the concepts. These changes increase the generalizability of the findings and their applicability.

4 | METHOD

4.1 | Participants

A group of 81 people completed the experiment following recruitment through Amazon's Mechanical Turk (https://www.mturk.com/). Participants were randomly assigned to one of the two study sequence conditions (N = 41 for the Blocked condition and N = 40 for the interleaved condition). Data from 16 participants were excluded from analyses due to non-compliance following the same procedure as in the previous experiment (see below for details). Data from one additional participant were excluded due to self-reported previous participation in another experiment with the same materials. The final sample included 64 participants (N = 32 in each of the conditions). Characteristics of participants in the overall sample are presented in Table 1.

4.2 | Stimuli and procedure

The stimuli and procedure were the same as in Experiment 1, except for the differences noted below.

Participants completed the following phases, always in this order: Pretest–Reading Definitions–Study–Test. Participants completed only one study phase, either interleaved or blocked, and the corresponding test phase. The pretest phase was similar to the one in the previous experiment except that participants were shown only five concepts (the five they would study).

During the Reading Definitions phase, participants were told that they would be presented with the definitions of the five concepts, that these were important for the task and that they should read them carefully. Participants then saw one definition at a time, along with the name of the concept it defined and could study it for as long as they wished. Immediately following the Reading Definitions phase, participants were given the instructions for the Study phase and completed the study phase just like in Experiment 1.

Following the study phase participants played a Tetris game for 1 min before the instructions for the test phase were presented. During the test phase participants completed the following tests, always in this order: Multiple-Choice, Writing Definitions, Match Definitions, Write Examples (see Table 4 for a description of each test). All tests used the same procedure as in the previous experiment. During the Write Examples test, participants were asked to write a good example describing a situation that illustrated the concept presented (no limitations on the examples were enforced, thus studied examples or slight variations were allowed). Participants were presented with the name of a concept one at a time. Which group of five concepts was used was counterbalanced across participants.

5 | RESULTS AND DISCUSSION

We followed the same procedure as in the previous experiment to identify potentially non-compliant participants. The sample's median response time to complete the study phase was 13.3 s per problem (max: 23.0 s/problem; min: 0.63 s/problem). The 10th and 90th percentiles for the distribution of median response times were 4.0 s/problem

TABLE 4 Description of what was presented and requested of participants for each test task (Experiments 1, 1b, and 2)

	Multiple-choice test	Writing definitions test	Match definitions test	Write examples test
Presented	A vignette representing a situation corresponding to one of the studied concepts is presented.	Name of a studied concept.	The definition of one of the studied concepts.	Name of a studied concept.
Requested	Participants classified it by pressing the corresponding concept name among the options.	Participants are asked to write the best definition possible for the concept.	Participants classified it by pressing the corresponding concept name among the options.	Participants are asked to write the best example possible for the concept.

Note: Not all tasks were used in all experiments, see text for details.

and 15 s/problem respectively. Sixteen participants were identified as outliers based on this analysis and their data excluded from further analyses.

All the analyses of variance presented below are ANCOVA analyses including mean pretest score, counterbalancing condition and mean time spent reading the definitions as covariates.

5.1 | Pretest

As in the previous experiment, participants showed little to no knowledge of the to-be-studied concepts (see Table 3).

5.2 | Study phase

Mean performance during the study phase was higher (M = 88%, SEM = 1.4%) for the blocked group compared to the interleaved group (M = 77%, SEM = 3%), $F(1, 62) = 4.21$, $p = .044$, $\eta^2_G = .064$. The amount of time participants spent reading the definitions just before the study phase was not correlated with performance during the study phase, $r = .048$, $p = .707$.

5.3 | Test phase

Two trained coders, blind to condition assignment, rated as correct or incorrect the responses to the Writing Definitions and Writing Examples tests. For the Writing Definitions test, the two coders agreed 84% of the time and inter-coder reliability was high, Cohen's Kappa = .679, $p < .0001$. For the Writing Examples test, the two coders agreed 88% of the time, and inter-coder reliability was also high, Cohen's Kappa = .764, $p < .0001$. Disagreements were resolved by a third coder, also blind to condition assignment of the responses.

Overall, the time participants spent reading the definitions initially presented was not correlated with test performance, $r = .064$, $p = .309$, or with performance on any of the tests separately, $-.065 < r < .214$, $ps > .090$.

For analyses, we classified each test based on the type of content being probed (Definitions vs. Examples) and the modality of the test (Classify vs. Write). The results of the test phase are presented in Figure 4.

