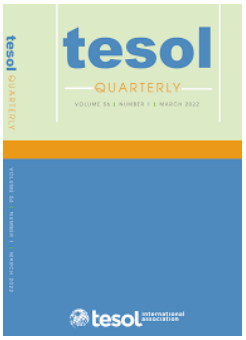
Does the reuse of constructions promote fluency development in task repetition? A usage-based perspective



Yuichi Suzuki
Kanagawa University, Japan
Masaki Eguchi
University of Oregon
Nel de Jong
University of Amsterdam Netherlands

Abstract

In this task-repetition intervention study, L2 learners' reuse of linguistic constructions was analyzed to investigate to what extent recurring reliance on specific constructions during the same task repetition predicts fluency development. English-as-a-foreign-language (EFL) learners performed oral narrative tasks three times per day under two task repetition schedules: blocked (Day 1: Prompt A-A-A, Day 2: B-B-B, Day 3: C-C-C) versus interleaved (Day 1: Prompt A-B-C, Day 2: A-B-C, Day 3: A-B-C). From a usage-based perspective, their reuse of constructions across the same prompt was examined at both concrete (lexical unigram [e.g., "bicycle"] and trigram [e.g., "behind the bicycle"]) and abstract (parts of speech trigram [e.g., "preposition determiner noun"]) level. Subsequent analyses revealed that blocked practice led to higher reuse of both concrete and abstract constructions than interleaved practice. Reuse frequency was correlated with during-training and pretest−posttest fluency changes. Particularly, greater reuse of lexical and abstract trigrams during interleaved practice led to improvements in speed and breakdown fluency (i.e., shorter mean syllable duration and fewer mid-clause pauses) after the intervention, albeit with higher effort (indicated by longer mid-clause and clause-final pauses). Taken together, these findings indicate that manipulating task-repetition schedule may systematically induce reuse of linguistic constructions, which may promote proceduralization (entrenchment) of constructional knowledge at both concrete and abstract levels.

Introduction

A large body of second language (L2) research indicates that task repetition improves fluency (e.g., Bygate, 2018; N. De Jong & Perfetti, 2011). When L2 learners narrate the same cartoon story multiple times, they are likely to reuse the same linguistic items across repeated task performances. Presumably, this process activates and strengthens access to those linguistic constructions that vary in size (e.g., single vs. multi-word items) and abstractness (e.g., concrete vs. abstract patterns) (Ellis & Wulff, 2020; Goldberg, 1995, 2006; Tomasello, 2003). Repeated use of specific constructions is gradually entrenched (Langacker, 2008; Schmid, 2017) and leads to faster and more efficient processing, the concept known as proceduralization (DeKeyser, 2018, 2020). Proceduralization of linguistic knowledge is considered essential for developing fluent L2 speech (Kahng, 2014; Kormos, 2006). However, this potential link between proceduralization of specific linguistic constructions and fluency development has not been sufficiently explored.

From a pedagogical perspective, it is worth investigating how learning conditions can be manipulated to induce greater reuse of linguistic constructions and improve specific aspects of utterance fluency (e.g., speed, breakdown, and repair fluency). Findings yielded by an emerging line of L2 research suggest that more frequent reuse of constructions across multiple speaking task performances may relate to L2 fluency development (e.g., N. de Jong & Perfetti, 2011; N. de Jong & Tillman, 2018). The current investigation expands the scope of these previous studies by focusing on the extent to which timing of speaking task repetition influences construction reuse across multiple performances and how this practice relates to L2 fluency development.

Literature Review

A Usage-based Constructional Approach to L2 Teaching and Learning

In the last few decades, constructionist usage-based approaches have emerged as an overarching theory of first language acquisition (Goldberg, 1995, 2006; Tomasello, 2003). In this theoretical framework, grammatical structures are analyzed in the units of constructions. In this context, constructions are defined as pairings of form and meaning that vary in complexity and abstractness and range from concrete lexical items (e.g., morphemes, words, idioms, multi-word sequences) to more abstract, rule(-like) schematic patterns (e.g., ditransitive, active, and passive sentence frames). In other words, “the network of constructions captures our grammatical knowledge *in toto*, that is, it is constructions all the way down” (Goldberg, 2006, p. 18).

The usage-based approach has profound implications for L2 teaching and

acquisition (Ellis, 2002; Ellis & Wulff, 2020; Robinson & Ellis, 2008; Tyler & Ortega, 2018), as frequency (e.g., token and type input frequency) is a key factor for L2 construction learning. Although the effects of frequency have been examined extensively in relation to receptive (input-based) L2 learning mode (Ellis, 2009), production (output) mode of L2 construction learning remains insufficiently studied. Nonetheless, it is reasonable to assume that producing the same constructions repeatedly (i.e., increasing token frequency) also plays an important role in strengthening the memory traces and enhancing the retrieval of those constructions.

When a novel linguistic construction is used repeatedly, it gradually becomes entrenched (Langacker, 2008). Arguably, the degree of construction entrenchment may be linked to the efficiency with which such construction is retrieved from memory (Schmid, 2017). In earlier stages of L2 acquisition, inexperienced learners effortfully produce constructions that are not sufficiently entrenched. Thus, using the same linguistic constructions repeatedly in output allows learners to produce them more quickly and efficiently. This gradual learning process through repetition is essential and is linked to a putative learning stage called proceduralization—creation of procedural knowledge that can be used to execute L2 skills more efficiently—which is a prerequisite for further automatization (DeKeyser, 2018, 2020; Suzuki, in press).

Proceduralization and automatization of linguistic construction have been postulated to underlie the development of oral L2 fluency through task repetition (Kormos, 2006; Wood, 2006). When certain linguistic constructions are reused through task-repetition practice, their retrieval becomes faster and more efficient, which is likely to exert a positive influence on multiple aspects of utterance fluency (e.g., faster articulation rate, shorter pauses, fewer self-repairs). When concrete items are reused through task iteration (e.g., single words, multi-word sequences, n-grams), this may also contribute to more efficient encoding of abstract, schematic constructions. For instance, when a learner repeatedly uses different lexical trigrams (e.g., *rang a bell*, *drive a car*, *saw a tiger*) in which the same schematic construction depicting a transitive event underlies, an abstract construction (e.g., verb−determiner−noun) may be extracted and gradually entrenched or proceduralized (cf., Tomasello, 2003). In sum, a usage-based constructionist perspective offers a useful vantage point to analyzing linguistic items that vary in size and abstractness, which may be linked to L2 fluency development through task repetition.

Constructions and L2 Utterance Fluency

Task repetition is effective for L2 utterance fluency development (see Bygate,

2018 for a review). The speech processes enhanced by task repetition is explained by speech production model (Kormos, 2006; Levelt, 1989). These speech models essentially consist of three stages: (a) the conceptualization (e.g., generating preverbal message), (b) the formulation (e.g., encoding of lexical and grammatical knowledge), and (c) the articulation. Most relevant to the current study's aim is the formulator stage where speakers need to retrieve linguistic constructions (e.g., single- and multi-word items, as well as abstract patterns) efficiently for fluent speech production. Reuse of the same constructions across repeated task performances presumably enhances the access to them (Ellis & Wulff, 2020; Goldberg, 1995, 2006; Tomasello, 2003).

Previous L2 research indicates that constructional knowledge (i.e., mental representation consisting of both single lexical and multi-word items) is tightly linked to utterance fluency, particularly speed and breakdown aspects of fluency (N. H. de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013; Koizumi & In'nami, 2013; Tavakoli & Uchihara, 2020). Speed fluency is typically gauged by articulation rate, whereas breakdown fluency is typically distinguished for mid-clause pauses (i.e., within the Analysis of Speech [AS] unit) and clause-final pauses (i.e., at the boundary of the AS unit) (Tavakoli & Skehan, 2005). The former presumably reflects linguistic breakdown (e.g., lexical and syntactic) in the linguistic formulation, whereas the latter typically reflects planning of content in the conceptualization stage (N. H. de Jong, 2016; Kahng, 2018; Lambert, Aubrey, & Leeming, 2020). In addition, accumulating evidence suggests that directly teaching linguistic constructions beyond the single words (e.g., multi-word expressions, formulaic sequences) can accelerate articulation rate and reduce mid-clause pauses (Nergis, 2021; Wood, 2010). Taken together, these previous findings suggest that systematic interventions that enhance the retrieval processes of linguistic constructions can exert positive influence on L2 fluency.

