Does Interleaved Practice Enhance Foreign Language Learning?

The Effects of Training Schedule on Spanish Verb Conjugation Skills

Steven C. Pan[a]

Jarrett Lovelett[a]

Jahan Tajran[a]

Jessica Osuna[a]

Timothy C. Rickard[a]

[a]Department of Psychology,

University of California San Diego

This manuscript was accepted for publication in the *Journal of Educational Psychology* on November 7, 2018. This document may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final version is available at: http://dx.doi.org/10.1037/edu0000336

Author Note

Steven C. Pan, Department of Psychology, University of California San Diego; Jarrett Lovelett, Department of Psychology, Department of Psychology, University of California San Diego; Jahan Tajran, Department of Psychology, University of California San Diego; Jessica

Osuna, Department of Psychology, University of California San Diego; Timothy C. Rickard, Department of Psychology, University of California San Diego.

Steven C. Pan is now at the Department of Psychology, University of California Los Angeles; Jahan Tajran is now at Wayne State University School of Medicine; and Jessica Osuna is now at the Department of Psychiatry and Veterans Medical Research Foundation, University of California San Diego.

This research was supported by an American Psychological Association (APA) Early Graduate Student Researcher Award, a National Science Foundation (NSF) Graduate Research Fellowship, and a Psychonomic Society award to Steven C. Pan. The authors gratefully acknowledge Robert Bjork, Elizabeth Bjork, Sean Kang, Barbara Knowlton, Doug Rohrer, Veronica Yan, and CogFog attendees for helpful comments; Dina Rodgers for valuable consultations regarding subject pool management; Vicky Phun for assistance with checking Spanish textbooks; and Anastasia Bogozova, Jon Clausen, Dominic D'Andrea, Danielle Emmar, Kellie King, Courtney Lukitsch, Ikjot Thind, Thomas Ting, Daanish Unwalla, and other lab members for assistance with running the experiments.

Correspondence concerning this article should be addressed to Steven C. Pan, Department of Psychology, University of California Los Angeles, 2434 Franz Hall, Los Angeles, CA 90095. E-mail: stevencpan@ucsd.edu

Abstract

Do the cognitive benefits of *interleaving*—the method of alternating between two or more skills or concepts during training—extend to foreign language learning? In four experiments, we investigated the efficacy of interleaved vs. conventional blocked practice for teaching adult learners to conjugate Spanish verbs in the *preterite* and *imperfect* past tenses. In the first two experiments, training occurred during a single session and interleaving between tenses began during the presentation of introductory content (Experiment 1) or during randomly-ordered verb conjugation practice trials at the end of the training session (Experiment 2). This yielded, respectively, numerically higher performance in the blocked group and equivalent performance in the interleaved and blocked groups on a two-day delayed test. In Experiments 3 and 4, the amount of training was increased across two weekly sessions in which the blocked group trained on one tense per session and the interleaved group trained on both tenses per session, with random interleaving occurring during verb conjugation practice trials. Interleaving yielded substantially better performance on a one-week delayed test. Thus, although interleaving did not confer an advantage over blocking under two different single-session training schedules, it improved learning when used to practice conjugating verbs across multiple training sessions. These results constitute the first demonstration of an interleaving effect for foreign language learning.

Keywords: interleaving, interleaved practice, language learning, verb conjugation, Spanish tense

Educational Impact and Implications Statement

The current study examined whether interleaving, a learning technique which involves alternating between two or more skills or concepts during training, improves foreign language learning. In many foreign language courses, interleaving is rarely used; rather, one-skill-at-a-time blocked practice (blocking) is more common. Across four experiments, college students used interleaving or blocking to learn how to conjugate verbs in the Spanish preterite and imperfect past tenses. Interleaving yielded better verb conjugation skills than blocking when it was used to practice those skills across multiple training sessions. These results suggest that interleaving can be beneficial for foreign language learning.

Does Interleaved Practice Enhance Foreign Language Learning?

The Effects of Training Schedule on Spanish Verb Conjugation Skills

Which is more effective: learning one skill or concept at a time, or learning multiple related skills or concepts concurrently? In conventional educational practice, the former method—also called *blocked practice* (or *blocking*)—is frequently used due to its seemingly obvious efficacy and ease of scheduling. However, a growing body of research has shown that the latter method—also called *interleaved practice* (or *interleaving*)—can have surprising benefits over blocking (Battig, 1972; Carpenter, 2014; Kornell & Bjork, 2008; for reviews see Kang, 2017; Rohrer, 2012). With interleaving, students alternate between a set of to-be-learned skills during training. For instance, if the goal is to learn to calculate the volume of cylinders, spheres, and cones, then interleaving may involve practicing with a problem involving a cylinder, then a problem involving a sphere, then a problem involving a cone, then a problem involving a cylinder, and so on (e.g., Rohrer & Taylor, 2007). By contrast, blocking involves practicing on an entire set of problems involving cylinders, then a set of problems involving spheres, and then a set involving cones. Interleaving tends to be more difficult and often yields lower performance during training than blocking. However, it can generate better long-term memory—an advantage called the *interleaving effect*—as evidenced by higher accuracy on a subsequent test featuring either novel problems requiring the trained skills or the same problems but with new numerical values (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Kang, 2017; Soderstrom & Bjork, 2015; Yan, Soderstrom, Seneviratna, Bjork, & Bjork, 2017).

The interleaving effect has been repeatedly demonstrated for motor skill learning (e.g., Goode & Magill, 1986; Hall, Domingues, & Cavazos, 1994; Shea & Morgan, 1979; for reviews see Brady, 1998; Magill & Hall, 1990), inductive visual category learning (e.g., Hatala, Brooks, & Norman, 2003; Kornell & Bjork, 2008; Vlach, Sandhofer, & Kornell, 2008; Wahlheim,

Dunlosky, & Jacoby, 2011), and mathematics learning (e.g., Rohrer, Dedrick, & Burgess, 2014; Rohrer, Dedrick, & Stershic, 2015; Taylor & Rohrer, 2010). Based on those results, many cognitive scientists have highlighted interleaving as a highly promising method for improving education and training (e.g., Carpenter, 2014; Brown, Roediger, & McDaniel, 2014; Kang, 2017; Roediger & Pyc, 2012; Schmidt & Bjork, 1992). However, some researchers have called for more research on interleaving with new types of tasks (e.g., Rohrer, 2012) and flagged studies showing null or even detrimental effects of interleaving (e.g., Dunlosky et al., 2013; Pan, 2015).

One notable area in which interleaving has largely failed to demonstrate robust benefits is second language (L2) learning. For instance, Schneider, Healy, and Bourne (1998, 2002) had college students learn French-English word translations using interleaving or blocking. In Schneider et al. (2002; Experiment 1), students in the blocked condition, who studied translations grouped by semantic category (e.g., tableware, foods, etc.), performed better on an immediate test than did students in the interleaved condition, who studied translations in random order. Retention of learning in the two conditions was equivalent, however, on a one-week delayed test. In another example, Carpenter and Mueller (2013) had college students learn French pronunciation rules using either interleaving, where different rules were represented on successive practice trials, or blocking, where practice trials were grouped by rule. Across four experiments involving high vs. low amounts of training, implicit vs. explicit instructions, and easy vs. difficult tests, a blocking advantage for correct word pronunciation was consistently observed on immediate or 5-min delayed tests. Although the materials in these studies are far from the only skills that L2 learners must master, the results suggest limitations of interleaving and invite further research into when the technique is beneficial. We addressed that issue in this manuscript by exploring interleaving's efficacy for the promotion of grammar learning, and specifically for foreign language verb conjugation skills.

Process Accounts of the Interleaving Effect

Two prominent accounts of the interleaving effect, namely spaced practice and the discriminative contrast hypothesis, suggest circumstances under which interleaving benefits will be observed.

The spacing account. The earliest hypothesis of the interleaving effect posits that it is solely a *spacing effect*—i.e., the finding that, given the same overall duration of practice, temporally *distributed practice* results in better long-term retention than does temporally *massed practice* (Carpenter, 2014; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1996; Ebbinghaus, 1885). Interleaving necessarily incorporates spacing due to the fact that successive trials on a specific skill or concept are separated in time by intervening trials on other skills or concepts (e.g., given to-be-learned concepts A, B, and C, an interleaved schedule may be ABCABCABC, such that there are two trials in between successive exposures to the same concept). According to spacing-based accounts of the interleaving effect, the same cognitive mechanisms that underlie the spacing effect, such as study-phase retrieval processes or encoding variability (Benjamin & Tullis, 2010; Cepeda et al., 2006; Dempster, 1996), may also underlie the interleaving effect. However, it should be noted that evidence is mixed for the efficacy of spacing for foreign language learning (Bird, 2010; Lapkin, Hart, & Harley, 1998; Lightbown & Spada, 1994; Serrano & Muñoz, 2007; Suzuki & DeKeyser, 2017), due perhaps to the varied learning tasks investigated to date and to the limited number of studies (for a review of the applicability of spacing and testing effects to L2 learning, see Ullman & Lovelett, 2018).

The discriminative contrast hypothesis. This hypothesis posits that the interleaving effect is due to the juxtaposition of items from different categories on successive trials (Kang & Pashler, 2012). As such, it predicts that interleaving's benefits are likeliest when categories have high *between-category similarity* (i.e., Birnbaum, Kornell, Bjork, & Bjork, 2013; Rohrer, 2012;

Sana, Yan, & Kim, 2017). For example, the *simple past* and the *present perfect* tenses in English both refer to relatively subtle differences in past actions that can be difficult to discriminate between (e.g., "I went to the store yesterday" vs. "I have gone to the store many times"). By comparison, the *simple past* and *simple future* grammatical tenses refer to past and future events, respectively, and should be easier to tell apart (e.g., "I went to the store yesterday" vs. "I will go to the store tomorrow"). According to the discriminative contrast hypothesis, interleaving should be especially beneficial for learning in the former case.

Supporting evidence for the discriminative contrast hypothesis stems from studies of visual category learning in which the degree of between-category similarity has been manipulated (e.g., Carvalho & Goldstone, 2014; see also Zulkiply & Burt, 2013). When between-category similarity is high, an interleaving effect is typically obtained, and when it is low, it is not (and in fact a blocking advantage is often observed, e.g., Carvalho & Goldstone, 2014; Kurtz & Hovland, 1956; Zulkiply & Burt, 2013). Thus, for the case of grammatical tenses that are easily confused with one another (which is a property of the tenses that were examined in the current research), the discriminative contrast hypothesis predicts that an interleaving advantage should be observed on a delayed test.