Overall learners were more accurate when the test required classification by choosing an option among several than writing a response, $F(1, 62) = 67.24$, $p < .0001$, $\eta^2_G = .184$. That is, writing tests were, overall, more difficult. Similarly, learners were overall more accurate when the test asked about definitions compared to examples, $F(1, 62) = 10.66$, $p = .002$, $\eta^2_G = .019$. Importantly, as can be seen in Figure 4, there is a three-way interaction between type of content, modality of test and the sequence of study, $F(1, 62) = 4.00$, $p = .049$, $\eta^2_G = .007$. Following interleaved study, participants are better at Classification tests about Definitions than about Examples, $t(31) = 3.56$, *corrected* $p = .002$, $d = .629$, whereas there is no difference in performance between Writing tests about Definitions and Examples, $t(31) = 0.22$. *corrected* $p = .827$, $d = .039$. Conversely, following blocked study, regardless of the type of test, participants are

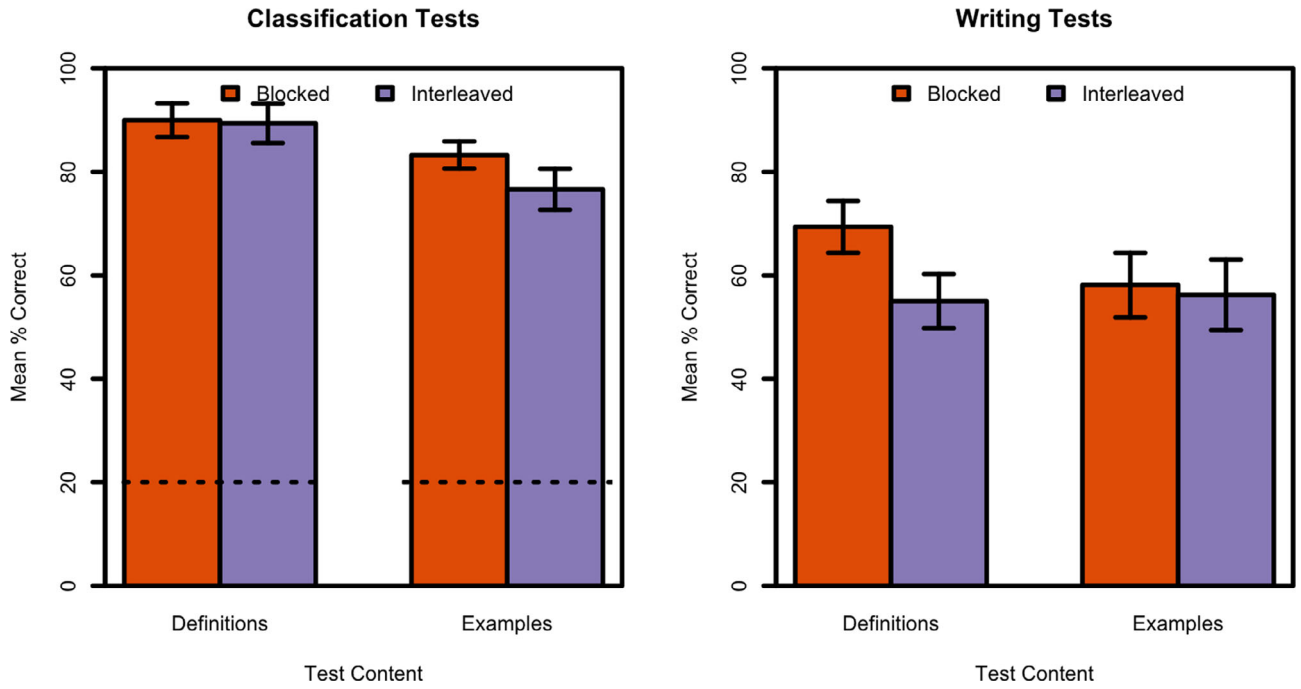

FIGURE 4 Results for the Test phase of Experiment 1b. Results for the Classification Tests are presented on the left and results for the Writing Test are presented on the right. The four types of tests were classified based on their characteristics: Multiple-Choice (Class cation of examples), Writing Definitions (Writing Definitions), Match Definitions (Classification of Definitions), Write Examples (Writing Definitions). Dotted lines indicate chance level of performance. Error bars represent standard errors of the mean [Colour figure can be viewed at wileyonlinelibrary.com]

more accurate when asked about Definitions compared to Examples, $F(1, 31) = 8.75$, $p = .006$.

Furthermore, for this study the critical contrasts are those between performance in writing definitions and writing examples tests following blocked vs. interleaved study. Planned contrasts correcting for multiple comparisons using FDR correction show no difference between the two study sequences in tests that ask learners to write examples, $t(62) = 0.203$, corrected $p = .840$, $d = 0.050$, but a marginally significant difference for tests that ask learners to write definitions, $t(62) = 1.99$, corrected $p = .103$, $d = 0.496$.

In sum, although one of the critical analyses falls below statistical significance, as a whole the results of this follow up experiment do suggest that the findings of Experiment 1 are more likely due to blocked study leading to isolated representations that help learners in tasks about definitions. If the results were due to the difficulty of the writing definitions task compared to the other tasks in Experiment 1, we would have seen an overall benefit of blocked study for all harder tasks (writing examples and writing definitions) in Experiment 1b. Instead we see that participants are overall more accurate when asked about definitions than examples, even when writing examples is harder and classifying definitions is overall easier.

Interestingly, although the inclusion of correct definitions at the start of the study might have contributed to the smaller effect size found in this experiment compared to Experiment 1, it did not eliminate it, indicating that even when learners are provided with the definitions, further study with examples affects the type of representations that are built and available at test.