Construction Reuse in Task Repetition

While task repetition is effective in improving L2 utterance fluency, the underlying mechanisms that support L2 fluency development remain unclear. Available empirical evidence nonetheless indicates that when linguistic constructions that vary in size and abstractness (i.e., single lexical items, multi-word expressions, sentence patterns) are used during repeated task performance, this practice contributes to L2 speech development (Boers, 2014; N. de Jong & Perfetti, 2011; N. de Jong & Tillman, 2018; Thai & Boers, 2016).

In a study conducted by N. de Jong and Perfetti (2011), ESL learners performed a speech task multiple times under either (a) a content-repetition condition in which

narration of one topic (e.g., sports) was repeated three times or (b) a no-content-repetition condition in which each speech related to a different topic (e.g., sports, learning English, and travel). The authors found that the content-repetition group exhibited greater lexical overlap (i.e., reused the same single lexical items more frequently) than did participants assigned to the no-content-repetition condition. Furthermore, the lexical overlap was moderately correlated with pretest–posttest fluency changes (i.e., phonation time and mean pause length). Because the lexical items that were recycled were high-frequency words (e.g., good, have, make, and think) that are found in various syntactic structures (e.g., think that . . .), N. de Jong and Perfetti (2011) "speculated that it is not the words themselves but the processing of sentence constructions and expressions they are used in that was proceduralized" (p. 557). These assertions indicate that larger units of construction that are beyond the single lexical items (sequences of two or more words, known as n-grams) must be investigated more thoroughly.

In response to this need, Boers (2014) and Thai and Boers (2016) examined the n-grams that were recycled though task iteration. In both studies, the authors compared two task-repetition conditions involving monologue performance that differed in time pressure: (a) constant time (e.g., repeating a three-minute speech three times) and (b) decreasing time (e.g., performing the speech for four, three, and two minutes). Their results revealed that (1) the decreasing time condition resulted in higher n-gram reuse than the constant time condition, and (2) n-gram reuse was weakly-to-moderately correlated with pruned speech rate improvement from the first to the third performance. These findings suggest that repeated use of the same constructions (e.g., n-grams) is a major factor influencing fluency changes, particularly in the decreasing time condition, although increasing the time pressure may adversely affect the accuracy and complexity of speech performance (Thai & Boers, 2016).

More recently, N. de Jong and Tillman (2018) adopted the Natural Language Processing (NLP) approach to conduct a sophisticated analysis of linguistic constructions during task-repetition practice. ESL learners that narrated the same six-frame picture stories three times each under constant and decreasing time conditions. Three levels of linguistic analyses were conducted to measure the between-performance similarity of (a) lexical unigrams (e.g., *car*), (b) lexical trigrams (e.g., *the red car*), and (c) parts-of-speech (POS) trigrams (e.g., determiner–adjective–noun). Using the NLP technique, the authors computed a cosine similarity score (described in detail in the Method section) that captures the reuse of concrete and abstract constructions. Their analyses revealed that, in both constant and decreasing time condition, the cosine similarity score of POS trigrams was positively correlated with two aspects of utterance fluency (i.e., proportion of time

filled with speech and mean syllable duration). These results suggest that repeated POS trigram reuse, presumably leading to proceduralization of abstract constructions, may eventually contribute to L2 fluency development.

Motivations for the Current Study

The goal of the current study is to extend the emerging line of work aiming to elucidate how recycling of various construction types through task-repetition practice is related to L2 fluency development (Boers, 2014; N. de Jong & Perfetti, 2011; N. de Jong & Tillman, 2018; Thai & Boers, 2016). The study was particularly motivated by three observations pertaining to these investigations. First, given that the novel NLP technique used by N. de Jong and Tillman (2018) focused on ESL learners, there is a need to validate the utility of their linguistic analysis approach for different L2 populations (e.g., EFL learners). Second, extant task-repetition studies focused on comparisons between specific task-repetition conditions, such as constant vs. decreasing time (Boers, 2014; N. de Jong & Tillman, 2018; Thai & Boers, 2016) and content-repetition and no-content-repetition (N. de Jong & Perfetti, 2011). In contrast, in the present study, blocked and interleaved task repetition—a promising pedagogical option to enhance L2 fluency—are compared (Y. Suzuki, 2021). Third, with the exception of N. de Jong and Perfetti (2011), none of the other authors employed a pretest−posttest design to examine the relationship between construction reuse and L2 fluency changes assessed by a new (untrained) task. Because proceduralization of constructions (abstract ones in particular) through repetition should ideally transfer to new task performance, it is important to explore how construction reuse during training contributes to L2 fluency changes in an unfamiliar context.

The Current Study

The aim of the present study was advancing our understanding of the link between construction reuse during task-repetition practice and L2 fluency development. For this purpose, the data gathered as a part of the fluency intervention study conducted by Y. Suzuki (2021) was re-analyzed following N. de Jong and Tillman's (2018) NLP methodology. In this study, Japanese EFL learners were assigned to two groups which engaged in either a blocked practice or an interleaved task-repetition practice, based on the following formats:

Blocked practice: Day 1: AAA, Day 2: BBB, Day 3: CCC
Interleaved practice: Day 1: ABC, Day 2: ABC, Day 3: ABC

As illustrated above, participants assigned to the blocked practice condition performed the same oral narrative task three times as a part of a single training session, while those in the interleaved practice condition repeated the same task on three consecutive days. Their fluency changes were assessed both before (pretest) and after (posttest) practice schedule completion by presenting them with novel oral narrative tasks (i.e., different from those used in the training sessions). The results showed that blocked practice led to greater fluency development (i.e., shorter syllable duration and shorter mid-clause pauses) than interleaved practice both during training and from the pretest to the posttest (see Appendix S1 in Online Supplementary File). However, it remained unclear why blocked practice was more effective than interleaved practice. Because reuse of specific constructions presumably plays an important role in promoting proceduralization underlying fluent L2 speech, detailed analysis of recycled constructions (i.e., lexical unigrams, lexical trigrams, and POS trigrams) allows us to probe the potential links between the construction reuse and L2 fluency development. The following three research questions (RQs) were addressed:

1. Is there a difference between the blocked and the interleaved practice condition in the extent of construction reuse (lexical unigrams, lexical trigrams, and POS trigrams) during the training phase?
2. To what extent does the frequency of construction reuse contribute to the during-training fluency changes in the blocked and interleaved practice conditions?
3. To what extent does the frequency of construction reuse contribute to the pretest−posttest fluency changes in the blocked and interleaved practice conditions?

With regard to RQ1, it was hypothesized that blocked practice would induce a greater propensity for construction reuse than interleaved practice. The superiority of blocked practice for fluency development could be ascribed to the reuse of the same (or highly similar) constructions during task repetition. This link is postulated because immediate repetition of the same task would temporarily ease the retrieval of the same (or highly similar) constructions due to the priming effects (Bock & Griffin, 2000; Jacobs, Cho, & Watson, 2019).

In addressing RQ2 and RQ3, the aim was to further explore the extent to which reuse frequency would be related to fluency development both during- and post-training. Given a systematic relationship reported by N. de Jong and Tillman (2018), we expect a significant correlation between reuse frequency and fluency changes both during training and between pretest and posttest. Due to the paucity of previous research on blocked and

interleaved task repetition, no strong prediction could be made. Nonetheless, construction reuse was expected to influence some aspects of fluency under both practice conditions, such as syllable duration and mid-clause pause duration, for which Y. Suzuki (2021) reported greater improvement in the blocked practice condition. The systematic relationships, if found, would lend support to the assumption that construction reuse contributes to their proceduralization, which underlies fluency development through task repetition.

Methods

Participants

The study sample comprised of 50 L2 English learners studying at a Japanese university (aged 18−22 years) who were recruited through announcements in their regular EFL classes. Their English proficiency was estimated to fall between A2 (elementary) and B1 (intermediate) level on the Common European Framework of Reference for Languages (CEFR) benchmark. They were randomly assigned to either a blocked task repetition (n = 24) or an interleaved task repetition (n = 26) condition. No significant difference between the English proficiency levels of students assigned to the two groups was found based on an objective proficiency test called the Junior English Minimal English Test (Goto, Maki, & Kasai, 2010).

