Studies in Second Language Acquisition **41** (2019), 999–1026
doi:10.1017/S0272263119000159

Research Article

PRACTICE IS IMPORTANT BUT HOW ABOUT ITS QUALITY?

CONTEXTUALIZED PRACTICE IN THE CLASSROOM

Masatoshi Sato*

Universidad Andres Bello

Kim McDonough

Concordia University

Abstract

This study explored the impact of contextualized practice on second language (L2) learners' production of *wh*-questions in the L2 classroom. It examined the quality of practice (correct vs. incorrect production) and the contribution of declarative knowledge to proceduralization. Thirty-four university-level English as a foreign language learners first completed a declarative knowledge test. Then, they engaged in various communicative activities over five weeks. Their production of *wh*-questions was coded for accuracy (absence of errors) and fluency (speech rate, mean length of pauses, and repair phenomena). Improvement was measured as the difference between the first and last practice sessions. The results showed that accuracy, speech rate, and pauses improved but with distinct patterns. Regression models showed that declarative knowledge did not predict accuracy or fluency; however, declarative knowledge assisted the learners to engage in targetlike behaviors at the initial stage of proceduralization. Furthermore, whereas production of accurate *wh*-questions predicted accuracy improvement, it had no impact on fluency.

Practice has been shown to positively impact second language (L2) speakers' accuracy during memory retrieval and their processing speed. Drawing on skill-learning theories

This work was partially supported by grants awarded to the first author by the Fondo Nacional de Desarrollo Científico y Tecnólogico from the Ministry of Education of Chile (FONDECYT: 1181533) and PIA (CIE160009) from the Chilean National Commission of Science and Technology (CONICYT) as well as funding awarded to the second author from the Canada Research Chairs program (950-231218).

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://www.iris-database.org/iris/app/home/detail?id=york%3a936167.

We would like to thank the teachers who generously supported our project and the research assistants who helped with data collection and coding: Mayuri Kewlani, Estefanía Valencia, Camila Valenzuela, Mélanie Vergara, and Paula Viveros.

*Correspondence concerning this article should be addressed to Masatoshi Sato, Department of English, Universidad Andres Bello, Fernández Concha 700, Las Condes, Santiago, 7550000, Chile. Email: masatoshi.sato@unab.cl

Copyright © Cambridge University Press 2019

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

(Anderson, 1985; Schneider & Fisk, 1983), research in cognitive psychology has typically examined error rate and reaction time during paired-associate tasks. Neuroimaging evidence has shown that during skill development, learners increasingly become reliant on procedural memory over declarative memory, resulting in more skilled performance, that is, faster and more accurate performance (Eichenbaum, 2012; Ullman, 2001). What drives the change in memory retrieval is practice; by being engaged in repeated practice, learners develop memory routines based on the procedural system more and more, that is, proceduralization. Ultimately, L2 learners may automatize L2 knowledge by restructuring the components of procedural memory. To that end, the vast majority of L2 practice studies have examined the spacing effect, that is, the length of intervals between practice sessions (e.g., Nakata, 2015; Rodgers, 2011; Schuetze, 2015; Serrano, Stengers, & Housen, 2015; Suzuki & DeKeyser, 2017a; Toppino & Gerbier, 2014), with the findings revealing positive effects of practice on L2 learning overall.

However, previous L2 research faces several methodological issues. First, the practice tasks in empirical studies have been largely mechanical and decontextualized. Given the evidence related to skill specificity (Healy & Bourne, 1995) and transfer-appropriate processing (Morris, Bransford, & Franks, 1977), practice conditions should approximate the context where L2 learners will use the learned skills in the real world. Specifically, if the instructional goal is to help learners develop L2 communicative skills, repeated practice should elicit memory retrievals of those skills. Therefore, the current study implemented contextualized practice activities as part of classroom instruction and examined their effect on English as a foreign language (EFL) learners' production of *wh*-questions. Second, previous research has examined learners' correct responses to experimental stimuli only. However, during the learning processes, L2 learners may draw upon their interlanguage systems and produce incorrect utterances. It is unclear whether practicing inaccurate forms has any longer-term effects. Third, while skill-learning theories claim that declarative knowledge supports the development of procedural knowledge (DeKeyser, 2017), evidence for their association has been scarce. Therefore, the current study considered the potential role of declarative knowledge, along with both correct and incorrect production, as potential predictors of increased accuracy and fluency.

BACKGROUND AND MOTIVATION

THE DECLARATIVE/PROCEDURAL MODEL AND SKILL LEARNING

Drawing on theories of skill learning (Anderson, 1985) and automaticity (Schneider & Fisk, 1983) from cognitive psychology, skill acquisition theory states that L2 skill development involves two types of memory systems. On the one hand, the declarative memory system (knowledge *that*) is characterized by controlled processing that is slow, serial, and effortful. Though often thought to hold rules of grammatical systems, declarative representations may contain a broader range of factual knowledge, such as the meaning of words, their phonological forms, and grammatical specifications (Morgan-Short & Ullman, 2012). On the other hand, the procedural memory system subsumes fast, parallel, and fairly effortless processes. The advantage of procedural processing is that it frees up space in short-term memory so that an individual can handle other tasks

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

concurrently. Hence, procedural knowledge (knowledge *how*) is thought to be responsible for actual actions, such as speaking a L2 fluently (DeKeyser, 2017; Gade, Druey, Souza, & Oberauer, 2014).