6 | EXPERIMENT 2

Our main proposal in this paper is that blocked study creates relatively independent representations of each concept studied which emphasizes the concept's characteristic features. These independent representations include more details from each concept than what is fostered by the relatively interrelated representations created during interleaved study. In the context of studying examples of different concepts, we proposed that blocked study allows learners to more successfully write definitions of the concepts because a definition requires the type of knowledge that blocked study promotes; it is generally possible to write good definitions for the learned psychology concepts without mentioning other psychology concepts learned at the same time. Consistent with this hypothesis, Experiment 1 showed that following blocked rather than interleaved study, learners were more successful at writing definitions of concepts, but the groups did not differ on classifying examples.

However, when two concepts are highly related (e.g., "foot-in-the-door technique" and "door-in-the-face technique") their definitions can be aptly construed concerning each other. If they are studied together, one central feature to include in the definition is the feature that *discriminates* them. Thus, the fact that in the previous experiments learners studied in the same session concepts that were dissimilar from each other and varied in many properties (see Table 2) might have contributed to the pattern of results seen. Would studying similar concepts together change the pattern of results observed?

Studying related concepts together changes the learning task in at least two critical ways. First, studying similar concepts in the same session might result in the necessity to discriminate between similar situations to find the subtle differences between the two types of concepts. It has been shown before that the interrelated representations promoted by interleaved study are likely to improve test performance when learning highly similar concepts (Carvalho & Goldstone, 2014b). Second, the features that discriminate these related concepts are also characteristic features of the concept, unlike what is the case when the concepts are dissimilar (see Table 2). This means that interleaved study could promote representations appropriate for a writing definitions test through identification of differences between concepts, whereas these differences were unlikely to be highlighted in the previous experiment.

In sum, when similar items are studied in the same session, there are several reasons to believe that performance would benefit from interleaved study, even when the test requires learners to write definitions. However, when similar items are studied in separate sessions, as in Experiment 1, blocked study would promote best performance in a test requiring isolated representations, such as writing definitions.

To test this, we used a procedure similar to how students often organize their study. In most natural situations, students are likely to randomly assign the topics to be studied to a study session or to follow the sequence of their textbook or instructor. Therefore, in this experiment we randomly assigned concepts to being studied either interleaved or blocked, instead of using different pre-defined groups of concepts that guarantee low between-category overlap as in the previous experiment. This results in a situation in which similar concepts might be studied together or separately. We compare performance on multiple-choice and writing definitions tests following blocked or interleaved study in each one of these situations.

7 | METHOD

7.1 | Participants

A group of 36 people completed the experiment following recruitment through Amazon's Mechanical Turk (https://www.mturk.com/). Data from three participants were excluded due to self-reported previous participation in another study with the same materials. Data from an additional eight participants were excluded from analyses because of possible compliance issues (see below for details). The final sample included 25 participants. Table 1 includes the demographic characteristics of participants in the overall sample.

7.2 | Stimuli and procedure

In this experiment, we used the same set of materials as in the previous experiment, but concepts were randomly assigned to be studied

interleaved or blocked. Thus, in this experiment we did not force related concepts to be studied in separate phases.

The procedure was similar to the procedure used in Experiment 1 except for the following differences. Participants studied only eight concepts, four interleaved and four blocked in a single study session. During the study session, participants saw four situations depicting each one of the concepts. After study, participants played a game of Tetris for 30 s. The test phase included only a multiple-choice test and a writing definitions test, always presented in that order. During the multiple-choice test participants saw a total of four novel examples of the concepts studied, presented one at a time, and were asked to indicate which concept it illustrated.

8 | RESULTS AND DISCUSSION

We identified potentially non-compliant participants using the participants' response times during study. The sample's median response time to complete the study phase was 8.4 s per problem (max: 22.8 s/problem; min: 0.47 s/problem). The 10th and 90th percentiles for the distribution of median response times were 2.6 s/problem and 16 s/problem respectively. Eight participants were identified as outliers based on their response times falling outside of this range and their data were excluded from further analyses.

In all the analyses presented below, mean pretest score and counterbalancing condition were included as covariates.

8.1 | Pretest

As in the previous experiments, participants showed little to no pre-training knowledge of the to-be-studied concepts (see Table 3).

8.2 | Study phase

Mean performance during the blocked study phase was 79% (SEM = 2.5%), while during interleaved study it was 73% (SEM = 4%). However, this difference was not statistically significant, $F(1, 25) = 2.65$, $p = .116$, $\eta^2_G = 0.04$.

8.3 | Test phase

Two trained coders, blind to condition assignment, rated as correct or incorrect each of the Written Definitions provided. The two coders agreed 84% of the time and inter-coder reliability was high, Cohen's Kappa = .611, $p < .0001$. Disagreements were resolved by a third coder, also blind to the condition assignment of the responses.