Materials

Training materials. Three prompts (Bicycle, Tiger, and Race) were used for oral narrative practice. These prompts were originally developed by Heaton (1996) and all three picture stories and the guiding questions were adopted from previous research on L2 fluency (N. de Jong & Tillman, 2018). Each prompt comprised of six-panel picture stories based on a tight sequential structure with similar narrative form (involving little causal reasoning) and number of elements (e.g., main characters, locations).

Pretest and posttest. In order to assess fluency development, two different six-panel picture stories (Street and Airport) were used in the pretest and posttest. Both stories also had a tight sequential structure with a similar narrative layout involving little causal reasoning (i.e., a thief steals the main character's purse/suitcase, and another main character helps capture the thief). Each of the prompts involved three main characters (thief, victim, and helper) at a different location (street and airport). The picture prompts used for pretest and posttest were counterbalanced. All instruments are available in the IRIS digital repository of data collection instruments (Mackey & Marsden, 2016) and are also provided as Supplementary Information for External Review Only.

Procedure. As shown in Figure 1, a pretest−training−posttest design was used in

the present study. One week prior to the training session, participants took a pretest in a computer lab. After they were randomly assigned to one of the experimental conditions (blocked or interleaved practice), they engaged in a three-day fluency training program outside the lab (Day 1, 2, and 3) by following the instructions and recording their narratives using a digital recorder. They were told that the objective of this intervention is to improve their general speaking skills. Participants assigned to the blocked practice condition performed the same narrative task three times in a single day (e.g., AAA−BBB−CCC), whereas those in the interleaved practice condition performed three different narrative tasks on each of the three days (e.g., ABC−ABC−ABC). The order of the three prompts (i.e., Bicycle, Tiger, and Race) was counterbalanced for each participant. Both groups were allowed the same amount of time (i.e., 3 minutes) for each narration throughout the experiment. To ensure that the guidelines related to fluency training were followed, all participants received daily reminders from a research assistant. On the day following the last training session (Day 4), participants assigned to both groups took part in the posttest.

		Blocked	**Interleaved**
Week 1		Pretest	
Week 2	Day 1	AAA	ABC
	Day 2	BBB	ABC
	Day 3	CCC	ABC
	Day 4	Posttest	

Figure 1. Experimental schedules.

Analysis

Data coding. A total of 550 speech datasets, derived from pretest (50 learners), training (50 learners × 9 deliveries), and posttest (50 learners), were coded by three trained coders. The unpruned transcripts were prepared and further coded using PRAAT (Boersma & Weenink, 2016) with the assistance of the script for automatically detecting pauses (N.H. de Jong & Wempe, 2009). Seven fluency measures were derived from the speech data to capture multiple dimensions of fluency (Table 1).

Table 1

List of Fluency Measures

Category	Measure	Operationalization
Speed fluency	Mean syllable duration	Inverse score of articulation rate (i.e., number of syllables per minute of speech, excluding pauses)
Breakdown fluency	Mid-clause pause duration	Mean duration of mid-clause filled and unfilled pauses
	Clause-final pause duration	Mean duration of clause-final filled and unfilled pauses
	Mid-clause pause frequency	Number of mid-clause filled and unfilled pauses per minute
	Clause-final pause frequency	Number of clause-final filled and unfilled pauses per minute
Repair fluency	Repetition frequency	Number of self-repetitions per minute
	Repair frequency	Number of reformulations and replacements per minute

Note. Instead of articulation rate, in correlation analyses, mean syllable duration (i.e., the inverse score of articulation rate) was used to align the direction of the score with those pertaining to the other fluency measures. Hence, for all fluency measures included in the current analyses (speed, breakdown, and repair fluency), smaller values indicate higher utterance fluency and vice versa.

These seven measures related to speech, breakdown, and repair fluency (Housen & Kuiken, 2009; Skehan, 2009). For speed fluency, mean syllable duration (the inverse of articulation rate) was computed (N. H. de Jong et al., 2013), whereas four pause-related indices (mid-clause and clause-final pause duration and frequency) were computed for breakdown fluency. Pauses were defined as the filled and unfilled (silent) pauses lasting at least 200 ms (N. de Jong & Perfetti, 2011) and were further coded as mid-clause pauses (i.e., within the Analysis of Speech [AS] unit, Foster, Tonkyn, & Wigglesworth, 2000) and clause-final pauses (i.e., at the boundary of the AS unit). Mid-clause pauses indicate linguistic (e.g., lexical and syntactic) breakdown, whereas clause-final pauses presumably reflect conceptualization, including content planning (N. H. de Jong, 2016; Kahng, 2018; Lambert et al., 2020). For repair fluency, repetition and repair

frequency were counted. Repetition refers to the number of self-repetitions (e.g., the man hit . . . hit the tiger's head), presumably indicating disfluency and/or a coping strategy, whereby speakers are trying to buy time for linguistic encoding. In the context of the current investigation, repair frequency refers to both overt reformulations (e.g., the poster which caution the…which tell us the tiger is so danger) and replacements (e.g., feel sleep . . . sleepy; owner tell . . . told). These self-repair behaviors are considered to reflect the degree to which learners' attentional resources are directed to monitoring their speech and reformulating initially encoded language (Kormos, 1999; Lambert et al., 2020).

Computation of similarity score between repeated task performances (cosine similarity). The measurement of constructional reuse adopted by N. de Jong and Tillman (2018) was used in the present study, which involved computing cosine similarity within pairs of task performances (i.e., transcribed texts). Following N. de Jong and Tillman's procedure, the pruned transcripts were prepared in CHAT transcription format (MacWhinney, 2000) after removing fillers, false starts, repetitions and corrected words. The analysis consisted of two steps: (a) extraction of constructions and (b) calculation of cosine similarity between participants' repeated narrations.

In the first step, three types of constructions were extracted from transcripts using CLAN program (MacWhinney, 2000): lexical unigram (individual lexical items), lexical trigram (contiguous sequences of three lexical words), and POS trigram (contiguous sequences of part of speech comprising of three words). For example, utterance *and a car was also running behind the bicycle* was parsed to generate a list of three levels of construction.

- Lexical unigram: and; a; car; was; also; running; behind; the; bicycle
- Lexical trigram: and a car; a car was; car was also; was also running; also running behind; running behind the; behind the bicycle
- POS trigram: conjcoo det n; det n aux; n aux adv; aux adv part; adv part prep; part prep det; prep det n

In the second step, the lists of extracted constructions were used to calculate cosine similarities between participants' repeated narrations using the python code that de Jong and Tillman (2018) made available for researchers (https://bitbucket.org/philtillman/fluencysimilarity/src/master/). The similarity scores between repeated tasks were computed as illustrated in Figure 2. In the blocked condition, narrations related to the same prompt performed on the same day (e.g., Day 1−1 & Day

1−2, Day 1−2 & Day 1−3, Day 1−1 & Day 1−3) were compared. In the interleaved condition, the same procedure resulted in comparisons across three training days (e.g., Day 1−1 & Day 2−1, Day 1−1 & Day 3−1, Day 2−1 & Day 3−1). This strategy allowed us to establish whether practice condition influenced the degree to which the participants reused the same constructions during task repetition based on the same prompt. An illustrative example of similarity score calculation is presented in Table 2.

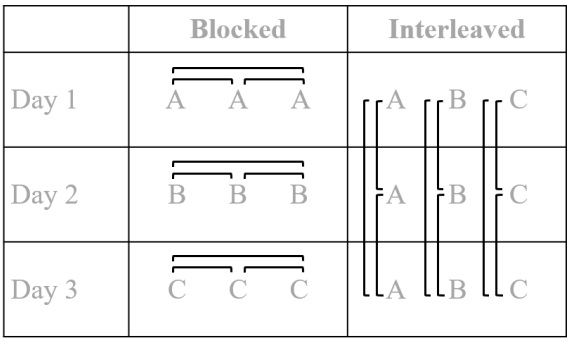


Figure 2. Comparisons of training performance for computing the similarity scores for blocked and interleaved condition.