To differentiate between the discriminative contrast and spacing accounts, Kang and Pashler (2012) as well as Birnbaum et al. (2013) investigated interleaving for visual category learning in which there was (a) interleaving between items on successive, contiguous trials vs. (b) interleaving between items but with additional spacing between trials (where irrelevant materials, such as cartoons or trivia questions, were shown). Both found that the interleaving effect was eliminated when additional spacing was introduced (which by the spacing account should have enhanced the effect), suggesting that discriminative contrast is most likely to occur on successive trials that are in close temporal proximity, and that, in at least some contexts, it is

the critical factor underlying the interleaving effect (see also Taylor & Rohrer, 2010; Zulkiply & Burt, 2013).

When and how much interleaving should be used. The point at which interleaving is introduced during training may also impact its efficacy. Most interleaving studies incorporate the technique throughout the entire training session (e.g., Kornell & Bjork, 2008; Sana et al., 2017). However, some researchers have hypothesized that providing a certain amount of blocked practice prior to interleaving may yield even better learning (Carpenter & Mueller, 2013; Dunlosky et al., 2013; Rohrer, 2012). That early *blocking* may aid initial learning of a series of to-be-interleaved topics. Indeed, in a recent study, the use of interleaving only after a specified amount of blocking—a form of *hybrid* blocked-to-interleaved training schedule—yielded better learning of verbal categories (i.e., lists of words grouped by invented category names) than did interleaving from the beginning of the training session (Sorensen & Woltz, 2016). That finding led the authors to hypothesize that for some learning tasks, and particularly those involving explicit rule learning, interleaving throughout training disrupts the cognitive processes that are necessary to develop a complete understanding of the categories being learned (e.g., working memory, attention, hypothesis-testing). There is also evidence from the motor skills literature that transitioning from initial blocked to subsequent interleaved practice can be helpful (e.g., Porter & Magill, 2010; for further discussion, see Kang, 2017). In the present study, the interleaved group in each experiment learned at least some introductory materials in blocked fashion prior to the onset of interleaving.

Learning Spanish Verb Conjugation Skills

We investigated the effects of interleaving for the acquisition of *verb conjugation skills*—i.e., the modification of root verbs to reflect tense and other syntactic properties. Developing the ability to conjugate verbs is one crucial step in learning to speak and understand a second

language. We used the world's second most widely-spoken native language, Spanish, which more than 21 million students study as a second language annually (Fernández & Roth, 2013; Fernández-Vítores, 2015). Spanish can be especially difficult for native English speakers because of differences in the way that grammatical tense is represented in that language relative to English. Specifically, Spanish relies on verb suffixes and grammar rules that in many cases have no clear analogues in English (Castañeda, 2011; Frantzen, 1995).

Verb conjugation in English vs. Spanish. In English, conjugated verb forms reflect tense but often ostensibly ignore person (e.g. first-person, third-person) or number (singular, plural). All three characteristics are explicitly marked as part of the verb itself in Spanish conjugation. Consider the English verb "to use." In English, there is one *simple past* tense form of that verb (i.e., "used") and it is always used regardless of the subject of the sentence. In contrast, there are at least six past tense forms of the equivalent Spanish root verb "usar" (to use); these vary from "usaba" (I used) to "usaron" (they used) depending on grammatical features of the sentence and the relationship of the past event to other events and/or to the present. When conjugating Spanish verbs, each of those characteristics must be attended to. For the beginning learner, that may yield a three-step process (see Figure 1): identify grammatical tense, identify the subject (i.e., pronoun), and then recall and use the correct suffix to conjugate the root verb.

The challenge of the *preterite* and *imperfect* tenses. Conjugating Spanish verbs in two particular grammatical tenses—the *preterite* and *imperfect* past tenses (or more formally, *aspects*)—is an especially difficult skill for many Spanish L2 learners to master (Castañeda, 2011). Broadly, the preterite tense refers to temporally specific past events, whereas the imperfect tense refers to temporally ambiguous past events. There are also other defining characteristics (see Table 1 for a list of rules; for further details see Frantzen, 1995; Iguina & Dozier, 2008; Westfall & Foerster, 1996). The difficulty lies in the potential for considerable

confusion between the two tenses—i.e., high between-category similarity—as evidenced by sentences that, in the absence of close inspection or sufficient Spanish experience, appear to maintain both their meaning and their grammaticality when expressed in either tense (but actually do not).

In current educational practice, the preterite and imperfect tenses are often learned using blocked training. Our examination of 25 common Spanish textbooks found that the two tenses are usually segregated into separate and non-adjacent chapters (e.g., Nissenberg, 2013), separate but adjacent chapters (e.g., Goodall & Lear, 2017), or separate sections within the same chapter (e.g., Blanco & Colbert, 2009). In nearly all cases, each tense is learned separately (although some books include "preterite vs. imperfect" subsections at the end of a chapter or in later chapters). The lone exception, Iguina and Dozier (2008), introduced both tenses in parallel and emphasized the need to distinguish between the two throughout. Spanish instructional guides also recommend introducing both tenses separately (e.g., Westfall & Foerster, 1996).

The Current Experiments

The primary question addressed in the present research was: (a) Does interleaving benefit the learning of Spanish verb conjugation skills among English speakers, and specifically for the preterite and imperfect tenses? In each of four experiments, after interleaved or blocked practice, retention of verb conjugation skill was measured via a delayed test wherein participants had to conjugate verbs in both tenses. That delayed test also enabled us to examine two related questions: (b) Does the manner in which interleaving is integrated into training affect the acquisition of verb conjugation skills?; and more specifically, (c) Is there is an interleaving benefit for verb conjugation skills when training takes place across more than one weekly session, as is common in language courses?

Across the experiments, we investigated the relative benefits of interleaving under single

(Experiments 1-2) vs. multisession (Experiments 3-4) training conditions, the latter being relatively rare in the current interleaving literature, and in cases where the introduction of interleaving occurred relatively early during training (Experiment 1) vs. later (Experiments 2-4). Thus, these experiments explored several implementations of interleaved practice for learning verb conjugation skills. The literature makes differing predictions as to whether interleaving may improve learning in the current research; the spacing and discriminative contrast accounts generally imply that a benefit will be observed, whereas prior studies showing limits of interleaving for foreign language materials (e.g., Carpenter & Mueller, 2013) and for materials in which explicit learning is involved (e.g., Sorensen & Woltz, 2016) suggest otherwise. It should be noted, however, that verb conjugation skills are more complex than the materials that have been used in prior studies of interleaving and L2 learning (e.g., vocabulary words) and differ from the category-learning materials that comprise much of the interleaving literature.

Design. In each experiment, students with no prior Spanish language experience were randomly assigned to an *interleaved* group or a *blocked* group. In Experiments 1 and 2, all training (interleaved or blocked) occurred within a single session and was followed by the delayed test 48 hr later. The primary difference between those experiments was the manner in which interleaving was implemented (e.g., when it was introduced during training and how it occurred at the trial level). In Experiments 3 and 4, we extended both the training process and retention interval: training occurred across two sessions in consecutive weeks, followed by the delayed test one week later. The only design difference between those experiments was whether a short answer or multiple-choice format was used for the delayed test.

The dependent measure in each experiment was delayed test performance in terms of proportion correct over all test items.

Overview of procedure. Each tense was trained across three phases that were derived

from Spanish language textbooks: *tense rules* (Phase 1), *suffixes* (Phase 2), and *verb conjugation practice* (Phase 3). For each tense, the following occurred:

Phase 1 involved learning the four defining rules for the tense (Table 1). After those rules were presented, participants completed a series of practice trials in which they determined whether an English sentence was an example of that tense or not (on the basis of those rules; see Table 2 for examples).

Phase 2 involved learning the suffixes that are to be used to conjugate verbs for different pronouns in the tense (Table 1). Three suffixes were learned per tense (one corresponding to each of three pronouns: "I", "you", and "we").[1] Each of those suffixes was appropriate for conjugating Spanish root verbs that had the common "-ar" ending, such as "hablar" (to speak). Participants completed one practice trial per suffix. That trial involved appending the correct suffix to a given root verb (see Table 2 for examples). Hence, across tense and suffix, there were six possible correct answers (i.e., 2 tenses x 3 suffixes per tense).

Phase 3 involved participants practicing what they had learned by conjugating Spanish "-ar" root verbs into new Spanish fill-in-the-blank sentences (see Table 2 for examples).

Whether all three phases occurred in succession for one tense, or occurred in a manner which alternated between tenses, depended on training group assignment (i.e., interleaved or blocked). After participants completed all three phases for both tenses, they provided a metacognitive judgment of difficulty (e.g., "How easy was it to learn Spanish verb conjugation?") and/or learning.

The delayed test resembled Phase 3 of training and involved participants conjugating new Spanish "-ar" root verbs into new Spanish fill-in-the-blank sentences in either multiple-choice (Experiments 1-3) or short answer (Experiment 4) format. Delayed test questions were presented in random order. This method has ecological validity given that speakers regularly use multiple

tenses within a single conversation and many exams do not block questions by topic, although training usually involves blocking.

Experiment 1

The first experiment was our initial attempt to investigate whether interleaving or blocking yields better learning when both the preterite and imperfect tenses are learned in a single training session. In this experiment, interleaving between tenses began relatively early during that session (i.e., with the learning of suffixes in Phase 2). This training schedule resembled that in the research literature for other task domains, in which interleaving occurs during a single training session and is used for all trials throughout the majority or the entirety of that session (e.g., Carpenter & Mueller, 2013).

Methods

Participants. In this and all subsequent experiments, undergraduate students recruited from a large U.S. research university participated in exchange for course credit. Students could participate only if they had no prior Spanish language experience or instruction and no family members who speak the language. They were also required to be fluent English speakers. The entire study was conducted with the approval of the university's Institutional Review Board.

The target sample size in this and subsequent experiments was determined using a priori power analysis. Based on the standard deviations of the test scores in the interleaved and blocked conditions of Carpenter and Mueller (2013; Experiment 4, between-participants), a sample size of 42 per group is needed to achieve power of 0.80 to detect a mean proportion correct difference of 0.05 or greater (based on a two-tailed, two-sample *t* test, $\alpha = .05$). Ninety-four participants (47 in each group) participated in Experiment 1. All but eight completed both sessions of the experiment, leaving 86 participants (*interleaved* group, $n = 44$; *blocked* group, $n = 42$) that were included in the data analyses.

Across this and the subsequent experiments, participant mean ages ranged from 20.1 to 20.9 yrs., with an overall range of 17-53 yrs. Most were female (68-74%). Ethnic and/or racial composition was approximately 82% Asian/Asian-American, 13% Caucasian, and 6% African-American or of other groups. That composition differs from that of the university's student body and was due to our language experience exclusionary criteria. All participants were fluent in English (37-51% natively).[2] Of non-native English speakers, the most common native language was Mandarin Chinese (69%), followed by Korean (11%) and various other languages ($\leq$ 5%). Demographic and language characteristics were similar across all experiments.