To analyze the results from the two tests used in this experiment we classified each concept based on whether it had been studied blocked or interleaved and whether its related concept (see Table 2) had been studied in the same sequence or different sequences. When both related concepts were studied in the same phase and the same sequence (e.g., "foot-in-the-door technique," "door-in-the-face

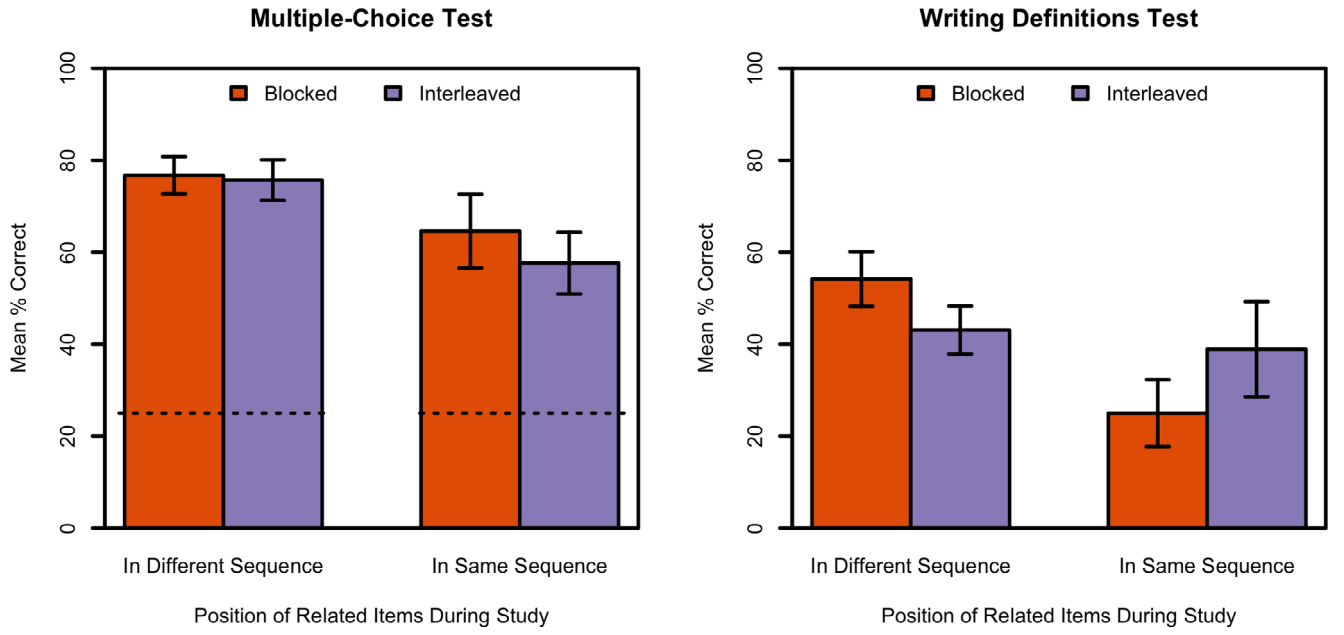

FIGURE 5 Results for the Test Phase of Experiment 2. Results for the Multiple-Choice test are presented on the left panel and results for the Write Definitions test are presented on the right panel. Pairs of similar concepts were classified as having been studied in the same sequence or different sequences (see text for details). The dotted line represents chance performance. Error bars represent standard errors of the mean [Colour figure can be viewed at wileyonlinelibrary.com]

technique" studied blocked), they were both classified with the sequence they were studied with and with "Same Sequence" (e.g., "Blocked" and "Same Sequence.") However, when only one of the related concepts was studied, or the two related concepts were studied in different phases/sequences, each was classified in the sequence studied and both were classified as "Different Sequences."

This classification of the concepts resulted in empty cells for learners who did not have both concepts studied in the Same Sequence and concepts studied in Different Sequences for both interleaved and blocked study. Because traditional repeated-measures ANOVA does not allow for the existence of empty cells and we wanted to maximize the inclusion of all data collected, here we used mixed model analyses (including pretest score as a covariate) and report Wald *F* tests and respective *p*-values using Kenward-Roger's approximation (Kenward & Roger, 1997; for applications to psychological research see e.g., Judd, Westfall, & Kenny, 2012). The results are depicted in Figure 5.

As we saw in the previous experiments, overall learners performed better on the Multiple-Choice test than when writing definitions, *Wald* $F(1, 33.842) = 91.68$, $p < .0001$. Similarly, the sequence of study had no overall effect on performance, *Wald* $F(1, 33.948) < 1$. No interaction was found between these two variables, *Wald* $F(1, 92.007) < 1$.

However, the relatedness between concepts presented in the same sequence influenced performance. Overall, when learners studied the two related concepts in the same sequence their performance was lower ($M = 46.52\%$, SEM = 1.50%) than when related concepts were not in the same sequence ($M = 62.41\%$, SEM = 1.04%), *Wald* $F(1, 24.284) = 15.60$, $p = .0006$. Item relatedness also interacted with sequence of study and type of test, *Wald* $F(1, 98.522) = 6.83$, $p = .010$.

To further analyze this interaction, we explored the test results for each type of test separately. For the Multiple-Choice test, only the effect of relatedness reached statistical significance, *Wald* $F(1, 24.264) = 9.43$, $p = .005$. However, for the Writing Definitions test, in addition to a significant effect of item relatedness, *Wald* $F(1, 25.649) = 9.81$, $p = .004$, we also found a significant interaction between item relatedness and sequence of study, *Wald* $F(1, 29.282) = 6.71$, $p = .015$ (see right panel of Figure 5). As predicted by the results of Carvalho and Goldstone (2014a, 2014b), the relative relatedness between items modulates the relative benefit of each sequence for the Writing Definitions test. Moreover, consistent with the results of Experiment 1, we see that when similar items are not studied in the same sequence, performance in the Write Definitions test benefits from blocked study, although this effect was only marginally significant, $t(35) = 1.90$, $p = .066$, $d = 0.317$.