Table 2

Illustrative Example of Similarity Score Calculation

[Unigram]

The similarity score (between the first and second performance) = .894

The similarity score (between the second and third performance) = 898

The similarity score (between the first and third performance) = .861

Performance	Lexical Unigram
1st	and; a; car; was; also; running; behind; the; bicycle
2nd	and; the; car; was; also; running; behind; the; bicycle
3rd	and; the; car; was; also; running; behind; the; boy

[Lexical Trigram]

The similarity score (between the first and second performance) = .858

The similarity score (between the second and third performance) = .805

The similarity score (between the first and third performance) = .669

Performance	Lexical Trigram
1st	and a car; a car was; car was also; was also running; also running behind; running behind the; behind the bicycle
2nd	and the car; the car was; car was also; was also running; also running behind; running behind the; behind the bicycle
3rd	and the car; the car was; car was also; was also running; also running behind; running behind the; behind the boy

[POS Trigram]
The similarity score (between the first and second performance) = 1
The similarity score (between the second and third performance) = 1
The similarity score (between the first and third performance) = 1

Performance	POS Trigram
1st	conjcoo det n; det n aux; n aux adv; aux adv part; adv part prep; part prep det; prep det n
2nd	conj:coo det n; det n aux; n aux adv; aux adv part; adv part prep; part prep det; prep det n
3rd	conj:coo det n; det n aux; n aux adv; aux adv part; adv part prep; part prep det; prep det n

Note. For more details on the similarity score computation method, see N. de Jong and Tillman (2018) and Appendix S2 in Online Supplementary File.

Statistical analysis. To address RQ1, independent samples *t*-tests were conducted to compare the similarity scores of blocked and interleaved practice groups. For this purpose, the similarity score was aggregated (averaged) across all three prompts (see Figure 2 above) for each participant. According to the Shapiro-Wilk test of normality, all three similarity scores were normally distributed for each group ($p > .10$). The Levene's tests indicated that the homogeneity of variances assumption was violated for the lexical and POS trigrams. The Welch *t*-tests were thus conducted for these two indices. According to the L2-specific research benchmark (Plonsky & Oswald, 2014), the magnitude of effect size (*d*) was interpreted as small (.40), medium (.70), or large (1.00).

To answer RQ2 and RQ3, a series of correlation analyses was conducted to examine the associations between similarity scores and fluency changes during the training phase. Fluency change scores were composite scores of gains from the first to

the third performance related to the three prompts, as well as those from the pretest to the posttest. Because the sample size was relatively small and the visual inspection of the scatterplots showed potential outliers in some cases (see Appendix S4 and S5 in Online Supplementary File), a robust statistical procedure (Wilcox, 2017) was adopted to estimate the strengths of associations between cosine similarity and fluency measures. Using the "pbcor" function in the WR2 package (Mair & Wilcox, 2020) in R (R development Core Team, 2019), we computed the percentage bend correlation. This method addresses potential outliers by mitigating their effects on correlation coefficient estimates. Coupled with the bootstrapping procedure, this robust method controls for Type I error and provides better correlation coefficient estimates for relatively small samples than the traditional methods such as Pearson's correlations (Pernet, Wilcox, & Rousselet, 2013).

The strengths of percentage bend correlation coefficients were interpreted based on the L2-specific research benchmark (Plonsky & Oswald, 2014) as small (r = .25), medium (r = .40), or large (r = .60). Additionally, in their study, which is most relevant to the goals of the present investigation, N. de Jong and Tillman (2018) computed a grand mean correlation coefficient between the similarity scores and fluency changes of .28.[1] This value corresponds to small effect size according to the Plonsky and Oswald's (2014) benchmark. Hence, this correlation magnitude was interpreted as minimally meaningful (.25 < r < .39) in this study.

Results

Similarity Scores between Blocked and Interleaved Task Repetition Conditions

Figure 3 illustrates the similarity scores between the blocked and interleaved practice conditions.[2] For all three indices (lexical unigram, lexical trigram, and POS trigram), blocked practice led to higher similarity scores than interleaved practice. According to the independent samples t-test, the lexical unigram similarity score was significantly higher in the blocked practice condition than in the interleaved practice condition with a large effect size, $t(48) = 4.37$, $p < .001$, $d = 1.24$, 95% confidence interval (CI) [0.62, 1.84]. In addition, blocked practice resulted in a significantly higher lexical

[1] This correlation coefficient was obtained by averaging the correlation coefficients (i.e., absolute value) across fluency measures (i.e., phonation time and mean syllable duration), the training days (i.e., Day 1, 2, and 3), and construction types (i.e., lexical unigram, lexical trigram, and POS trigram).

[2] No significant group differences were detected in the total number of syllables or lexical diversity (i.e., the measures of textual lexical diversity, Zenker & Kyle, 2021) across the training performances ($p > .05$).

trigram similarity score than interleaved practice with a large effect size, $t(35.27) = 4.19$, $p < .001$, $d = 1.20$, 95% CI [0.57, 1.81]. Similarly, for the POS trigram, blocked practice resulted in higher similarity score than interleaved practice with a large effect size, $t(37.52) = 2.96$, $p = .01$, $d = 0.85$, 95% CI [0.25, 1.43].

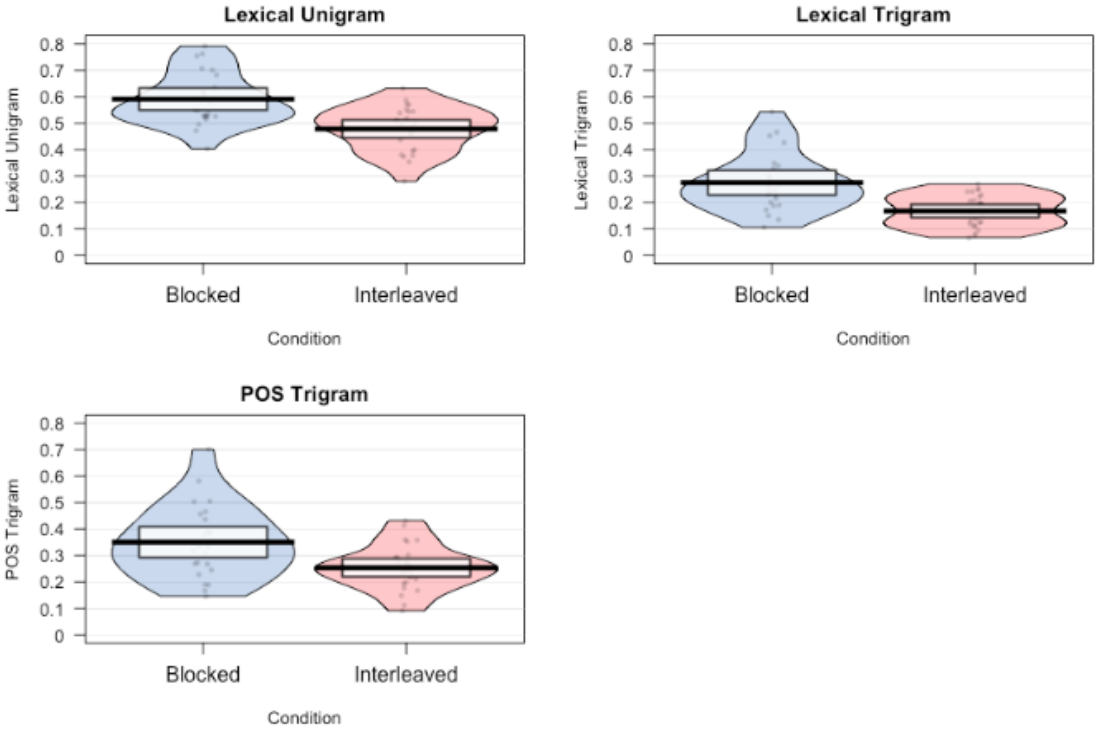


Figure 3. Similarity scores of blocked and interleaved task repetition conditions.
Note. Boxes indicate 95% confidence intervals (CIs). Numerical values for mean, SD and 95% CIs are presented in Appendix S2 in Online Supplementary File.