Materials. To facilitate learning among participants with no prior Spanish experience and to maintain consistency, all Spanish language materials were presented with accompanying English language translations, without diacritical marks (accent marks), and in some cases with simplified translations (i.e., some pronoun modifiers and/or prepositions were omitted). The linguistic accuracy of all materials, in the context of intentional deviations from conventional Spanish as just noted (including instances of further simplified translations; for details and examples of training materials, see Table 2) was independently verified by two of the authors with fluent or native Spanish language ability.

For Phase 1, twelve English sentences were created to serve as examples for each tense (three per rule). Eight additional sentences were constructed for use as practice trials for each tense. For Phase 2, an example sentence and a fill-in-the-blank practice question was created for each of the three suffixes per tense. These sentences were written in English, excepting a Spanish root verb. For Phase 3, nine fill-in-the blank practice questions in the same format as those in Phase 2 were created for each tense. The nine questions were comprised of three questions each for the "I", "you," and "we" pronouns, each involving a different root verb.

For the delayed test, 30 multiple-choice questions were developed (see Table 2). Each

question consisted of three parts: (a) a fill-in-the blank sentence that was written entirely in Spanish, (b) a to-be-conjugated Spanish root verb with an "-ar" ending, and (c) the English translation of the complete sentence. Root verbs were not repeated across questions. There were six answer choices for each question (corresponding to the six suffixes that were presented during training). Eighteen questions involved sentences in the preterite tense and 12 questions involved sentences in the imperfect tense; of these, each pronoun-tense combination and each of the four preterite and three imperfect rules was represented on at least three questions.[3]

Procedure. Participants completed the training and delayed test sessions at their own pace and at individual computer workstations. In both groups, training on either tense was prefaced by a series of introductory slides that provided a general overview of the Spanish verb conjugation process. After those slides, formal training began.

Training. The training schedules for both groups are depicted in Figure 2 (panels a and b). Participants in the blocked group completed Phases 1-3 for one tense before completing Phases 1-3 for the other tense. Thus, one tense was entirely learned before the other. In contrast, participants in the interleaved group completed Phase 1 for one tense followed by Phase 1 for the other tense, and then completed Phases 2 and 3 in interleaved fashion (alternating between tense within each phase). Thus, in the interleaved group, after an initial introduction to tense rules that occurred separately for each tense, participants learned and trained on both in a manner which alternated between tenses.

In both groups, the tense being learned was always identified at the top of the screen during Phases 1 and 2 (e.g., "How to conjugate verbs in the preterite tense"). During Phase 3, an introductory slide referred to the tense(s) that had just been learned (e.g., "You will now practice conjugating verbs in the tense that you just learned").

The training procedure in each group was as follows:

Blocked group. Phase 1 for the blocked group involved viewing each of four rules for a given tense, with examples, one at a time and on a single slide each (see Table 1). A summary slide featuring all four rules was then presented, followed by two cycles of eight randomly-ordered practice trials (i.e., eight trials were attempted twice across two cycles). On each trial, participants had to indicate whether or not the presented sentence reflected the tense that they had just learned (by typing Y or N). They were then given immediate correct answer feedback which (a) indicated whether the sentence matched the tense in question and (b) contained a statement of the rule that most closely applied to that sentence or a statement that none of the rules for that tense applied.

Phase 2 for the blocked group involved learning the verb suffixes for the tense introduced in the preceding phase. The suffixes to be used with the "I", "you", and "we" pronouns were learned in that order. Each was presented using two steps. First, a slide in which the pronoun and its respective suffix, as well as English translations and examples, was presented. Next, participants practiced applying that suffix to a root verb by typing its conjugated form into a fill-in-the-blank sentence, followed by immediate correct answer feedback.

Phase 3 for the blocked group began with a summary slide that reiterated the suffixes that had just been learned for a single tense. Nine practice trials, all involving that tense, followed. In each, participants attempted to modify a given root verb with the proper suffix to complete a fill-in-the-blank sentence. Correct answer feedback including the correctly conjugated verb, correct suffix, tense name, and relevant pronoun was provided on each trial. All practice trials were presented in a fixed order (which was unique to Experiment 1) wherein three consecutive trials each involved the "I", "you", and "we" pronouns and a different root verb was used on each consecutive trial.

Once Phases 1-3 were completed for a given tense, the same procedure was repeated for

the other tense. Afterwards, participants provided a metacognitive judgment of difficulty ("In the activities that you just experienced, how easy was it to learn Spanish verb conjugation?" on a five-point scale) and were dismissed.

Interleaved group. Phase 1 for the interleaved group was identical to the blocked group except that Phase 1 for one tense was immediately followed by Phase 1 for the other tense. Phase 2 was also largely identical except that all six suffixes from the two tenses were learned in the following order: "I" (preterite), "I" (imperfect), "you" (preterite), "you" (imperfect), "we" (preterite), and "we" (imperfect), with counterbalancing of the tense that was presented first in that sequence. That pattern, which was unique to Experiment 1, maintained the "I'-"you"-"we" order used in the blocked group but with the addition of alternation between tenses.

Phase 3 in the interleaved group also resembled that in the blocked group but with the following exceptions: (a) the phase began with two summary slides, one per tense; (b) the instructions stated that the practice trials would involve both tenses; and (c) 18 practice trials were presented consecutively (i.e., 9 per tense x 2 tenses) using the same general "I"-"you"-"we" pronoun order as in Phase 2 but with the tense changing every 3 trials and the pronoun changing every 6 trials (e.g., 3 "I"-preterite trials, then 3 "I"-imperfect trials, then 3 "You"-preterite trials, then 3 "You"-imperfect trials, and so on). This pattern, which was also unique to Experiment 1, maintained a consistent rate of switching between tenses and is comparable to the fixed patterns used in several prior interleaving studies (e.g., Kang & Pashler, 2012; Sana et al., 2017; Taylor & Rohrer, 2010), albeit with a more complex nested structure. Once participants completed Phase 3, they provided the same metacognitive judgment as in the blocked group and were dismissed.

Delayed test. Two days after training, participants returned to the laboratory for a delayed test that was identical for both groups and involved verb conjugation in the preterite and imperfect tenses. Thirty multiple-choice questions were presented in random order determined

anew for each participant. On each question, participants were presented with a Spanish fill-in-the-blank question, its English translation, a Spanish root verb, and six possible answer options. They had unlimited time to select one of those answer options. No feedback was provided.

Delayed test measures. As previously described, the dependent variable was delayed test proportion correct over all 30 delayed test questions.

Analysis plan. To analyze the delayed test results, we performed independent-samples *t*-tests on the factor of Training Group. No formal analyses were performed on training data except for (a) a chi-square test on metacognitive judgment data and (b) an analysis of a possible interaction between Phase 3 training performance and delayed test results. The same analysis plan was used in the subsequent experiments.

Results

Training. We performed exploratory analyses on the training data solely to examine patterns of performance during each phase. Descriptive statistics (mean proportion correct and *SE*) for each phase are presented for all experiments in Table 3. In Experiment 1, Phase 1 performance was comparable across groups, as expected given that Phase 1 training was blocked for both groups. However, both Phase 2 and Phase 3 performance were better in the blocked group. Analogous findings of poorer training phase performance in the interleaved group are common in the literature (e.g., Rohrer & Taylor, 2007; Taylor & Rohrer, 2010), including in cases where an interleaving effect is ultimately observed on a delayed test. In terms of total training duration, the blocked and interleaved groups were highly similar (mean and *SE*) at 10.29 (0.49) and 10.89 (0.48) min, respectively.

In a supplementary analysis we examined the potential influence of the fixed practice trial pattern in Phase 3, where the three consecutive trials per pronoun-suffix combination could have facilitated a working memory-based strategy of participants reusing suffixes on successive

trials. A visual evaluation of Phase 3 data revealed that mean accuracy improved from trial 1 to trial 2 of each three-trial sequence, was relatively stable from trial 2 to trial 3 of the same sequence, and dropped from trial 3 to trial 1 of the next sequence, in both groups. Although the potential use of that strategy ceased to be effective on every third trial, that pattern prompted our use of randomly-ordered Phase 3 trials in the remaining experiments.

With regard to metacognitive judgments of difficulty, a X^2 test for independence on participant ratings (difficulty on a scale of 1-5) and group was significant, $X^2(4) = 22.55$, $p < .0001$ (Table 4). Not surprisingly based on the literature, participants in the interleaved group were more likely to assign greater difficulty ratings to their training experience. Metacognitive results from all experiments will be considered further in the General Discussion.

Delayed test. Mean verb conjugation accuracy was 0.58 (SE = 0.033) in the blocked group and 0.48 (SE = 0.040) in the interleaved group, $t(84) = 1.93$, $p = .057$, $d = 0.42$ (Figure 3). It thus appears that, contrary to expectation, but consistent with some prior results involving foreign language learning (e.g., Carpenter & Mueller, 2013; Schneider et al., 1998, 2002), interleaving is not universally superior to blocking for the learning of verb conjugation skills. In fact, under some circumstances—and at least with the single-session training design used in this experiment—blocking may be as effective if not better.

Despite the finding of no interleaving benefit, to explore a possible interaction between performance on verb conjugation practice trials (Phase 3) vs. the delayed test, we performed a mixed-factors Analysis of Variance (ANOVA) with factors of Group (blocked vs. interleaved) and Session (Phase 3 vs. delayed test). That analysis revealed a significant Group x Session interaction, $F(1,84) = 10.01$, $p = .0022$, $MSE = 0.24$, $\eta_p^2 = 0.11$, reflecting the fact that the large blocked group performance advantage during Phase 3 was attenuated, but not entirely eliminated, on the delayed test.

The Kuder-Richardson (KR-20) reliability of the delayed test results (i.e., the internal consistency of the scores obtained with the sample that was used) was 0.86 and 0.92 for the blocked and interleaved groups, respectively.

Discussion

In other contexts involving interleaving, worse performance during training can co-occur with better performance on a delayed test (e.g. Bjork & Bjork, 2011; Rickard, Lau, & Pashler, 2008), suggesting a dissociation between factors that affect immediate performance and those that contribute to learning that survives after a delay. However, in Experiment 1 it appears that the poorer performance in the interleaved group reflected genuinely impaired learning during those phases that reduced delayed test performance. That impaired learning may have been due to the difficulty of switching between tenses (in Phases 2 and 3), the greater number of suffixes (i.e., six, rather than three) that were sequentially learned (in Phase 2), and the need to apply a greater number of suffixes (in Phase 3) in the interleaved group. Alternatively, if participants perceived the fixed trial patterns in Phase 3 and adopted a strategy of reusing suffixes across trials, then that may have attenuated benefits of interleaving.