EXPERIMENT 3

The results of the previous experiment suggest that whether interleaved or blocked study lead to best test performance depends on the similarity among concepts studied in the same session.We proposed that this is because interleaved study emphasizes discrimination and leads to creating interrelated representations that highlight the distinguishing features between the similar concepts. These differences can then be used to aptly construct a definition. Blocked study, on the other hand, will lead to the creation of isolated representations. Although generally better for producing definitions, if learners cannot tell the concepts apart in the first place, stand-alone representations of each concept will not be helpful.

In Experiment 3 we further test this hypothesis. If interleaved study leads to the creation of interrelated representations, the usefulness of these representations should be tied to the specific concepts studied (as seen in Experiment 2). That is, interrelated representations of concepts A and B emphasize their differences, and thus are unlikely to help to discriminate A from C when those two concepts do not share the same differences. To test this, interleaved and blocked study were followed by a two-alternative forced choice test varying the distractor items. In some tests the distractor is from the other category studied in the same interleaved or blocked sequence, whereas other times it is an item from a category studied in the other schedule or a new category. The main prediction is that if blocked study leads to isolated representations, the nature of the distractor should not affect test performance following blocked study. Similarly, if interleaved study leads to interrelated representations it should be particularly beneficial for tests that contrast the two concepts studied interleaved (for which the representations are interrelated).

9 | METHOD

9.1 | Participants

A group of 39 people completed the experiment following recruitment through Amazon's Mechanical Turk (https://www.mturk.com/). Data from three participants were excluded due to self-reported previous participation in another study with the same materials. Data from an additional eight participants were excluded from analyses because of possible compliance issues (see below for details). The final sample included 28 participants. Table 1 includes the demographic characteristics of participants in the overall sample.

9.2 | Stimuli and procedure

In this experiment, we used the same set of materials as in the previous experiments, but concepts were randomly assigned to be studied interleaved or blocked.

The procedure was similar to the procedure used in Experiment 2 except for the following differences. After completing the study phase and playing Tetris for 30 s participants completed a two-alternative forced choice test. In this task participants were shown an example of a situation depicting a psychology concept and asked to classify it into one of two concepts by selecting the name of the correct concept from two options. Participants classified a total of 32 novel situations in the 2AFC task at test. Eight situations (25%) depicted an example of one of four novel concepts not studied (Novel trials), 12 (38%) depicted examples of one of the four

concepts studied blocked (Blocked Trials) and the remaining 12 (38%) were examples of one of the four concepts studied interleaved. In each trial two labels appeared under the example for classification. In the case of studied concepts, the two options could be the names of two concepts studied in the same schedule (same trials) or from different schedules (different trials). In the case of novel concepts, in addition to the novel label to match the novel example, the distractor could be the name of a concept studied blocked or studied interleaved.

10 | RESULTS AND DISCUSSION

We identified potentially non-compliant participants using the participants' response times during study. The sample's median response time to complete the study phase was 8.4 s per problem (max: 21.5 s/problem; min: 0.70 s/problem). The 10th and 90th percentiles for the distribution of median response times were 3.8 s/problem and 19 s/problem respectively. Eight participants were identified as outliers based on their falling outside of this range and their data were excluded from further analyses. In all the analyses presented below, mean pretest scores were included as covariates.

10.1 | Pretest

As in the previous experiments, participants showed little to no pre-training knowledge of the to-be-studied concepts (see Table 3).

10.2 | Study phase

Mean performance during the blocked study phase was 69% (SEM = 4.0%), while during interleaved study it was 65% (SEM = 4%). However, this difference was not statistically significant, $F(1, 25) = 2.44$, $p = .130$, $\eta^2_G = 0.01$.

10.3 | Test phase

We start by analyzing performance for trials concerning studied concepts. According to the hypothesis that interleaved study would lead to the creation of interrelated concepts, we would expect that performance would be better on trials on which both options were studied interleaved and therefore are represented in contrast to one another. As show in the left panel of Figure 6, although participants are overall better at discriminating concepts studied in different sequences, $F(2, 54) = 4.98$, $p = .010$, $\eta^2_G = 0.02$, there is also an interaction between the type of schedule and the type of test, $F(2, 54) = 4.43$, $p = .016$, $\eta^2_G = 0.02$. Participants were more accurate at classifying studied concepts following interleaved study compared to blocked study when both options were concepts studied in the same sequence, than when they were from different sequences, $F(2, 54) = 4.30$, $p = .017$, $\eta^2_G = 0.09$. Finally, for novel concepts, which cannot be part of an interrelated concept created during study, we would predict that classification would be better when the foil is a concept studied blocked, because its stand-alone representation would allow for better