Relationship between Similarity Scores and During-training Fluency Change

Table 3 presents correlations between similarity scores and fluency changes during the training phase. No significant correlations (or CIs including zero) were detected in the blocked practice condition. Yet, based on the benchmark established for this study (see the section on statistical analysis), there were some meaningful, albeit small, correlations. Specifically, mid-clause pause duration was negatively related to all three construction types ($-.27 < r < -.34$) and mid-clause pause frequency was also negatively related to lexical unigrams ($r = -.31$, $p = .13$). Moreover, negative correlations were noted between self-repetition and similarity scores for lexical unigrams and trigrams ($r = .29$ and $-.27$, $p = .17$ and $.21$, respectively).

In contrast, one significant and stronger correlation was detected in the interleaved practice condition, whereby mean syllable duration was (marginally) significantly

correlated with lexical and POS trigrams (r = .37, p = .06; r = .49, p = .01). This unexpected positive correlation direction indicated that greater construction reuse led to lower speed fluency in the blocked practice condition. The coefficient between mean syllable duration and POS trigram was the only medium effect size and its CI did not overlap zero. It is also noteworthy that clause-final pause duration was positively related to lexical and POS trigrams (r = .27, p = .18; r = .30, p = .13), suggesting that greater reuse was linked to *longer* clause-final pauses. Conversely, negative small correlations were noted between mid-clause pause frequency and lexical unigram and POS trigram (r = -.33, p = .10; r = -.25, p = .22).

Table 3

Correlations between Similarity Scores and During-training Fluency Change

	Blocked			Interleaved		
	Lexical unigram	Lexical trigram	POS trigram	Lexical unigram	Lexical trigram	POS trigram
Mean syllable duration	-.12	.06	-.08	.09	.37+	.49*
	[-.35, .52]	[-.49, .36]	[-.39, .47]	[-.45, .27]	[-.68, .02]	[-.76, -.03]
Mid-clause pause duration	-.28	-.27	-.34	.20	.09	.16
	[-.61, .15]	[-.61, .19]	[-.68, .09]	[-.24, .57]	[-.36, .51]	[-.36, .55]
Clause-final pause duration	.11	-.13	-.12	.16	.27	.30
	[-.36, .59]	[-.57, .31]	[-.57, .37]	[-.20, .44]	[-.24, .64]	[-.15, .65]
Mid-clause pause frequency	-.31	-.18	-.19	-.33	-.14	-.25
	[-.66, .07]	[-.56, .23]	[-.57, .22]	[-.69, .08]	[-.59, .32]	[-.62, .23]
Clause-final pause frequency	.19	.07	.12	-.14	-.11	-.07
	[-.24, .60]	[-.36, .51]	[-.29, .54]	[-.50, .31]	[-.49, .33]	[-.46, .37]
Repetitions	-.29	-.27	-.19	-.16	-.19	.05
	[-.60, .10]	[-.64, .19]	[-.59, .24]	[-.62, .33]	[-.59, .28]	[-.39, .51]
Repairs	-.09	-.09	-.06	-.23	-.12	-.01
	[-.50, .36]	[-.51, .40]	[-.48, .40]	[-.59, .24]	[-.56, .33]	[-.46, .48]

Note. +p < .10, *p < .05, **p < .01. See Appendix S4 in Online Supplementary File for scatter plots.

Relationship between Similarity Scores and Pretest−Posttest Fluency Change

As shown in Table 4, correlations between similarity scores and pretest−posttest fluency changes exhibited more contrasting patterns between blocked and interleaved

practice conditions than those indicated in the training data. In the blocked practice condition, magnitudes of all meaningful associations were small ($.25 < |r| < .36$), whereas those for the interleaved practice condition ranged from small to large ($.31 < |r| < .63$).

While the weak correlations related to the blocked practice condition were not statistically significant, these potentially meaningful associations of similarity scores were linked to breakdown fluency measures (i.e., mid-clause pause duration and frequency as well as clause-final pause duration). Somewhat unexpectedly, correlation coefficients of similarity scores for the mid-clause pause frequency were negative ($-.36 < r < -.33$) but were positive for the mid-clause pause duration ($.12 < r < .28$). Furthermore, positive weak correlations were detected between clause-final pause duration and similarity scores ($.23 < r < .29$).

In contrast, significant and stronger correlations were noted between similarity scores and four fluency measures in the interleaved practice condition. First, although (unexpected) positive correlations were found between syllable duration and the similarity scores for the training data, analyses pertaining to the pretest−posttest data revealed that lexical trigram was negatively related to mean syllable duration ($r = -.40$, $p = .045$). This finding suggests that grater reuse of trigrams during training led to higher speed fluency on the posttest. Second, negative medium-to-large correlations were found for mid-clause pause frequency. The magnitudes of associations did not include zero in the confidence intervals and were the smallest for lexical unigrams, increasing for lexical trigrams and finally for POS trigrams ($r = -.43$, $-.60$, and $-.63$, respectively). Last, meaningful positive correlations were noted between similarity scores and mid-clause ($.31 < r < .35$) and clause-final pause duration ($.40 < r < .46$).

Table 4

Correlations between Similarity Scores and Pretest−Posttest Change

	Blocked			Interleaved		
	Lexical unigram	Lexical trigram	POS trigram	Lexical unigram	Lexical trigram	POS trigram
Mean syllable duration	.13	.09	.22	-.06	-.40*	-.15
	[-.53, .38]	[-.46, .36]	[-.61, .23]	[-.33, .47]	[-.01, .72]	[-.27, .62]
Mid-clause pause duration	.26	.12	.28	.35+	.36+	.31
	[-.22, .69]	[-.32, .49]	[-.19, .66]	[-.14, .67]	[-.09, .73]	[-.14, .69]
Clause-final pause duration	.29	.25	.23	.10	.40*	.46*
	[-.13, .63]	[-.19, .61]	[-.25, .61]	[-.27, .50]	[-.04, .71]	[-.01, .74]
Mid-clause pause frequency	-.36+	-.33	-.35+	-.43*	-.60**	-.63**
	[-.74, .07]	[-.70, .11]	[-.73, .06]	[-.68, -.05]	[-.85, -.23]	[-.81, -.32]
Clause-final pause frequency	-.14	-.03	-.04	.01	-.14	-.17
	[-.58, .33]	[-.46, .41]	[-.48, .39]	[-.40, .40]	[-.52, .30]	[-.54, .28]
Repetitions	-.17	-.21	-.13	-.05	-.04	.08
	[-.51, .21]	[-.58, .22]	[-.53, .28]	[-.46, .42]	[-.48, .38]	[-.42, .55]
Repairs	.20	.00	.00	-.01	-.06	-.23
	[-.23, .59]	[-.42, .44]	[-.42, .42]	[-.40, .38]	[-.46, .42]	[-.59, .25]

Note. +$p < .10$, *$p < .05$, **$p < .01$. See Appendix S5 in Online Supplementary File for scatter plot

Discussion

Blocked Practice Induces More Repetition of Concrete and Abstract Constructions

Findings related to RQ1 supported the hypothesis that blocked practice induces a greater degree of constructional recycling than interleaved practice. This observation could partially explain the advantage of blocked practice over interleaved practice for promoting utterance fluency (particularly, speed and breakdown fluency) established by Y. Suzuki (2021). Blocked and interleaved practice might have exhibited different degrees of construction self-priming (Bock & Griffin, 2000; Jacobs et al., 2019). Because the same prompt was narrated three times on the same day in the blocked practice condition, L2 learners were more likely to reuse pre-activated linguistic constructions than those assigned to the interleaved practice condition in which the same prompt was narrated on three consecutive days.

Findings yielded by previous studies indicate that the extent to which constructions are recycled during task repetition is influenced by the task type (N. de Jong & Perfetti, 2011) and the time allocated for repeated task performance (Boers, 2014; Thai & Boers, 2016). The current investigation revealed that task sequence (i.e., practice schedule) is another factor that influences L2 learners' propensity for construction reuse. In other words, manipulating the task sequence without changing the task type or increasing time pressure can systematically elicit repetition of constructions at both concrete and abstract levels.