Overall, Experiment 1's results suggest that the advantage of interleaving that has been observed in other task domains does not necessarily generalize to verb conjugation. These results, along with those of several other studies noted earlier, also raise the possibility that L2 learning generally does not benefit from interleaving. However, an alternative hypothesis is that other implementations of interleaving may yield different results. For example, interleaving did not have to begin while fundamental knowledge of both tenses was still being learned, nor did it have to involve a fixed trial pattern. In other studies for which interleaving effects have been observed, there is either no major foundational content to master (e.g., painting styles) or there is pretraining in the form of lessons that occur prior to interleaved practice trials (e.g., math

problems). Moreover, interleaving effects have sometimes been attributed to the unpredictability of training that occurs during random practice trials (e.g., Bjork, 1999), which this experiment lacked. Accordingly, in the subsequent experiments, we withheld the use of interleaving until the onset of verb conjugation practice in Phase 3, plus dropped the fixed ordering of practice trials in favor of full randomization.

Experiment 2

The second experiment was designed to investigate whether a different implementation of interleaving within a single training session would increase the technique's competitiveness relative to blocking. Specifically, the interleaved group did not experience any interleaving between tenses until Phase 3 (i.e., after all foundational content has been covered), and all practice trials were fully randomized.

Methods

Participants. Ninety-four undergraduate students, recruited in the same manner as in the preceding experiment, participated for course credit. All but nine participants successfully completed both sessions of the experiment (*interleaved* group, $n = 44$; *blocked* group, $n = 41$).

Materials. Materials were identical to those of the preceding experiment except that the summary slides for Phase 3 not only displayed the suffixes for the tense being practiced, but also displayed the rules for that tense.

Procedure. The procedure was modified from the preceding experiment as follows.

Training. Training schedules for both groups are depicted in Figure 2 (panels a and c).

Blocked group. The procedure for the blocked group was identical to the prior experiment except for three changes. First, each informational slide (i.e., rules or suffixes) was programmed to display for a minimum of 12 s before advancing was allowed. This helped ensure that participants read all content (in Experiment 1, our experimenters observed that some

participants may have rushed through several slides, although the same performance patterns were evident among those with the shortest reading times). Second, Phase 3 practice trials were randomized to preclude the aforementioned working memory-based response strategy that may have been used by some participants in each group in Experiment 1 (random practice trial ordering was implemented throughout all subsequent experiments). Third, participants were asked to provide an additional metacognitive judgment of learning ("How well did you learn Spanish verb conjugation today?" on a five-point scale) in addition to the judgment of difficulty.

Interleaved group. The interleaved training schedule was changed such that Phases 1 and 2 for both tenses were completed in blocked fashion before interleaving between tenses began in Phase 3. Consequently, all foundational materials (i.e., rules and suffixes for each tense) were learned before trial-level interleaving occurred. A minimum 12 s informational slide duration, random Phase 3 practice trial ordering, and a second metacognitive judgment question were implemented, just as in the blocked group. The random trial ordering, which was used in the interleaved group from this experiment onwards, prevented participants from being able to predict the tense or pronoun of any Phase 3 practice trial.

Delayed test. Forty-eight hr after training, participants completed a delayed test that was identical to that used in the preceding experiment.

Results

Training. Phase 1 performance was largely equivalent across both groups, just as in Experiment 1 (Table 3). However, unlike the preceding experiment, Phase 2 performance was also essentially equivalent between groups. That result is expected given that blocked training occurred in both groups. Phase 3 training performance, wherein blocked vs. interleaved training was first implemented, was again higher in the blocked group. Total training durations were again highly similar at (mean and *SE*) 16.99 (0.36) and 16.98 (0.39) min in the blocked and

interleaved groups, respectively.

As in Experiment 1, participants in the interleaved group assigned higher judgments of difficulty (Table 4) to their training experience, $X^2(4) = 10.92$, $p = .027$. Judgments of learning did not differ between training groups, $X^2(4) = 5.26$, $p = .26$.

Delayed test. Mean verb conjugation accuracy on the delayed test was nearly equivalent at 0.71 (*SE* = 0.039) in the blocked group and 0.73 (*SE* = 0.032) in the interleaved group, $t(83) = 0.51$, $p = .61$, $d = 0.11$ (Figure 3). An ANOVA analogous to that performed for Experiment 1 revealed a highly significant Group x Session interaction, $F(1,83) = 41.79$, $p < .0001$, $MSE = 0.75$, $\eta_p^2 = 0.33$, reflecting the fact that the large blocked group performance advantage during Phase 3 was attenuated on the delayed test. The reliability of the delayed test was 0.92 and 0.89 for the blocked and interleaved groups, respectively.

Discussion

In this experiment we observed that a single-session training schedule wherein interleaving did not occur until after foundational materials had been learned, and in which practice trials were fully randomized, still did not yield an interleaving advantage on a delayed test. However, unlike Experiment 1, delayed test performance was equivalent between groups. The parity between the blocked and interleaved groups on the delayed test raises the possibility that the elimination of interleaving in Phases 1 and 2 may have yielded better retention of learning in that group than in Experiment 1, a possibility that will we return to in the General Discussion.

Yet the apparently improved interleaved group performance did not translate into an interleaving advantage. Beyond the possibility that interleaving is not advantageous under any circumstances in this task domain, we considered two accounts of that result. First, it may be that there was an insufficient number of Phase 3 training trials to yield an interleaving advantage.

It is possible that the benefits of interleaving on retention become more robust with an increased amount of training and (or) at a higher level of achieved performance (for related findings, see Shea, Kohl, & Indermill, 1990). Second, it may be that a comparison of interleaving vs. blocking wherein two tenses are learned in a single session enables both groups to engage in discriminative contrast to varying degrees. Although the literature generally implies that the discriminative contrast effect requires trial-level interleaving between categories (i.e., information from one category may need to be held in working memory when another category is presented, a process that interleaving seems especially able to facilitate), the present study may be unique in that it involves a set of well-defined and explicitly-retrievable rules. This may support discriminative contrast even on non-adjacent trials. For example, when learning about the imperfect tense, a subject in the blocked group might have mentally compared that tense with the tense that had just been learned several minutes prior (i.e., the preterite tense). This could have involved contrasting similar suffixes (e.g., "-amos" vs. "-abamos"). By comparison, for other types of materials in the literature such as artists' painting styles, there are no explicitly instructed rules, and the learning about a given artist's painting style in a blocked group may not be available for retrieval and comparison to the style of another artist encountered after a delay.

Thus, with respect to the discriminative contrast hypothesis, the most powerful manipulation of interleaving vs. blocking should arguably involve "isolated" blocking in which each tense is learned in a separate session separated by days or weeks, and where Phase 3 interleaving occurs in the interleaved group in each of those sessions. Under those circumstances, which are also more ecologically valid, it should be more difficult for participants in the blocked group to integrate or contrast what they have learned for one tense with the other. We implemented this design, plus doubled the amount of Phase 3 practice, in the next experiment.

Experiment 3

In the third experiment we investigated whether interleaving or blocking yields better learning when blocking occurs across two training sessions separated by one week and interleaving occurs in both training sessions. In this experiment, (a) the blocked group learned one tense per weekly session; (b) the interleaved group trained on both tenses in the first session, followed by verb conjugation practice in the first and second sessions; and (c) the delayed test occurred one week after the second session. The total amount of Phase 3 practice trials was twice that of the prior experiments. This experiment thus addressed the ecologically relevant question of whether it is advantageous to learn one tense per session in blocked fashion, or both per session in interleaved fashion, while keeping the amount of training materials used constant in both groups.

Methods

Participants. Ninety-six undergraduate students, recruited in the same manner as in the preceding experiments, participated for course credit. All but eight participants successfully completed all three sessions of the experiment (*interleaved* group, $n = 41$; *blocked* group, $n = 47$).

Materials. These were identical to the preceding experiments except for 9 additional practice questions per tense during Phase 3 (18 per session; 36 in total). This doubled the amount of Phase 3 practice and potentially helped to ameliorate the greater amount of forgetting that is to be expected over longer retention intervals. No questions were repeated between sessions.

Procedure. The procedure resembled that of the preceding experiment, including fully randomized practice trials, but involved two training sessions spaced one week apart.

Training. Training schedules for both groups are depicted in Figure 2 (panels d and e).

Blocked group. The blocked training schedule was unchanged except for a one-week delay between Phases 1-3 for the first tense to be learned (Session 1) and Phases 1-3 for the second tense to be learned (Session 2).

Interleaved group. The interleaved training schedule resembled that used in the preceding experiment, including completion of Phases 1-3 for both tenses during an initial training session (Session 1) and in the same order as in Experiment 2. After a one-week delay, a second Phase 3 (Session 2) occurred.

Delayed test. One week after Session 2, participants completed a delayed test that was identical to that used in the preceding experiments.

Results

Training. Phase 1 and 2 practice trial performance for both groups, within either the first or second session, was similar (Table 3). Phase 3 practice trial performance was higher in the blocked than in the interleaved group in Sessions 1 and 2, mirroring the patterns observed in the prior experiments. Although Phase 3 performance in the interleaved group was numerically worse in session 2, that result cannot necessarily be interpreted as evidence that session 2 training had no effect on verb conjugation skill. Rather, forgetting between sessions 1 and 2 may have occurred, masking that Phase 2 learning effect.

The mean (*SE*) training durations in Sessions 1 and 2, respectively, were 15.04 (0.63) and 9.47 (0.20) min in the blocked group and 18.23 (0.52) and 5.09 (0.16) min in the interleaved group, the differences reflecting the divergent training schedules that were used for each group. However, total mean training duration (Sessions 1 and 2 combined) of 24.52 (0.71) and 23.32 (0.50) min in the blocked and interleaved groups, respectively, was highly similar.

With regard to metacognitive judgments (Table 4), participants in the interleaved group assigned higher judgments of difficulty to their training experience in Session 1, $X^2(4) = 42.85$, p

< .0001, as well as in Session 2, $X^2(4) = 29.36$, $p < .0001$. Although the groups did not differ in their judgments of learning in Session 1, $X^2(4) = 1.28$, $p = .86$, participants in the blocked group gave higher judgments of learning in Session 2, $X^2(4) = 19.33$, $p < .001$.

Delayed test. Mean verb conjugation accuracy on the delayed test was 0.52 ($SE = 0.034$) in the blocked group and 0.64 ($SE = 0.033$) in the interleaved group, $t(86) = 2.49$, $p = .015$, $d = 0.53$ (Figure 3), constituting a 23% proportion correct gain in the interleaved condition. Moreover, performance in the blocked group for whichever tense (counterbalanced) was learned in Session 1 ($M = 0.54$, $SE = 0.045$) or Session 2 ($M = 0.52$, $SE = 0.045$) was not significantly different from one another, $t(46) = 0.40$, $p = .69$, $d = 0.058$ (no such analyses are possible for the interleaved group as both tenses were learned during Session 1 and practiced only during Session 2). Counterbalancing of materials across retention intervals in the blocked group was thus not a complicating factor for interpretation.