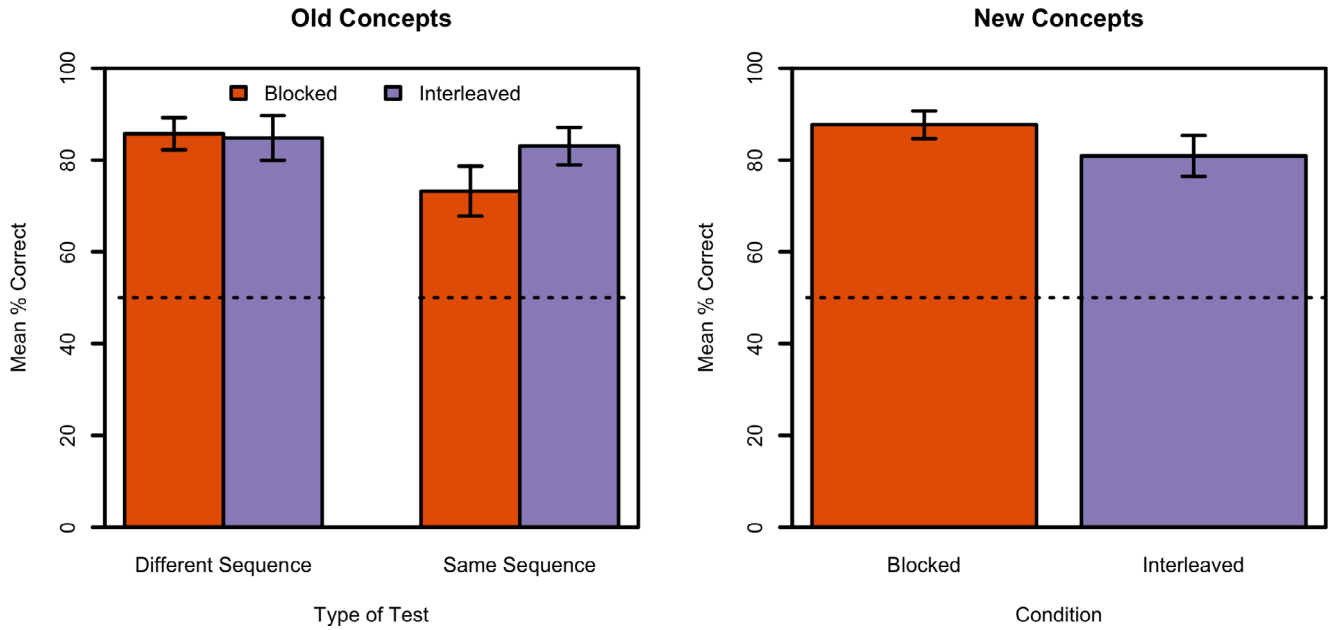

FIGURE 6 Results for the Test Phase (two-alternative forced choice; 2AFC) of Experiment 3. Results for tests involving studied concepts are presented on the left organized by whether both target and distractor of the 2AFC were presented in the same or different sequence. Results for test involving novel concepts not studied are presented on the right, organized by whether the foil option in the 2AFC was studied interleaved or blocked. The dotted line represents chance performance. Error bars represent standard errors of the mean [Colour figure can be viewed at wileyonlinelibrary.com]

comparison with a novel, never studied, concept. We found a marginally significant effect in that direction (see right panel of Figure 6), $F(2, 27) = 4.11$, $p = .052$, $\eta^2_G = 0.03$.

As a whole, Experiment 3 confirms the predictions of SAT and the results from Experiments 1 and 2: During blocked study learners create positive (focusing on features that are characteristic of the category and not on features that do not), stand-alone representations of the concepts studied which characterize each concept independently. Conversely, during interleaved study learners create interrelated representations that emphasize discriminating characteristics of each concept relative to the other simultaneously acquired concepts. Because of this we see that learners provide better definitions in tests following blocked study, but this effect is flipped when the stimuli studied are highly confusable. Furthermore, we see in Experiment 3 that the representations created during interleaved study do not contribute to discrimination of stimuli not studied interleaved, whereas blocked study provides a marginal benefit to classify novel concepts, providing further evidence for our hypothesis. To further support our hypothesis, we ran a follow up experiment that replicated Experiment 3 using a set of stimuli where all concepts are highly confusable. As predicted, we found an overall benefit of interleaved study for all tests (see Supporting Information).

11 | GENERAL DISCUSSION

The sequence of study has been shown to have a considerable impact on learning outcomes (e.g., Carvalho & Goldstone, 2014b; Kornell & Bjork, 2008). However, our theoretical understanding of how the sequence of study changes learning and our ability to use sequence of study to bolster learning outcomes is lacking. In our view, one limiting factor is the focus on understanding the benefits of one sequence over the other, instead of understanding how sequence shapes learning.

In this paper, we extended Carvalho and Goldstone's SAT model's proposal that the sequence of study changes attention and encoding during learning. SAT predicts that during interleaved study learners will attend and encode better differences between concepts, whereas during blocked study learners will attend and encode better similarities among items of the same concept. Following SAT, we proposed that, if indeed attention and encoding are changed during learning, then learners will end up with different representations of the concepts depending on whether interleaved or blocked study was used, despite having studied the same concepts.