Constructions that were recycled by the study participants were not only concrete (lexical) unigrams and trigrams but also abstract (POS) trigrams. Even if utterances involve superficially different trigrams, the schematic patterns (POS trigrams) may still encode abstract meaning depicted in the cartoons, including transitive events, relations among characters, objects, etc (Langacker, 2008). For instance, all L2 learners that took part in the current study used five types of POS trigrams during the training phase (i.e., adjective−noun−verb, determiner−adjective−noun, verb−determiner−noun, pronoun−verb−determiner, and verb−determiner−noun). When these abstract grammatical constructions were used repeatedly for expressing a similar propositional meaning (e.g., a construction such as verb−determiner−noun expressing a transitive event), the abstract construction might have been gradually committed to learner's memory (Langacker, 2008; Schmid, 2017). Greater frequency of active retrieval of the same constructions may be linked to proceduralization or potentially be a sign of incipient automatization, which is facilitated by extended repeated practice (DeKeyser, 2018, 2020).

Nevertheless, because the task-repetition intervention adopted in the current study involved mere repetitions of the same task, some portions of ungrammatical constructions (e.g., *was very regret* [used by five learners nine times in total], or *{he/man/driver} {is/was} glare* [used by nine learners 36 times in total]) were repeatedly used and could have contributed to the proceduralization (entrenchment) of ungrammatical constructions. Despite its effectiveness in promoting utterance fluency, this could be a major drawback of task repetition training (Boers, 2014; Thai & Boers, 2016). Thus, in order to assist the development of both fluency and accuracy, some form of accuracy enhancement (e.g., provision of models or corrective feedback) would be highly beneficial (Lynch, 2018; Tran & Saito, 2021).

Relationships between Constructional Recycling and Fluency Change

The goal of addressing RQ2 and RQ3 was elucidating the role of construction

reuse in the fluency changes during training and between pretest and posttest under the two practice conditions. The correlation patterns summarized in Table 5 indicate that the meaningful association magnitudes ranged from weak to large in effect size, and due to the small sample sizes, several meaningful coefficients were not statistically significant, with the 95% CIs including zero. Given these limitations, the reported findings should be interpreted with caution.

Table 5

Summary of Relationships Between Reuse and Fluency Changes

Effect size	Blocked			Interleaved		
	Unigram	**Lexical Trigram**	**POS trigram**	**Unigram**	**Lexical Trigram**	**POS trigram**
Training						
Small (.25 ≦ *r* ≦.39)	Mid pause dur. (↓) Mid pause freq. (↓) Repetition (↓)	Mid pause dur. (↓) Repetition (↓)	Mid pause dur. (↓)	Mid pause freq. (↓)	Syllable dur. (↑) Final pause dur. (↑)	Mid pause freq. (↓) Final pause dur. (↑)
Medium (.40 ≦ *r* ≦.59)						Syllable dur. (↑)
Large (*r* ≧.60)						
Pretest-Posttest						
Small (.25 ≦ *r* ≦.39)	Mid pause dur. (↑) Mid pause freq. (↓) Final pause dur. (↑)	Mid pause freq. (↓) Final pause dur. (↑)	Mid pause dur. (↑) Mid pause freq. (↓)	Mid pause dur. (↑)	Mid pause dur. (↑)	Mid pause dur. (↑)
Medium (.40 ≦ *r* ≦.59)				Mid pause freq. (↓)	Syllable dur. (↓) Final pause dur. (↑)	Final pause dur. (↑)
Large (*r* ≧.60)					Mid pause freq. (↓)	Mid pause freq. (↓)

Note. Upward arrows indicate a higher similarity score corresponding to an increase in fluency score (supposedly reflecting less fluent speech), whereas downward arrows indicate a higher similarity score corresponding to a decrease in fluency score (supposedly reflecting more fluent speech).

During the training phase (RQ2), the magnitudes of associations between reuse and fluency changes were small in the blocked practice condition and small-to-medium in the interleaved practice condition. These results are consistent with N. de Jong and Tillman's (2018) findings linking the similarity score to the fluency changes during training (the average magnitude of association was about .30).

Meaningful relationships were also noted for the pretest−posttest fluency changes assessed by a new transfer test (RQ3). Since single lexical unit reuse and fluency transfer to a new task was only examined by N. de Jong and Perfetti (2011), their prior finding was extended in the current study to the potential links with the reuse of constructions that are larger (lexical trigrams) and more abstract (POS trigrams). Based on the L2-specific benchmark (Plonsky & Oswald, 2014), the magnitudes of meaningful associations were larger (including medium-to-large effect) in the interleaved practice condition than in the blocked practice condition (small effect).

Among the evaluated fluency measures, speed (mean syllable duration) and breakdown fluency (mid-clause pause frequency and duration as well as clause-final pause duration, but not clause-final pause frequency) were found to be associated with constructional recycling.[3] In what follows, these meaningful patterns will be discussed in relation to fluency measure types. First, a systematic pattern for mean syllable duration is interpreted. Second, we focus on mid-clause pause frequency and duration, as a part of breakdown fluency measures, which were consistently related to construction reuse in both practice conditions. Finally, another breakdown fluency measure, clause-final pause duration, is discussed.

Higher constructional recycling led to initially longer but shorter mean syllable duration in the interleaved condition. In the present study, reuse of lexical and POS trigrams was positively related to mean syllable duration (r = .37 and .49, respectively) in the interleaved practice condition only. These positive relationships indicate that learners who used the same trigrams uttered their sentences more slowly at the end of the training phase. These findings may suggest that, when repeating the same task on the following day (in the interleaved practice condition), more effort was required

[3] For repair fluency, lexical unigram and trigram reuse frequency was weakly related to self-repetition change during the training phase in the blocked practice condition only. This finding indicates that learners who repeated the lexical unigrams and trigrams more often across task repetitions made fewer self-repetitions. Because self-repetition may reflect a coping strategy to compensate for disruptions in linguistic encoding, using identical unigrams and trigrams might have reduced the need for self-repetition, especially when the same task was immediately repeated in the blocked practice condition. This weak effect should be interpreted with extra caution, however.

to retrieve the previously used trigrams (a more complex construction), resulting in longer mean syllable duration. It is also worth noting that the systematic relationship between the reuse frequency and mean syllable duration was found only for trigrams, not for unigrams. This is to be expected, given that retrieving single lexical items was presumably not too cognitively demanding (challenging) to slow down the articulation speed even in the interleaved practice condition.

However, the pretest−posttest change indicated that reuse of lexical trigrams negatively contributed to the mean syllable duration with medium effect size (r = -.40). This finding could be ascribed to more efficient processing by learners who engaged in effortful retrieval of lexical trigram during training. This supposition is consistent with the view that suboptimal performance during the training phase, as a result of effortful and challenging processing, may prompt learners to practice more, leading to better learning in the end (Bjork, 1994; Suzuki, Nakata, & DeKeyser, 2019).

Nevertheless, a direct connection between the effortful retrieval of specific trigrams during the training and the more efficient processing of trigrams at the posttest may not be tenable. This is because the trigram used during the training were unlikely to be used for the posttest performance with different picture prompt. Perhaps, an "indirect" link could be presupposed. The effortful practice involving repeated use of the same trigrams might have contributed to the enhancement of the formulation stage in the speech model (e.g., general retrieval processes of constructions, encoding of schematic constructions), which could have resulted in the posttest performance improvement.

The pattern found in the interleaved practice condition was not evident in the blocked practice condition. In other words, performing the same task repeatedly on the same day might not have been cognitively demanding for learners, allowing them to effortlessly retrieve previously used constructions. Because the constructions used in the first performance were presumably activated or primed during subsequent performances (Bock & Griffin, 2000; Jacobs et al., 2019), learners might have been able to use the same constructions without slowing down their utterance speed.

Higher constructional recycling led to fewer mid-clause pause frequency but longer duration. Irrespective of the practice condition, L2 learners who repeated the same constructions made shorter and less frequent mid-clause pauses during training. Because mid-clause pauses is related to linguistic formulation, including lexical and syntactic encoding (N. H. de Jong, 2016; Kahng, 2018; Lambert et al., 2020), recycling of the same constructions throughout task repetition seems to facilitate linguistic encoding (e.g., proceduralization, indicated by more efficient retrieval of constructions).