An ANOVA analogous to that performed for the preceding experiments (with Phase 3 training data collapsed over both sessions) revealed a highly significant Group x Session interaction, $F(1,86) = 97.15$, $p < .0001$, $MSE = 1.92$, $\eta_p^2 = 0.53$, indicating a crossover interaction wherein the blocked group's performance advantage during Phase 3 was reversed on the delayed test. Reliability was again high at 0.89 and 0.87 for the blocked and interleaved groups respectively.

Discussion

In the third experiment we observed that interleaving yields better verb conjugation skills than blocking when training occurs over two weekly sessions. The trial-level implementation of interleaving in this experiment was similar to that of Experiment 2, including its use exclusively during Phase 3 and fully randomized practice trials. However, in this case interleaving occurred during each of two sessions, whereas the blocked group trained on only one tense per session,

and there were twice as many Phase 3 practice trials in both groups. Those design changes appear to have yielded markedly different results than in the preceding experiments.

Experiment 4

For the final experiment we investigated whether an interleaving effect would replicate under identical training conditions as in Experiment 3 but with a more difficult delayed test involving short answer format. Short answer tests are stricter assessments of learning due to the lack of provided answer choices and a chance accuracy rate of effectively zero (for related discussion see Pan & Rickard, 2017). Relative to the multiple-choice format, such tests better approximate how language skill is expressed in ecological circumstances.

Methods

Participants. One hundred and two undergraduate students, recruited in the same manner as in the preceding experiments, participated for course credit. All but eleven students (*interleaved* group, $n = 46$; *blocked* group, $n = 45$) completed all three sessions of the experiment.

Materials. These were identical to the prior experiment excepting a change in delayed test format (short answer, a change facilitated by removing any answer choices) and a greater number of delayed test questions (42, including 24 preterite questions and 18 imperfect questions). That increased amount enabled us to field two questions each involving the "I", "you", and "we" pronouns per tense and six questions invoking each assessed tense rule.

Procedure. The procedure was identical to Experiment 3 excepting the switch to short answer format on the delayed test.

Results

Training. Practice trial data patterns across all training phases were essentially identical to that of Experiment 3 (Table 3). The mean (*SE*) training durations in Sessions 1 and 2,

respectively, was 11.01 (0.27) and 9.68 (0.26) min in the blocked group and 17.78 (0.56) and 5.35 (0.79) min in the interleaved group. Total mean training duration (Sessions 1 and 2 combined) was modestly longer in the interleaved vs. blocked group at 20.69 (0.40) vs. 23.12 (0.59) min, respectively. That pattern differed from Experiment 3, wherein a slight difference was found in the opposite direction.

In terms of metacognitive data (Table 4), participants in the interleaved group assigned higher judgments of difficulty in Session 1, $X^2(4) = 33.50$, $p < .0001$, as well as in Session 2, $X^2(4) = 27.13$, $p < .0001$. Participants in the blocked group gave higher judgments of learning in Session 1, $X^2(4) = 13.01$, $p = .011$, and in Session 2, $X^2(4) = 12.55$, $p = .014$.

Delayed test. Mean verb conjugation accuracy on the delayed test was 0.30 ($SE = 0.032$) in the blocked group and 0.49 ($SE = 0.039$) in the interleaved group, $t(89) = 3.77$, $p < .001$, $d = 0.79$ (Figure 3). That 63% accuracy gain constitutes a larger interleaving effect than observed in Experiment 3 ($d = 0.53$). As with the prior experiment, the effect of counterbalancing across different retention intervals for the blocked group did not yield significant differences, $t(44) = 0.45$, $p = .65$, $d = 0.068$.

An ANOVA identical to that performed for the preceding experiment revealed a highly significant Group x Session interaction, $F(1,89) = 108.60$, $p < .0001$, $MSE = 2.41$, $\eta_p^2 = 0.55$, again indicating a crossover interaction wherein the blocked group's performance advantage during Phase 3 was reversed on the delayed test. The reliability of the delayed test was 0.92 and 0.94 for the blocked and interleaved groups, respectively.

Uniquely in this experiment, there was a non-trivial difference in training duration between the two groups, with that duration being 12% longer for the interleaved group. To explore the possibility that the interleaving effect on the delayed test was solely due to a "time-on-task" advantage for the interleaving group during training, we computed a retention rate

estimate for each participant, wherein delayed test proportion correct was divided by the corresponding training duration (hence measuring retention of learning per unit time spent in training). Mean retention rate was 0.015 (*SE* = 0.0018) in the blocked group and 0.021 (*SE* = 0.0017) in the interleaved group, $t(88) = 2.58$, $p = .012$, $d = 0.54$. Hence, there is a retention advantage for the interleaving group even after adjusting for the differences in training duration.

Discussion

In the fourth experiment we again observed a substantial interleaving effect on a more difficult short answer delayed test that better approximates actual language use. Total training duration was somewhat longer for the interleaved group in this experiment. However, the interleaving advantage remained in a retention rate analysis that adjusted for training duration differences. Moreover, an interleaving effect was observed in Experiment 3 despite the blocked group taking slightly longer on average during training. Hence, it is unlikely in our view that the results of Experiments 3 and 4 were substantially driven by differences in training duration. Rather, these results reflect a retention advantage for interleaving.

General Discussion

Does interleaving enhance the learning of Spanish verb conjugation skills among English speakers? In answer to the three questions posed at this manuscript's outset, (a) we did observe benefits of interleaving over blocking, but those benefits were not universal across all four experiments; (b) the apparent progressive optimization of interleaving across experiments revealed conditions under which the technique can benefit learning, including with an increased number of training trials and notably when (c) it was used for verb conjugation practice across two weekly sessions. It is also notable, and consistent with some prior interleaving results, that the high level of performance achieved by the blocked group at the end of training in Experiments 3 and 4 was not well retained, yielding a crossover interaction between training

group and experimental phase (Phase 3 vs. delayed test).

To our knowledge, this study contributes the first demonstration of an interleaving effect for foreign language learning, and it does so for materials that are substantially more complex than those used in prior studies of interleaving and language learning (e.g., vocabulary words as in Schneider et al., 1998, 2002). Our results also suggest that the interleaving effect for foreign language learning (and perhaps other skills) can be promoted by hybrid scheduling, a topic to which we return below.

The Roles of Spacing and Discriminative Contrast

What accounts for the absence of an interleaving effect in Experiments 1 and 2 versus its emergence in Experiments 3 and 4? One possibility is that spacing played an important role. In Experiments 3 and 4, whereas the blocked group completed all training trials for a given tense in a single session, the interleaved group trained on each tense twice over two weeks (with half as many trials per tense in each session). That spaced exposure to each tense across two sessions may have improved retention. It should be reemphasized, however, that in prior research there is evidence that interleaving's benefits exceed those conferred by spacing alone (Birnbaum et al., 2013; Kang & Pashler, 2012). On the other hand, those earlier experiments did not entail multi-session training and week-long spaced intervals as our Experiments 3 and 4 did.

Beyond a possible spacing effect, our results are also broadly consistent with the discriminative contrast hypothesis. Specifically, in Experiments 3 and 4, the interleaving group practiced on both tenses in each session in alternating fashion, whereas the blocked group trained on only one tense per session. As noted earlier, this design may constitute a more powerful manipulation of discriminative contrast than that used in Experiments 1 and 2.

Hybrid Interleaving Schedules

Interleaving throughout much of the training session, as occurred in Phases 2 and 3 of

Experiment 1, yielded poorer delayed test performance than did blocking. When trial-level interleaving was reserved until verb conjugation practice trials in Phase 3, it yielded performance that was on par with (Experiment 2) or better than (Experiments 3-4) blocking. Why might that type of blocked-to-interleaved training schedule, wherein blocking is used for Phases 1 and 2, yield better learning than the training schedule used in Experiment 1? The answer may stem from the fact that L2 learning of Spanish verb conjugation skills is a multi-stage process involving different cognitive skills at different stages (e.g., learning explicit rules vs. recalling and applying those rules; for an analogous example see Kole & Healy, 2013). For relatively complex skills that involve a transition from knowledge to application (e.g., Anderson & Krathwohl, 2001; Bloom, 1956, 1984), it may be the case that interleaving that is implemented too early impairs the acquisition of basic knowledge (possibly if explicit rules are involved, as suggested by Sorensen & Woltz, 2016), thus affecting learners' ability to later apply that knowledge. More research is needed to scrutinize that possibility across foreign language and other materials (e.g., other subdomains of language learning may also require the initial acquisition of basic knowledge before interleaving and other learning interventions are effective). Additionally, future work that manipulates varying amounts of blocked-to-interleaved practice within a single experiment is needed to directly test the hypothesis that interleaving "too early" may impair learning.

The results of Experiments 3 and 4 also stand in contrast to prior work on hybrid schedules that involved different designs and yielded divergent results. Specifically, the hybrid schedules used in Sorensen and Woltz (2016) and Yan et al. (2017; Experiments 1 and 2) did not involve multiple training phases on component tasks as in the present experiments. Rather, a single task type was first learned under blocked, and then interleaved, conditions. In the Sorensen and Woltz study, a blocked-to-interleaved training schedule yielded better test

performance than interleaving alone, whereas a purely blocked schedule yielded the best test performance overall. In the Yan et al. study, blocked-to-interleaved schedules yielded test performance that was as good as, but not better than, fully interleaved schedules.

Trial-Level Implementations of Interleaving

For various learning materials, it is possible to implement trial-level interleaving across a host of different *dimensions* and with divergent effects as a consequence (e.g., Rau, Aleven, & Rummel, 2013). We primarily implemented interleaving based on tense (which the Spanish instructional literature implies is the most crucial dimension), but in some cases also according to pronoun. Given that the choice of interleaved dimension may be highly influential, the effects of interleaving across different dimensions for learning verb conjugation skills warrant further investigation. Relatedly, Carpenter and Mueller (2013) interleaved training based on pronunciation rule and analyzed test performance in terms of whole word pronunciation; if their data are re-analyzed according to pronunciation rule, the blocking advantage is eliminated.[4]

In addition, trial-level interleaving involved a fixed pattern in Experiment 1 but was random in Experiments 2-4. If unpredictability is a driver of interleaving effects (Bjork, 1999), then random schedules should be more effective. The fact that the interleaved group's test results were on par with or better than the blocked group in Experiments 2-4 is consistent with this possibility. Trial-level randomization might also incorporate *constraints* in that certain types of category change are specified on successive trials (e.g., Sana et al., 2017). In particular, a random schedule that guarantees a pronoun change on each successive trial might yield even larger benefits.