Overall, the results presented here are consistent with our hypothesis. Studying examples of different concepts in a blocked sequence improves performance in a test requiring learners to define the concept studied, whereas for other tests there is no difference in performance between the two sequences of study, suggesting that blocked learners created isolated representations of the concepts that are more suitable for tests requiring the use of features of the concepts that do not emphasize differences. Furthermore, we found similar results in a second study when the concepts were unrelated, but the effect was reversed if the concepts were related and highly confusable. This finding provides further evidence that not only does blocked study lead to the creation of isolated representations, but interleaved study leads to the creation of interrelated representations that emphasize the differences between concepts. These differences will be useful even in a test emphasizing similarities when discrimination is challenging. We further confirmed this hypothesis in Experiment 3 by demonstrating that interleaved study improves performance only in tests involving concepts that were studied in the same sequence, but this benefit does not extend to unrelated concepts studied in a different sequence or novel concepts.

Our findings also extend the predictions of the transfer-appropriate framework beyond memory tasks: in a transfer task involving not only recollection of studied materials but also manipulation and use of this information, we saw that when encoding and test requirements overlapped, participants performance was improved compared to when there was less overlap.

Interestingly, across the first three experiments we did not see a benefit of interleaved study for multiple-choice tests. At first look this finding seems inconsistent with our predictions and previous findings showing that interleaved study improves discrimination between stimuli (which Multiple-Choice tests emphasize). However, upon closer scrutiny and the results of Experiment 3, an hypothesis becomes clear: because the stimuli selected did not require much discrimination, even in a task that emphasizes discrimination Interleaved study did not improve performance. In Experiments 1 and 1b hard-to-discriminate concepts were never presented in the same session, therefore discriminability pressures during training were low. In Experiment 2, although confusable concepts were studied together, because of the procedure employed, the multiple-choice options included mostly dissimilar concepts, making the task easier than if only hard-to-discriminate items had been studied. The results of Experiment 3 are consistent with this hypothesis: interleaved study improved test performance only when the test required discrimination between two concepts studied interleaved but not when one of the concepts had been studied in a different sequence or was a novel concept never studied before. Experiment 3b, included in Supporting Information provide further evidence: when using the same procedure as Experiment 3 but concepts that are all highly confusable–artist styles–we saw a benefit of interleaved study for all types of test.

Overall, the results presented here are consistent with SAT (Carvalho & Goldstone, 2017), but potentially challenging for other current theories of an interleaving benefit. First the results suggest that although interleaved study does indeed improve discrimination as seen in Experiments 2 and 3, it is very targeted to the stimuli studied and their characteristics. Discriminating two related concepts does not yield benefits in discrimination of those concepts to other concepts or novel concepts. Although overall consistent, this finding requires some refinement of theories proposing that interleaved study improves learning because it increases discrimination given that this effect is targeted to the concepts studied and does not generalize as previously implied. Second, the results suggest that blocked study can lead to improved *test* discrimination when comparing to a novel

concept (though this conclusion requires caution given the marginal effect and certainly requires further tests). This finding is also challenging for theories of sequencing effects focused on the relation between interleaved study and discrimination in multiple ways, one would not expect blocked study to ever improve learning as it does not involve any discrimination and one would expect that improved discrimination during study would also lead to improved discrimination during test. Finally, the results presented here further challenge proposals focusing only on the interleaving advantage. Instead, learning is affected by the sequence of study in a way that leads to different information being learned from different concepts. We observed reliable blocking advantages when tasks benefitted from independent characterizations of concepts.

Although the sample sizes used in the studies reported here might seem small, it is important to note that most critical comparisons were within-subject manipulations which increases the analytic power of the studies and that the effect sizes reported here are considerable and in line with previous similar research.

In sum, there are two main contributions of the present work. First, it goes beyond existing demonstrations that blocked study is better or worse than interleaved study by showing how the sequence of study affects *what* is learned by creating different representations given the same content. Second, it provides evidence for the context-dependent nature of learning and how the benefits of each sequence depend on the learning situation. This evidence adds to previous demonstrations that the best sequence of study depends on the type of material being studied (Carvalho & Goldstone, 2014a, 2014b; Patel, Liu, & Koedinger, 2015), the type of study task (Carvalho & Goldstone, 2015a; Rawson et al., 2015), learners' working memory capacity (Sana et al., 2016), and whether learners' actively decide how to organize their study (Carvalho et al., 2016). These results also show the importance of developing *theories* of *why* one intervention is better than another. We have proposed a theory based on the similarities of the materials being learned and the nature of the task. When concepts are similar to each other, learners prioritize learning discriminating features. Writing definitions generally benefits from standalone representations unless the concepts being defined are similar to each other and benefit by being contrasted. The study of how an intervention interacts with the learning situation, we would argue, has the potential to not only provide a fuller understanding of how learning takes place, but also provide richer, more precise, recommendations for practice (Jonassen, 1982).

The sequence in which examples are presented might often feel inconsequential for learning outcomes, but the increased interest in how interleaved and blocked practice affect learning has demonstrated the impact easy-to-implement and frequently-applicable interventions such as blocked study might have on learning (for better or worse). The results presented here further suggest that sequence of study can be used as an effective learning tool. Students and educators should take this importance into account when organizing studying materials and, we would argue, take the whole learning situation into account to decide how to best organize the materials: what is going to be studied, how is it going to be studied, and how is it going to be tested.

ACKNOWLEDGEMENTS

This work is part of a doctoral dissertation submitted by the first author to Indiana University as partial fulfilment of the degree of Doctor of Philosophy.