Comparisons between the pretest and posttest results revealed that construction

reuse was more strongly related to (lexical and POS) trigrams than unigrams in the interleaved practice condition, suggesting that recycling more complex constructions facilitated proceduralization. Furthermore, greater construction reuse appears to have exerted both positive and negative influence on learners' reliance on mid-clause pauses on the posttest relative to the pretest. Specifically, the learners who recycled constructions more paused *less* frequently within the clausal boundary but for *longer* periods. This contrasting pattern of pause frequency and duration may indicate potential developmental signature of L2 fluency. It is speculated that learners who are consolidating their knowledge of L2 constructions through repetition can succeed in linguistic encoding more frequently which would result in fewer pauses). However, their encoding system is still developing and any shortfalls may be compensated by longer mid-clause pauses. This trade-off between mid-clause pause frequency and duration seems consistent with ad-hoc correlation analyses indicating that these aspects are significantly and negatively correlated (interleaved: $r = -.67$, $p < .001$; blocked: $r = -.47$, $p = .02$).

Higher constructional recycling led to longer clause-final pause duration. Clause-final pauses are presumably related to content planning and conceptualization (Kahng, 2018; Kormos, 2006; Saito, Ilkan, Magne, Tran, & Suzuki, 2018) and thus seem to reflect different cognitive process from mid-clause pause phenomena. The association between reuse and clause-final pause duration was evident during the training phase for the interleaved, but not blocked, practice condition. Due to the 1-day intervals between consecutive performances of the same task, learners in the interleaved practice condition had to re-engage in the conceptual planning and (re)form a new proposition of the preverbal message. When they used the same lexical and POS trigrams, they could have paused at the appropriate clause boundary for a longer period to restructure and elaborate the narrative events they attempted to recall from the previous performance. For instance, one learner in the interleaved practice condition paused longer at the clausal boundary (as well as fewer mid-clause pauses) in the third performance relative to the second performance (silent mid-clause pause = MP, silent clause-final pause = FP, the numeric values indicate seconds):

Second performance:
he(MP 0.6)a tall boy is(MP 1.3)taking nap (FP 2.1) when the tall boy(MP 1.1)wake(MP 0.4) woke (MP 0.4) up (FP 1.5) it's too late

Third performance:
(FP 4.5) he was sleeping for(MP 2.4)a long time (FP 3.7) when he waked up (FP 2.0) he

was in the last

When inspecting the first part of the utterances, no clause-final pause was made in the second performance. In contrast, in the subsequent performance, the 4.5-second pause preceded the more elaborated utterance (*he was sleeping for a long time*) as well as the coherent pronoun (*he*) use. In addition, the clause-final pause was longer (2.1 vs. 3.7 seconds) prior to the adverbial clause (*when* clause), which could have helped eliminating the mid-clause pauses in the third performance. As a more in-depth analysis is beyond the scope of this study, this interpretation remains speculative and merits further exploration.

According to the posttest results, higher construction reuse was related to longer clause-final pause duration for both practice conditions, but the effect size was larger (i.e., medium effect size) in the interleaved practice condition. Additionally, in the interleaved practice condition, clause-final pause duration was related to the reuse of more complex constructions (i.e., lexical and POS trigrams) than lexical unigrams. As discussed in relation to the training results, the link between greater reuse and longer (rather than shorter) clause-final pause may appear somewhat puzzling. However, *longer* clause-final pause duration may not necessarily be unfavorable from the fluency development perspective. Indeed, when learners in the interleaved practice condition tried to use the same constructions that are complex (trigrams, rather than unigrams), they had to expend more cognitive resources for conceptualizing propositions (due to the 1-day lag). Consequently, they could have become more adept in pausing at the appropriate clausal boundary rather than within the clauses, which may suggest that they could have engaged more in encoding of more complex (grammatical) constructions than simpler lemma retrieval.

Suggestions for Future Research

As the current study is subject to several limitations, these can be addressed in future research in this domain. First, given the relatively small sample size, many of the correlation coefficients—considered meaningful in effect size—were not statistically significant. Although a robust statistical approach was adopted to reduce Type I error incidence, the current findings should be interpreted with caution. A replication study with a larger sample of L2 learners with different backgrounds is thus needed to attest the current and previous findings (e.g., N. de Jong & Tillman, 2018; Thai & Boers, 2016). More specifically, a priori power analysis was conducted using *pwr* package in R (Champely et al., 2020) to estimate the required sample sizes for future research. The results indicated that the sample sizes to achieve the statistical power of .80 for the weak

correlation coefficients found in the current study (.25 < *r* < .37) ranged from 54 to 122. This information may be informative for planning future replication and extension research. However, because a massive amount of labor-intensive coding is necessarily for this type of research, analyzing 100+ participants' speech data may not be realistically feasible in a single study led by one group of researcher(s). Because the analysis procedures regarding fluency and cosine similarity measures can be implemented using the openly available scripts (see the Analysis section), making a consorted effort such as multi-site replication with multiple research groups (Morgan-Short et al., 2018) may be a promising idea.

Second, while seven fluency measures were used in the current study, one of the reviewers pointed out that (mid-clause and clause-final) pauses could have been further divided into filled and silent pauses. Because silent and filled pausing behaviors may not necessarily be caused by the same speech production mechanisms, using separate measures would be useful in future research (see, e.g., S. Suzuki, 2021; Tavakoli, Nakatsuhara, & Hunter, 2020). Additionally, some intriguing trade-off patterns between pause frequency and duration were documented in the current study (see Appendix S6 in Online Supplementary File). From a developmental perspective, it is not entirely implausible that the duration increases when the frequency decreases (and vice versa) while the overall breakdown fluency is stable. To our knowledge, no L2 fluency intervention research, except for the current project (Y. Suzuki, 2021), has examined this possibility. Possibly, using a different pause frequency measure such as pause ratio (i.e., the mean number of pauses divided by the total number of syllables) may also be useful in future research (S. Suzuki, 2021).

Third, because the L2 learners that took part in the current study performed oral narrative monologue using a six-picture frame cartoon, more open-ended task design such as topic monologue (N. de Jong & Perfetti, 2011) and interactive information-gap speaking activities may be used in the future to examine the effects of reuse on L2 fluency development. Such task design may allow learners to use a larger variety of concrete linguistic constructions that could increase the importance of abstract-level analysis (e.g., POS trigram).

Fourth, the current intervention included only three task performances for each prompt. Thus, a more longitudinal research design with greater opportunities for repetition would be beneficial for establishing the effects of more intensive construction reuse on L2 fluency development as well as putative underlying proceduralization and possibly further automatization.

Fifth, the L2 learners in this study engaged in repeated practice with a broad aim

of improving their speaking proficiency. There may be other ways of providing useful instructions (e.g., narrate the story in a shorter time, describe the story in more detail) and feedback (e.g., directing learners' attention to linguistic accuracy) to fine-tune different aspects of speaking skills more effectively. We observed that one learner's reuse of the same trigram fluctuated in terms of tense and aspect (1st performance: "boy continued riding"; 2nd performance: *"boy continue ride"; 3rd performance: *"boy continue riding"). It is speculated that this learner might have been working hard to speak more accurately (despite the lack of instructions on accuracy enhancement) possibly using their declarative knowledge of present progressive and past tense for proceduralization (yet, this learner ended up speaking the partially inaccurate sentence at the third performance).

Last, because L2 speech was analyzed in terms of fluency only, no measures were adopted to ascertain whether the recycled constructions were accurate/appropriate. Some learners in the current study repeated the same trigram constructions incorrectly in all three performances (e.g., *"tiger was unconsciousED", *"glared TO him [the]"). Although reuse may promote proceduralization, if the recycled constructions are inaccurate, their reuse may be counterproductive (see Boers, 2014 for discussion). It is thus premature to make any strong practical recommendations based on the results obtained in this study, but immediate, blocked practice does seem to induce systematic linguistic construction recycling. However, mere repetition of the same task during the intervention may be considered to promote fluency development at best. With this limitation in mind, accuracy enhancement techniques (e.g., provision of narration models and corrective feedback) would be highly beneficial (Lynch, 2018; Tran & Saito, 2021), or even obligatory under some circumstances, for promoting the reuse and proceduralization of target-like linguistic expressions.