Metacognitive Judgments of Difficulty and Learning

Throughout all four experiments, participants in the interleaved group both performed worse and gave higher difficulty ratings during training. For judgments of learning (assessed in

all but Experiment 1), the similarity in those ratings between groups in Experiment 2 mirrored the delayed test results in that experiment. However, there was a disparity between ratings and delayed test results in Experiments 3 and 4. Specifically, the blocked group in both of those experiments tended to overestimate their mastery of the tense that was trained in each session (providing much higher ratings than the interleaved group). That pattern of responding represents an illusion of competence (Koriat & Bjork, 2005) which is akin to that in prior studies comparing interleaving vs. blocking (e.g., Kornell & Bjork, 2008), as well as massed vs. spaced practice (McCabe, 2011; Soderstrom & Bjork, 2015). It should however be noted that participants in the blocked groups of Experiments 3 and 4 were likely unaware of a forthcoming test involving both tenses (probably moreso than participants in any other condition of any of the experiments) given that they never practiced on both tenses in any training session. As such, they were rating their learning of a single tense at a time and not both tenses together. Nevertheless, it appears that blocked practice involving one tense per session yielded inflated estimates of learning for each tense.

Educational Implications

The present study is educationally relevant in at least four respects. First, it generalizes the interleaving effect to foreign language grammar learning, and to a skill that is widely regarded as one of the most difficult to master for L2 learners of Spanish (Castañeda, 2011; Frantzen, 1995; Iguina & Dozier, 2008; Westfall & Foerster, 1996). As such, this study illustrates the potential utility of interleaving for widely-learned topics beyond mathematics (e.g., Rohrer & Taylor, 2007; Rohrer et al., 2014), which currently stands as the primary example of common classroom materials for which an interleaving benefit has been demonstrated (cf. Hatala et al., 2003; Sana et al., 2017). Second, it raises the possibility that interleaving may benefit other aspects of language learning (although it is important to reemphasize that language is not a

single capacity but a collection of many skill subdomains, and interleaving's benefits are likely to vary by task type; for instance, it has not shown a benefit for learning vocabulary). Third, this study highlights the fact that not all implementations of interleaving guarantee learning benefits, illustrates the potential for hybrid interleaved schedules to combine the "best of both worlds" with regard to blocked and interleaved practice, and raises the possibility that for certain skills, interleaving after foundational materials have been learned may be more effective than interleaving that begins from the outset.

Finally, Experiments 3 and 4 served as a controlled laboratory test of the multi-session blocked training method that is commonly used in L2 Spanish conjugation instruction and found that method wanting. A provocative interpretation of those results is that such blocked training schedules should be abandoned entirely in favor of hybrid, multi-session practice. That conclusion however awaits confirmatory evidence in educational settings.

Limitations and Future Directions

As for any investigation of candidate interventions for improving learning, the conclusions of this study may be limited by factors such as materials, training schedules, and participant populations. Further investigative work stands to yield additional insights on hybrid scheduling, different implementations of interleaving, and other issues. For instance, the delayed test questions were not specifically designed to distinguish between the different types of errors that participants could make (such as incorrect tense selection and/or incorrect pronoun suffix usage; see Appendix for data and a supplementary analysis). A delayed test wherein participants must separately indicate both a tense and pronoun/suffix choice on each trial may reveal more about the nature of verb conjugation errors following interleaving vs. blocking.

Additionally, we were not able to fully disentangle the possible effects of interleaving vs. spacing, nor was that a goal of the current study. Although we converged on the finding that an

interleaved training schedule over two sessions yields substantial benefits over blocking, it should be reemphasized that more than one design variable was altered across experiments (i.e., early vs. later use of interleaving, 1 vs. 2 training sessions, moderate vs. more substantial Phase 3 practice, and a 48 hr. vs. one-week delay). Thus, it could be a combination of multiple factors that yielded the observed interleaving benefits in the latter experiments. Investigating different implementations of interleaving within a single experiment (e.g., manipulating amounts of interleaving or perhaps varying the total number of training sessions) could inform stronger causal inferences about each factor's effects on the efficacy of the technique for these materials. Further, a multi-session training design involving a blocked group that trains on each tense per session vs. an interleaved group based on that of Experiments 3-4 could further illuminate the roles of spacing and interleaving in the present experiments. Followup studies could also potentially adapt the paradigms used by Birnbaum et al., Kang and Pashler (2012), or Taylor and Rohrer (2010) to address the interleaving vs. spacing issue, as well as examine the roles of other aspects of training design.

Conclusions

The benefits of interleaved practice can be substantial for the learning of verb conjugation skills, such as those involving the preterite and imperfect past tenses in Spanish. These benefits are observable when verb conjugation practice occurs in a manner that randomly alternates between tenses and when training involves multiple sessions. From a practical standpoint, the present research reveals that the traditional blocked training approach may not be the most efficacious method of foreign language grammar instruction, and that a hybrid blocked-to-interleaved schedule can generate considerable improvements in learning.

References

Anderson, L., & Krathwohl, D. A. (2001) *Taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Battig, W.F. (1972). Intra-task interference as a source of facilitation in transfer and retention. In R.F. Thompson & J.F. Voss (Eds.), *Topics in learning and performance* (pp. 131-159). New York, NY: Academic Press.

Bird, S. (2011). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 32*(2), 435-452.

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.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance. Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435-459). Cambridge, MA, US: The MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, *2*, 55-64.

Blanco, J. A., & Colbert, M. (2009). *Ventanas: Curso intermedio de lengua española* (2nd ed.). Boston, MA: Vista Higher Learning.

Blanco, J. A., & Tocaimaza-Hatch, C. C. (2015). *Suena* (3rd ed.). Boston, MA: Vista Higher Learning.

Bloom, B. S. (1984). *Taxonomy of educational objectives*. Boston: Allyn and Bacon.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals (1st ed.)* Harlow: Longman Group.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Brady, F. (1998). A theoretical and empirical review of the contextual interference effect and the learning of motor skills. *Quest, 50*(3), 266-293.

Brown, P. C., Roediger, H. L., III., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning.* Cambridge, MA: Belknap Press of Harvard University Press.

Carpenter, S. K. (2014). Spacing and interleaving of study and practice. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying the science of learning in education: Infusing psychological science into the curriculum* (pp. 131-141). Washington, DC: Society for the Teaching of Psychology.

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

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

Castañeda, D. A. (2011). The effects of instruction enhanced by video/photo blogs and wikis on learning the distinctions of the Spanish preterite and imperfect. *Foreign Language Annals*, *44*(4), 692-711.

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. doi:http://dx.doi.org/10.1037/0033-2909.132.3.354

Comrie, B. (1976). *Aspect: An introduction to the study of verbal aspect and related problems* New York, NY: Cambridge Textbooks in Linguistics.

Delgado-Jenkins, H. (1990). Imperfect vs. preterit: A new approach. *Hispania, 73*(4), 1145-1146.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. *Memory, 10*, 317-344.

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. doi:http://dx.doi.org/10.1177/1529100612453266

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Eglington, L. G., & Kang, S. H. (2017). Interleaved Presentation Benefits Science Category Learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 475-485.

Fernández, M., & Roth, O. (2013) *Atlas de la lengua española en el mundo*. Madrid: Fundación Telefónica.

Fernández Vítores, D. (2015). *El español: una lengua viva*. Available at: http://www.icex.es/icex/wcm/idc/groups/public/documents/documento_anexo/mde2/njm1/~edisp/dax2016635284.pdf

Frantzen, D. (1995). The effects of grammar supplementation on written accuracy in an intermediate Spanish content course. *The Modern Language Journal, 79*(3), 329-344.

Goodall, G., & Lear, D. (2017). *Conéctate: Introductory Spanish* (2nd ed.). New York, NY: McGraw-Hill Education.

Goode, S., & Magill, R. A. (1986). Contextual interference effects in learning three badminton serves. *Research Quarterly for Exercise and Sport*, *57*(4), 308-314.

Hall, K. G., Domingues, D. A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills*, *78*(3), 835-841. doi:http://dx.doi.org/10.2466/pms.1994.78.3.835

Hatala, R. M., Brooks, L. R., & Norman, G. R. (2003). Practice makes perfect: The critical role of mixed practice in the acquisition of ECG interpretation skills. *Advances in Health Sciences Education*, *8*(1), 17-26.

Iguina, Z, & Dozier, E. (2008). *Manual de gramática: Grammar reference for students of Spanish* (4th ed.). Boston, MA: Thomson Higher Education.

Kang, S. H. K. (2017). The benefits of interleaved practice for learning. In J. C. Horvath, J. Lodge, & J. A. C. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 79-93). New York: Routledge.

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. doi:http://dx.doi.org/10.1002/acp.1801

Kole, J. A., & Healy, A. F. (2013). Is retrieval mediated after repeated testing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 462-472.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187-194.

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

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*(4), 239-243.

Lapkin, S., Hart, D., & Harley, B. (1998). Case study of compact core French models: Attitudes and achievement. *French second language education in Canada: Empirical studies*, 3-30.

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL quarterly*, *28*(3), 563-579.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9*(3-5), 241-289. doi:http://dx.doi.org/10.1016/0167-9457(90)90005-X

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

Nissenberg, G. (2013). *Practice makes perfect: Complete Spanish all-in*-one (1st ed.). New York, NY: McGraw Hill.

Ozete, O. (1988). Focusing on the preterite and imperfect. *Hispania*, 71(3), 687-691.

Pan, S. C. (2015). The interleaving effect: mixing it up boosts learning. In G. Cook (Ed.), *Scientific American, Mind Matters*. Available at: http://www.scientificamerican.com/article/the-interleaving-effect-mixing-it-up-boosts-learning/

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3), 278-292.

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 108*(4), 563-575.

Porter, J. M., & Magill, R. A. (2010). Systematically increasing contextual interference is beneficial for learning sport skills. *Journal of Sports Sciences*, *28*(12), 1277-1285.

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave?. *Learning and Instruction*, *23*, 98-114.

Rickard, T. C., Lau, J. S. H., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic Bulletin & Review*, *15*(3), 656-661.

Roediger, H. L., III., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248. doi:http://dx.doi.org/10.1016/j.jarmac.2012.09.002

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguist, 38*(6), 906-911. doi:http://dx.doi.org/10.1093/applin/amw052

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*(3), 355-367. doi:10.1007/s10648-012-9201-3

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481-498. doi:http://dx.doi.org/10.1007/s11251-007-9015-8

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review, 21*(5), 1323-1330. doi:http://dx.doi.org/10.3758/s13423-014-0588-3

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*(3), 900-908. doi:http://dx.doi.org/10.1037/edu0000001

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*(1), 84-98.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217. doi:http://dx.doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (1998). Contexual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 77-90). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*(2), 419-440. doi:http://dx.doi.org/10.1006/jmla.2001.2813

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL?. *System*, *35*(3), 305-321.

Shea, C. H., Kohl, R., & Indermill, C. (1990). Contextual interference: Contributions of practice. *Acta Psychologica*, *73*(2), 145-157.