The authors would like to thank the members of the Percepts and Concepts Lab for discussion. Dustin Finch, Kaley Liang, Ashton Moody, Alifya Saify, and Shivani Vasudeva assisted with response coding. We are grateful to Katherine Rawson for sharing the stimuli set with us.

This work was supported in part by the National Science Foundation [grant #0910218 to RG and grant #1824257 to PC]; the Department of Education [IES grant # R305A1100060 to RG]; Portuguese Foundation for Science and Technology, co-sponsored by the European Social Fund [Graduate Training Fellowship grant # SFRH/BD/78083/2011 to PC].

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

All data reported in this manuscript is available online at the Open Science Initiative (https://osf.io/y4a2r/).

ORCID

Paulo F. Carvalho https://orcid.org/0000-0002-0449-3733

REFERENCES

Anbar, M. (1991). Comparing assessments of students' knowledge by computerized open-ended and multiple-choice tests. *Academic Medicine*, *66*(7), 420–422.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

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

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2015). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition*, *43*(2), 193–205. https://doi.org/10.3758/s13421-014-0462-6

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671–682.

Carvalho, P. F., Braithwaite, D. W., de Leeuw, J. R., Motz, B. A., & Goldstone, R. L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLoS One*, *11*(3), e0152115. https://doi.org/10.1371/journal.pone.0152115

Carvalho, P. F., & Goldstone, R. (2019, September 13). A computational model of context-dependent encodings during category learning. Retrieved from https://doi.org/10.31234/osf.io/8psa4

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, *5*, 936. https://doi.org/10.3389/fpsyg.2014.00936

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015a). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*(1), 281–288. https://doi.org/10.3758/s13423-014-0676-4

Carvalho, P. F., & Goldstone, R. L. (2015b). 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. (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*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Corneille, O., Goldstone, R. L., Queller, S., & Potter, T. (2006). Asymmetries in categorization, cerceptual discrimination, and visual search for reference and nonreference exemplars. *Memory & Cognition*, *34*(3), 556–567.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608–628.

Jonassen, D. H. (1982). Aptitude-versus content treatment interactions: Implications for instructional design. *Journal of Instructional Development*, *5*(4), 15–27.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

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. https://doi.org/10.1002/acp.1801

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory*, *10*(5-6), 397–403.

Patel, R., Liu, R., & Koedinger, K. (2015). When to block versus interleave practice? Evidence against teaching fraction addition before fraction multiplication. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.). Paper presented at: Proceedings of the 38th annual conference of the cognitive science society (pp. 2069–2074). Austin, TX: Cognitive Science Society.

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review*, *27*(3), 483–504. https://doi.org/10.1007/s10648-014-9273-3

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Sana, F., Yan, V. X., & Kim, J. A. (2016). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*. *109*(1):84–98.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Carvalho PF, Goldstone RL. The most efficient sequence of study depends on the type of test. *Appl Cognit Psychol*. 2021;35:82–97. https://doi.org/10.1002/acp.3740

APPENDIX: MATERIAL EXAMPLES

Concept	Definition	Example 1	Example 2
Availability heuristic	The tendency to estimate the likelihood that an event will occur by how easily instances of it come to mind.	Alan and Adam worked together on a project, each putting in about half of the amount of time and effort required to finish. At the end, they have to decide who should get the most credit. Alan and Adam both claim that they did a majority of the work, probably because it is easier for them to remember their own experiences working on the project.	After the columbine shootings and the extensive press coverage, people were more likely to overestimate the amount of teen violence and to fear school violence.
Mere exposure effect	The phenomenon whereby the more people are exposed to a stimulus, the more positively they evaluate that stimulus.	In one study, college students were told that they were participating in an experiment on how people learn foreign language. They were then shown Chinese-like characters for 2 s at a time, with the characters presented either once, twice, 5 times, 10 times, or 25 times. Participants then rated the characters, including some characters that they had not seen, on a good-bad scale. The results indicate that the more often a character was repeated, the more positively participants rated it.	Many people find that their liking of a piece of music, a work of art, or even a kind of food increases with repeated experiences with them, which is why we often call these preferences acquired tastes.
Door-in-the-face technique	A strategy to increase compliance based on the fact that refusal of a large request increases the likelihood of agreement with a subsequent smaller request.	Richard received a telephone call from a college alumni association asking him to show his loyalty by contributing $1,000. When he apologetically declined, the caller acted sympathetic and then asked whether he could contribute $500. And if not $500, how about $200? Richard agrees to donate $200.	Carrie needs about $10 for a shopping trip, so she asks her mother for $50, but her mother refuses. Carrie then asks for $10 and her mother agrees.
Foot-in-the-door technique	A strategy to increase compliance, based on the fact that agreement with a small request increases the likelihood of agreement with a subsequent larger request.	Imagine that you work for the local animal shelter. Your goal is to increase the number of people who are willing to adopt a dog from the shelter. To maximize your success you should first ask people if they would be willing to wear a button that says, “adopt a dog today.” A couple of weeks later, you should then ask these people to adopt a dog themselves.	Many cult leaders gain followers by having them make many small sacrifices one after the other, which eventually get bigger and bigger and bigger until the followers are willing to do anything–even take their own lives.