Conclusions

The aim of the present study was enhancing the current understanding of L2 learners' reliance on construction reuse during task repetition and its role in L2 fluency development. The findings reported in this work indicate that blocked practice can systematically enhance linguistic construction recycling in task-repetition practice. Immediately repeating the task seems beneficial in promoting proceduralization of constructions that vary in size (unigrams and trigrams) and abstractness (POS trigrams). In contrast, participants assigned to the interleaved practice condition were less reliant on previously used constructions. Nonetheless, those who reused lexical and POS trigrams more frequently tended to articulate faster and made fewer mid-clause pauses, possibly at the expense of longer mid-clause and clause-final pauses. Given that proceduralization

of construction should allow for the gained knowledge to be transferred to a new context, the current NLP-based analyses may indicate that the transfer effects of fluency training are mediated by systematic repetition of both concrete and abstract constructions. In sum, a closer look into repetition of constructions in task repetition allowed us to elucidate the underlying cognitive restructuring process of L2 speech fluency.

Acknowledgements

This study was supported by Grant–in–Aid for Scientific Research (KAKENHI) from Japan Society for the Promotion of Science (JP18K12470). We would like to express our deepest gratitude to Atsushi Miura, Misaki Kuratsubo, Miyu Koyama, Taeko Hosaka, Kazuma Arai for their dedicated assistance in data collection and coding.

Author Bio

Yuichi Suzuki is Associate Professor at Kanagawa University. His research focuses on explicit and implicit learning and knowledge, skill acquisition theory, automatization, and optimization of L2 practice. He received the Valdman's Award from Studies in Second Language Acquisition (2017) and the IRIS Replication Award (2018).

Masaki Eguchi is a PhD candidate at the Department of Linguistics at the University of Oregon. His research interests focus on the constructionist, functional approaches to L2 lexico-grammar learning, triangulating corpus, psycholinguistic, and classroom research. He is also interested in applications of educational measurement and statistics in applied linguistics research.

Nel de Jong was a lecturer of Linguistics at the University of Amsterdam. Her research interests include second language acquisition, oral fluency development, and language assessment. She is now an independent language consultant.

References

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

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General, 129*, 177-192. doi:10.1037//0096-3445.129.2.177

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal, 45*, 221-235. doi:10.1177/0033688214546964

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. Version 6.0.14. Retrieved from http://www.praat.org/

Bygate, M. (2018). *Learning language through task repetition* Amsterdam, the Netherlands: John Benjamins Publishing Company.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., . . . De Rosario, H. (2020). Pwr: Basic functions for power analysis (Version 1.3).

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning, 61*, 533-568. doi:10.1111/j.1467-9922.2010.00620.x

de Jong, N., & Tillman, P. C. (2018). Grammatical structures and oral fluency in immediate task repetition: Trigrams across repeated performances. In M. Bygate (Ed.), *Language learning through task repetition* (pp. 43-73). Amsterdam, The Netherlands: John Benjamins.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*, 113-132. doi:10.1515/iral-2016-9993

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*, 893-916. doi:10.1017/S0142716412000069

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods, 41*, 385-390. doi:10.3758/BRM.41.2.385

DeKeyser, R. M. (2018). Task repetition for language learning: A perspective from skill acquisition theory. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 27-42). Amsterdam, the Netherlands: John Benjamins Publishing Company.

DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed.,

pp. 83-104). New York, NY: Routledge.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*, 143-188. doi:10.1017/S0272263102002024

Ellis, N. C. (2009). Optimizing the input: Frequency and sampling in usage-based and form-focused learning. In C. Doughty & M. Long (Eds.), *The handbook of language teaching* (pp. 139-158). Oxford: Blackwell.

Ellis, N. C., & Wulff, S. (2020). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 63-82). New York, NY: Routledge.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354-375. doi:10.1093/applin/21.3.354

Goldberg, A. E. (1995). *Construction: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*: Oxford University Press, USA.

Goto, K., Maki, H., & Kasai, C. (2010). The minimal English test: A new method to measure English as a second language proficiency. *Evaluation & Research in Education, 23*, 91-104. doi:10.1080/09500791003734670

Heaton, J. B. (1996). *Composition through pictures*. Essex: Longman.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*, 461-473. doi:10.1093/applin/amp048

Jacobs, C. L., Cho, S.-J., & Watson, D. G. (2019). Self-priming in production: Evidence for a hybrid model of syntactic priming. *Cognitive Science, 43*, e12749. doi:10.1111/cogs.12749

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*, 809-854. doi:10.1111/lang.12084

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics, 39*, 569-591. doi:10.1017/S0142716417000534

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching & Research, 4*, 900-913. doi:10.4304/jltr.4.5.900-913

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning, 49*, 303-342. doi:10.1111/0023-8333.00090

Kormos, J. (2006). *Speech production and second language acquisition*. New York:

Routledge.

Lambert, C., Aubrey, S., & Leeming, P. (2020). Task preparation and second language speech production. *TESOL Quarterly, 55*, 331-365. doi:10.1002/tesq.598

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lynch, T. (2018). Promoting learning from second language speaking tasks. In V. Samuda, K. Van den Branden, & M. Bygate (Eds.), *TBLT as a researched pedagogy* (Vol. 12, pp. 213-234). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Mackey, A., & Marsden, E. (2016). *Advancing methodology and practice: The IRIS repository of instruments for research into second languages*: Routledge.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Transcription format and programs* (Vol. 1): Psychology Press.

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the wrs2 package. *Behavior Research Methods, 52*, 464-488. doi:10.3758/s13428-019-01246-w

Morgan-Short, K., Marsden, E., Heil, J., Issa II, B. I., Leow, R. P., Mikhaylova, A., . . . Szudarski, P. (2018). Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning, 68*, 392-437. doi:10.1111/lang.12292

Nergis, A. (2021). Can explicit instruction of formulaic sequences enhance L2 oral fluency? *Lingua, 255*, 103072. doi:10.1016/j.lingua.2021.103072

Pernet, C., Wilcox, R., & Rousselet, G. (2013). Robust correlation analyses: False positive and power validation using a new open source matlab toolbox. *Frontiers in Psychology, 3*, 606. doi:10.3389/fpsyg.2012.00606

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878-912. doi:10.1111/lang.12079

R development Core Team. (2019). R: A language and environment for statistical computing. Vienna, austria: R foundation for statistical computing. Retrieved from http://www.r-project.org/

Robinson, P., & Ellis, N. C. (2008). *Handbook of cognitive linguistics and second language acquisition*. New York: Routledge.

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics, 39*, 593-617. doi:10.1017/S0142716417000571

Schmid, H.-J. (2017). A framework for understanding linguistic entrenchment and its psychological foundations *Entrenchment and the psychology of language learning* (pp. 9-36). Washington, DC: De Gruyter Mouton.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics, 30*, 510-532. doi:10.1093/applin/amp047

Suzuki, S. (2021). *The multidimensionality of second language oral fluency: The interface between cognitive, utterance, and perceived fluency.* (Doctoral dissertation), Lancaster University, Lancaster, UK.

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning, 71*, 285-325. doi:10.1111/lang.12433

Suzuki, Y. (in press). Practice and automatization. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics*. New York: Routledge.

Suzuki, Y., Nakata, T., & DeKeyser, R. M. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal, 103*, 713-720. doi:10.1111/modl.12585

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*, 169-191. doi:10.1111/modl.12620

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Amsterdam: John Benjamins.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning, 70*, 506-547. doi:10.1111/lang.12384

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly, 50*, 369-393. doi:10.1002/tesq.232

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard university press.

Tran, M. N., & Saito, K. (2021). Effects of the 4/3/2 activity revisited: Extending Boers (2014) and Thai & Boers (2016). *Language Teaching Research, Early View*. doi:10.1177/1362168821994136

Tyler, A. E., & Ortega, L. (2018). Usage-inspired L2 instruction: Some reflections and a heuristic. In A. E. Tyler, L. Ortega, M. Uno, & H. Park (Eds.), *Usage-inspired L2*

instruction: Researched pedagogy (Vol. 49, pp. 315-321). Amsterdam: John Benjamins Publishing Company.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.): Academic Press.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*, 13-33.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. New York, NY: Continuum.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, 100505. doi:10.1016/j.asw.2020.100505