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*(2), 179-187. doi:http://dx.doi.org/10.1037/0278-7393.5.2.179

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199. doi:http://dx.doi.org/10.1177/1745691615569000

Sorensen, L. J., & Woltz, D. J. (2016). Blocking as a friend of induction in verbal category learning. *Memory & Cognition, 44*(7), 1000-1013. doi:http://dx.doi.org/10.3758/s13421-016-0615-x

Suzuki, Y. (2017). The Optimal Distribution of Practice for the Acquisition of L2 Morphology: A Conceptual Replication and Extension. *Language Learning 67*(3), 512-545. doi:http://dx.doi.org/10.1111/lang.12236

Suzuki, Y., & DeKeyser, R. (2017). Exploratory research on second language practice distribution: An aptitude × treatment interaction. *Applied Psycholinguistics, 38*(1), 27-56. doi:http://dx.doi.org/10.1017/S0142716416000084

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*(6), 837-848. doi:http://dx.doi.org/10.1002/acp.1598

Ullman, M. T., & Lovelett, J. T. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, 34(1) 39-65.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*(1), 163-167. doi:http://dx.doi.org/10.1016/j.cognition.2008.07.013

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750-763.

Westfall, R., & Foerster, S. (1996). Beyond aspect: New strategies for teaching the preterite and the imperfect. *Hispania, 79*(3), 550-560.

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How Should Exemplars Be Sequenced in Inductive Learning? Empirical Evidence Versus Learners' Opinions. *Journal of Experimental Psychology: Applied*, *23*(4), 403-416.

Footnotes

1. Although conjugated verbs in Spanish differ across at least seven different pronoun types and more than three root verb endings, for logistical reasons our materials included only suffixes corresponding to the pronouns “I”, “you [singular],” and “we”, and only for regular verbs whose infinitive forms end in “-ar”.
2. In all experiments, the relative difference in delayed test performance between the blocked vs. interleaved groups did not differ as a function of native English speaking ability.
3. Of the four rules learned per tense, all but the third rule of the imperfect tense (“stating one’s age in the past”) were represented during Phase 3 trials and on the delayed test. Although we introduced that rule in Phase 1 for completeness and to equalize the number of rules in that phase, it is often easy to identify sentences that mention age. As such, no practice trials invoking that rule appeared outside Phase 1, although it was still included on summary slides.
4. We thank Veronica Yan for contributing this insightful observation.

Table 1.

Preterite and Imperfect Past Tense Rules and Verb Suffixes

Tense		Detail
Tense rules		
	Preterite	
		1. For past actions that had a specific and clear beginning and/or end.
		2. To specifically state the beginning and end of a past action.
		3. For past actions that were repeated a specific number of times.
		4. For past actions that occurred during a specific period of time.
	Imperfect	
		1. For past actions that lack a specific and clear beginning or end.
		2. For past actions that were repeated habitually.
		3. For stating one's age in the past.
		4. For past actions that "set the stage" for another action.
Suffixes		
	Preterite	
		If the pronoun is "I" ("*yo*"), replace "*-ar*" with "*-e*"
		If the pronoun is "you" ("*tu*"), replace "*-ar*" with "*-aste*"
		If the pronoun is "we" ("*nosotros*"), replace "*-ar*" with "*-amos*"
	Imperfect	
		If the pronoun is "I" ("*yo*"), replace "*-ar*" with "*-aba*"
		If the pronoun is "you" ("*tu*"), replace "*-ar*" with "*-abas*"
		If the pronoun is "we" ("*nosotros*"), replace "*-ar*" with "*-abamos*"

Note. Verb suffixes were limited to those used for the "I", "you [singular]," and "we" pronoun equivalents only. Rules adapted from Frantzen (1995), Iguina and Dozier (2008), and Westfall and Foerster (1996).

Table 2.

Training and Delayed Test Example Materials

	Tense	Example sentence or question (*answer*)
Phase 1 (Rules)		
	Preterite	
		Rule 1 example: "I spoke with my mother yesterday."
		Rule 2 example: "Yesterday I began studying at 8 o'clock."
		Rule 3 example: "Last week you ate cookies three times."
		Rule 4 example: "We worked together for six months."
	Imperfect	
		Rule 1 example: "I used to speak with my friend."
		Rule 2 example: "We used to lunch together every day."
		Rule 3 example: "You were three years old when you started."
		Rule 4 example: "You were eating when you received the phone call."
Phase 1 (Practice trials)		
	Preterite	
		Is the following sentence *preterite*? "On Tuesday I ate four tacos." (*Yes*)
		Is the following sentence *preterite*? "I used to walk in the park." (*No*)
	Imperfect	
		Is the following sentence *imperfect*? "I used to read in my free time." (*Yes*)
		Is the following sentence *imperfect*? "We slept for eight hours." (*No*)
Phase 2 (Suffixes)		
	Preterite	
		"I" example: "I *hable* with my mother yesterday."
		"you" example: "You *hablaste* with my mother yesterday."
		"we" example: "We *hablamos* with my mother yesterday."
	Imperfect	
		"I" example: "I used to *hablaba* with my mother."
		"you" example: "You used to *hablabas* with my mother."
		"we" example: "We used to *hablabamos* with my mother."
Phase 2 (Practice trials)		
	Preterite	
		Conjugate *bailar* into: "I ____ with my friend last month." (*baile*)
		Conjugate *bailar* into: "You ____ with my friend last month." (*bailaste*)
		Conjugate *bailar* into: "We ____ with my friend last month." (*bailamos*)

(table continues)

Table 2. *(continued)*

	Tense	Example sentence or question (*answer*)
	Imperfect	
		Conjugate *bailar* into: "I used to ____ with my friend." (*bailaba*)
		Conjugate *bailar* into: "You used to ____ with my friend." (*bailabas*)
		Conjugate *bailar* into: "We would ____ together every day." (*bailabamos*)
Phase 3 (Practice trials)		
	Preterite	
		Conjugate *hablar* into: "We ____ with two doctors last week." (*hablamos*)
		Conjugate *jugar* into: "You ____ for the team for 2 years." (*jugaste*)
	Imperfect	
		Conjugate *hablar* into: "I used to ____ with my teacher." (*hablaba*)
		Conjugate *jugar* into: "We would ____ together every day." (*jugabamos*)
Delayed Test		
	Preterite	
		Conjugate *apoyar* (to support) into: "Yo ____ el por tres anos." / "I supported him for three years." (*apoye*) a. *apoye* b. *apoyaste* c. *apoyamos* d. *apoyaba* e. *apoyabas* f. *apoyabamos*
		Conjugate *parar* (to stop) into: "Nosotros ____ la semana pasada." / "We stopped last week." (*paramos*) a. *pare* b. *paraste* c. *paramos* b. *paraba* e. *parabas* f. *parabamos*
	Imperfect	
		Conjugate *llamar* (to call) into: "Tu ____ ella cada dia." / "You used to call her every day." (*llamabas*) a. *llame* b. *llamaste* c. *llamamos* b. *llamaba* e. *llamabas* f. *llamabamos*
		Conjugate *usar* (to use) into: "Nosotros ____ lápices cada dia." / "We used pencils every day." (*usabamos*) a. *use* b. *usaste* c. *usamos* b. *usaba* e. *usabas* f. *usabamos*

Note. Where multiple-choice questions were used (Experiments 1-3), the six answer options were randomly ordered on each trial. Diacritical marks (accent marks) and tense labels (i.e., preterite or imperfect) were not shown to participants in the actual experiment. Translations were simplified in some cases to maintain consistency across all materials in the experiment (e.g., for sentences involving the

phrase “used to”, the correctly translated sentence is usually prefaced by “antes”; however a translation lacking that word was used such that all delayed test translations began with “yo,” “tu”, or “nosotros” prior to a blank; similarly, in the above example with “llamabas”, the fully translated sentence begins with “Tu la llamabas a ella…”).

Table 3.

Training Session Practice Trial Means (SE)

	Training session	Group	Phase 1: Rules		Phase 2: Suffixes	Phase 3: Verb Conjugation Practice
			First cycle	Second cycle		
Experiment 1						
	1	Blocked	0.75 (0.023)	0.86 (0.020)	0.89 (0.025)	0.91 (0.018)
		Interleaved	0.82 (0.021)	0.89 (0.018)	0.78 (0.031)	0.66 (0.033)
Experiment 2						
	1	Blocked	0.85 (0.015)	0.93 (0.014)	0.91 (0.019)	0.88 (0.017)
		Interleaved	0.80 (0.017)	0.92 (0.012)	0.91 (0.021)	0.64 (0.033)
Experiment 3						
	1	Blocked	0.77 (0.024)	0.90 (0.018)	0.89 (0.029)	0.90 (0.016)
		Interleaved	0.81 (0.022)	0.91 (0.015)	0.87 (0.027)	0.64 (0.032)
	2	Blocked	0.81 (0.022)	0.93 (0.015)	0.91 (0.028)	0.89 (0.023)
		Interleaved	—	—	—	0.55 (0.033)
Experiment 4						
	1	Blocked	0.79 (0.026)	0.91 (0.019)	0.87 (0.041)	0.86 (0.026)
		Interleaved	0.82 (0.016)	0.91 (0.016)	0.89 (0.025)	0.62 (0.035)
	2	Blocked	0.78 (0.024)	0.89 (0.019)	0.81 (0.042)	0.87 (0.024)
		Interleaved	—	—	—	0.63 (0.031)

Note. For simplicity, Phases 1-3 data are collapsed across tenses in all cases (the overall patterns of training results did not differ by tense in any of the experiments). In Experiments 3-4, there was no Phase 1 or Phase 2 for the interleaved group in session 2.

Table 4.

Frequency of Metacognitive Judgments Collected During Training

			Judgments of Difficulty					Judgments of Learning				
	Training session	Group	Very easy	Easy	Moderate	Somewhat difficult	Very difficult	Excellent	Good	Average	Fair	Poor
Experiment 1												
	1	Blocked	1	23	14	4	0	—	—	—	—	—
		Interleaved	1	6	17	14	6	—	—	—	—	—
Experiment 2												
	1	Blocked	5	23	10	2	1	4	17	9	6	0
		Interleaved	2	13	18	6	5	2	15	12	5	4
Experiment 3												
	1	Blocked	22	21	4	0	1	4	15	8	15	8
		Interleaved	2	8	21	8	3	4	11	8	15	4
	2	Blocked	18	18	9	1	1	10	22	9	4	2
		Interleaved	2	8	11	19	2	1	10	12	9	10
Experiment 4												
	1	Blocked	19	19	6	1	0	9	21	9	5	1
		Interleaved	3	9	26	4	4	2	15	9	16	4
	2	Blocked	14	24	4	3	0	10	16	8	8	3
		Interleaved	3	12	13	13	5	1	11	14	14	6

Note. Judgments of difficulty were collected at the end of each training session and immediately prior to judgments of learning (if collected). Judgments of learning were not collected in Experiment 1 and were not administered to 11 participants in Experiment 2.

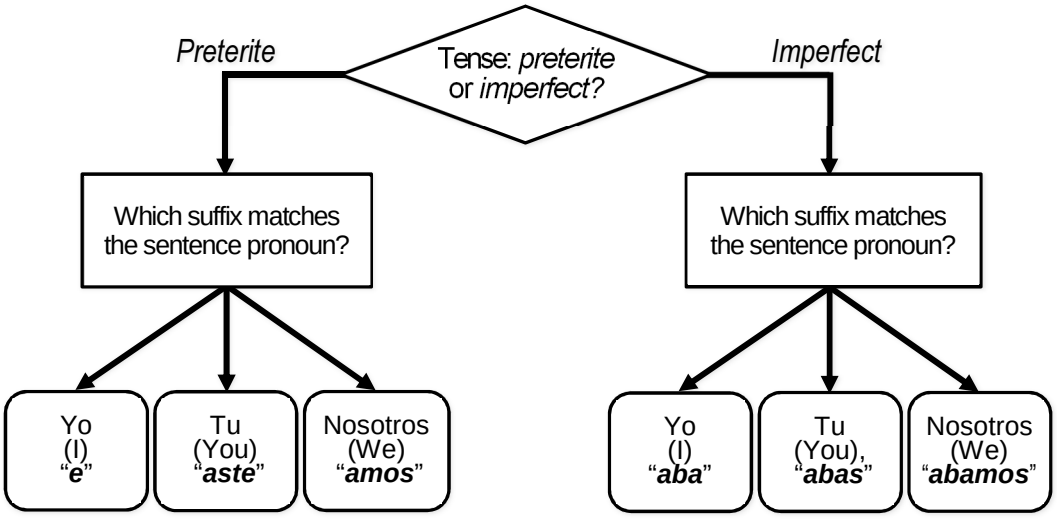

Figure 1. Flowchart depicting a process of conjugating Spanish "-ar" root verbs in the preterite and imperfect tenses for sentences in which the subject is the Spanish equivalent of "I", "you", or "we". On the bottom level of the figure, the correct Spanish suffix is listed below the corresponding pronoun.

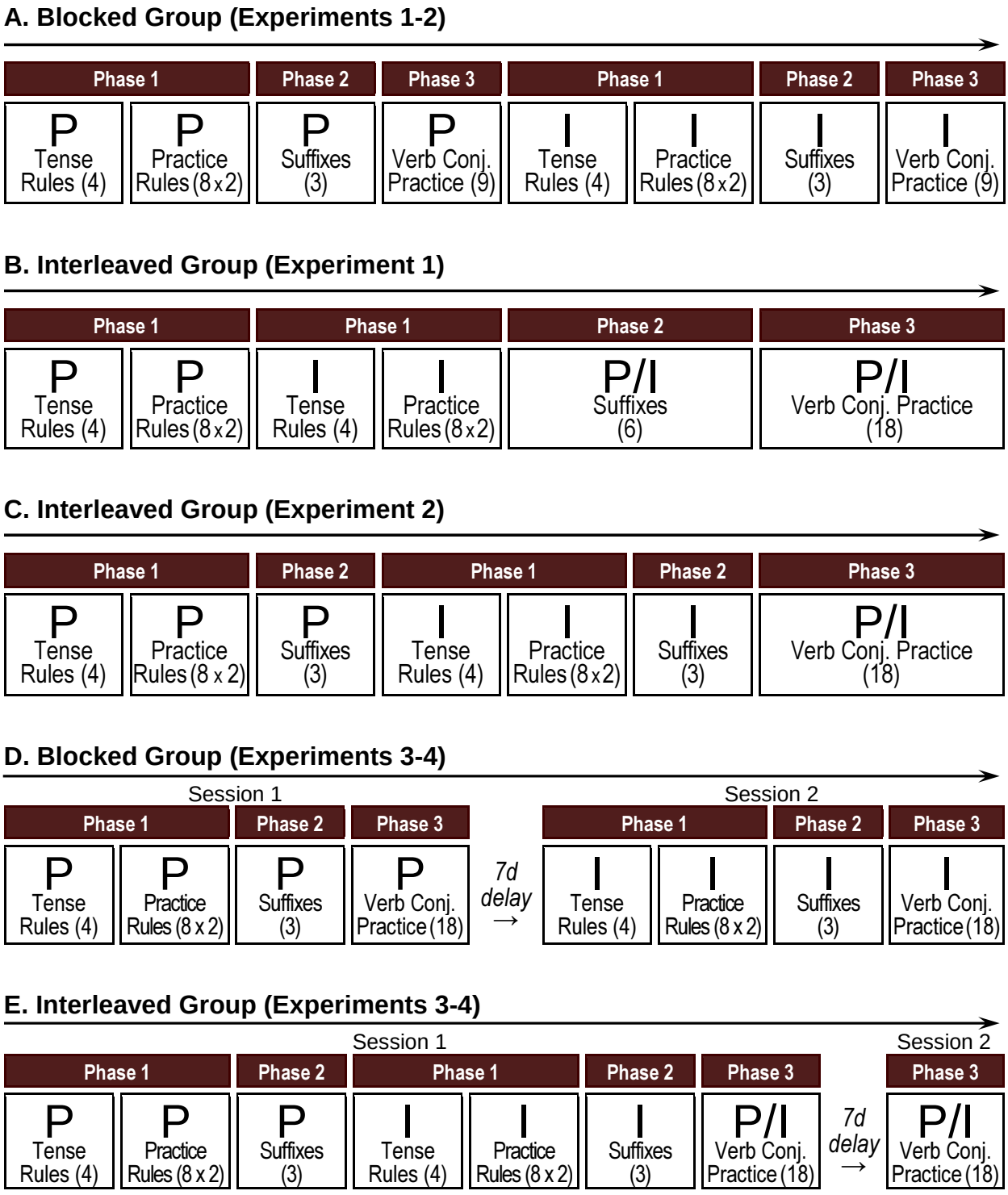

Figure 2. Schematic timeline of the training session designs used in the blocked and interleaved groups of Experiments 1-4. Each box represents a separate stage of training, with large capital letters indicating tense (P = *preterite*; I = *imperfect*). The number(s) in parentheses indicate the

number of presentation slides or practice trials (Note: 8 x 2 refers to two cycles of eight trials each). P/I within a single box indicates trial-level interleaved practice (alternating between tense). Summary slides were presented in Phases 1 (after presentation of the tense rules) and 3 (prior to the start of verb conjugation practice). Experiments 1-2 involved one training session and Experiments 3-4 involved two training sessions separated by one week. Only one of two counterbalanced tense orders (preterite or imperfect first) is depicted.

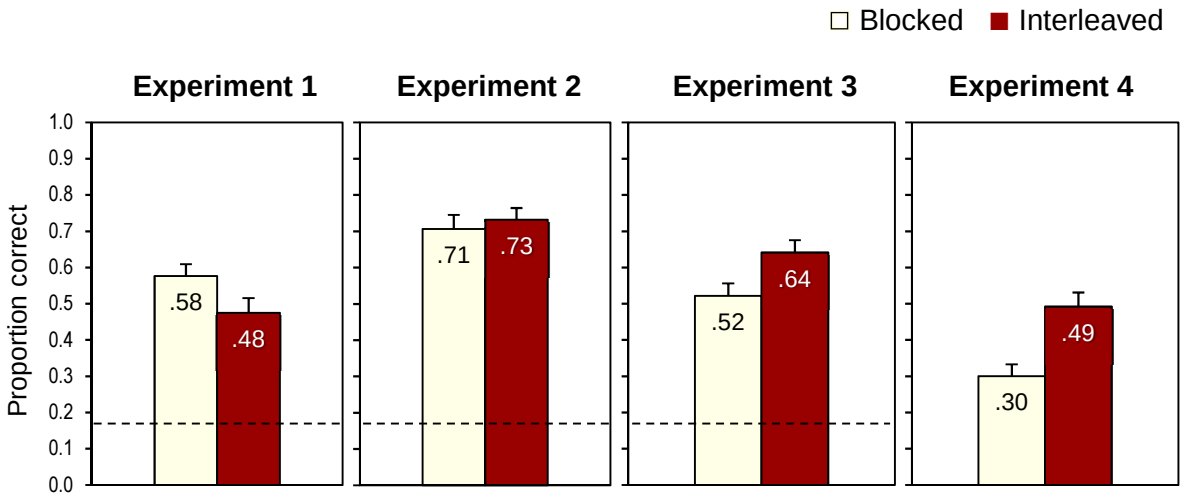

Figure 3. Delayed test performance in the blocked vs. interleaved groups of Experiments 1-4. The retention interval was 48 hr in Experiments 1-2 and one week in Experiments 3-4. A multiple-choice test format was used in all but Experiment 4, which involved short answer. The dotted line refers to the expected accuracy rate that would be expected from pure guessing on the multiple-choice delayed test of Experiments 1-3 (given a one-in-six chance of randomly selecting the correct answer). Data are collapsed across tense for simplicity (overall patterns did not markedly differ by tense). Error bars indicate standard errors of the means.

Appendix

Supplementary Analysis of Delayed Test Errors in Experiments 1-4

There were group differences in the frequencies of delayed test errors made in Experiments 1, 3, and 4 (X^2 test for independence, ps < .0001). In Experiment 1, that difference appeared to be driven by an increased number of errors involving incorrect suffix selection among the interleaved group. In Experiments 3 and 4, increased numbers of tense and/or pronoun suffix errors among the blocked group appeared to be the basis for the group difference.

Mean Proportion of Errors (*SE*) on the Delayed Test in Experiments 1-4

	Group	Tense suffix errors	Pronoun suffix errors	Both tense suffix and pronoun suffix errors
Experiment 1	Blocked	0.29 (0.026)	0.087 (0.009)	0.10 (0.012)
	Interleaved	0.26 (0.024)	0.17 (0.023)	0.18 (0.016)
Experiment 2	Blocked	0.21 (0.027)	0.11 (0.019)	0.092 (0.029)
	Interleaved	0.15 (0.017)	0.10 (0.017)	0.10 (0.021)
Experiment 3	Blocked	0.32 (0.023)	0.16 (0.023)	0.11 (0.016)
	Interleaved	0.24 (0.021)	0.10 (0.015)	0.10 (0.016)
Experiment 4	Blocked	0.23 (0.024)	0.30 (0.029)	0.23 (0.021)
	Interleaved	0.22 (0.023)	0.18 (0.02)	0.16 (0.018)

Note. *Tense suffix errors* = verb conjugations that corresponded to the given pronoun but were in the incorrect tense, *pronoun suffix errors* = verb conjugations that were in the correct tense but had the incorrect ending for that given pronoun, *both* errors = verb conjugations in incorrect tense and did not correspond to given pronoun.