The declarative/procedural model explains certain L2 development phenomena (Ullman, 2001). For instance, child learners in an immersion school develop a procedural system that may contain fossilized L2 errors, in which case declarative knowledge may be necessary to restructure the corresponding procedural system. Another phenomenon can be found in foreign language learning situations in which learners store what they learn using grammar-translation methods in the declarative system; however, instructional activities do not help them develop the procedural memory that is required for spontaneous production (see DeKeyser, 2007a; Lyster & Sato, 2013). Meanwhile, the development of procedural memory takes longer than that of the declarative system. This is because procedural memory is less amenable to changes because its activation involves different areas of brain; in contrast, retrieval of declarative memory is associated with a narrower region in the brain (Chein & Schneider, 2005; Eichenbaum, 2012).

Important to note is that the declarative/procedural model or automaticity theory does not predict that declarative knowledge becomes or is converted to procedural knowledge as a result of skill development; rather, depending on the task to be performed, "the relative functionality of the two systems can affect which one is relied on more" (Ullman & Lovelett, 2018, p. 42). During the skill learning process, a person becomes more reliant on procedural knowledge while declarative knowledge supports the shift in dependence on memory retrieval routines. In other words, the controlled and automatic processes constantly interact and controlled processing "modifies memory and leads to the development of automatic processing" (Schneider, Dumais, & Shiffrin, 1984, p. 23).[1] In the context of L2 skill learning, therefore, it can be hypothesized that having grammatical knowledge of a specific structure would support the development of the procedural memory system, which controls faster processing. In addition, it may be the case that declarative grammatical knowledge enhances accuracy increases while developing procedural representations. The current study tested practice effects while considering the declarative knowledge that the learners had possessed prior to the practice sessions.

PROCEDURALIZATION AND THE DEVELOPMENT OF ACCURACY AND FLUENCY

The premise of skill-learning theories is that as a person develops procedural representations of a skill, the error rate of performing the relevant task decreases and the speed of memory access increases. Those patterns conform to the power law of practice—an initial rapid change followed by gradual changes (Langley, Laird, & Rogers, 2009; Newell & Rosenbloom, 1981). In DeKeyser's (1997) experiment, the participants were first taught morphosyntactic rules and vocabulary of an artificial language. The participants then engaged in both comprehension and production practice sessions. In production tasks, the participants completed gap-fill sentence tasks that described pictures. The changes of error rate followed the power function of practice whereby the participants became dramatically more accurate from the first to the second sessions while the following decline in error rates was more gradual (see also Ferman, Olshtain, Schechtman, & Karni, 2009).

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

Similar findings have been reported in real languages as well. For instance, Bird (2010) investigated the acquisition of an English morphosyntactic structure (i.e., tense/aspect of verb forms). The learners engaged in practice sessions involving error correction of ungrammatical forms on worksheets. In Suzuki and DeKeyser (2017a), grammatical accuracy of a Japanese morphology (*te*-form) was targeted. Regardless of the practice conditions (spaced or massed), learners showed improvement over four practice sessions, with changes conforming to the power law of practice. With a cross-sectional design involving three different proficiency groups, Rodgers (2011) compared oral production during picture description tasks. In these tasks, learners completed sentences by filling in a gap related to the target structure (the regular present tense of the indicative in Italian). The analysis of error rates showed that the three groups significantly differed and Rodgers concluded that "learners become more automatic … as they become more proficient in the language" (pp. 311–312).

With regards to the speed of cognitive processing, skill-learning research has primarily examined reaction time when responding to stimuli or performing a task, based on a premise that reaction time provides insight into "how the mind works, and infer about the cognitive processes or mechanisms involved in language processing" (Jiang, 2012, p. 2). Various studies reported practice effects as evidenced by a decline in reaction time (see DeKeyser, 2017). However, the speed of cognitive processing can be operationalized as oral fluency as well, as Anderson (2005) illustrated in relation to foreign language learning: "the procedural, not the declarative, knowledge governs the skilled [fluent] performance" (p. 282). Indeed, oral fluency is a more relevant construct when considering classroom L2 learning, which is the focus of the current study.

L2 fluency was conceptualized into three types by Segalowitz (2010): cognitive, utterance, and perceived. Utterance fluency involves temporal aspects of L2 production and cognitive fluency is arguably manifested in some of those indices. Perceived fluency is based on listeners' impressions of speeches. The development of utterance fluency can be characterized by temporal measures, such as speed (e.g., speech rate), pauses (filled or unfilled), and repair (e.g., repetitions) phenomena (see Derwing, 2017). Research suggests that certain aspects of utterance fluency represent cognitive fluency, although findings have been mixed. For instance, in Segalowitz and Freed's (2004) study, cognitive processing was investigated through lexical access (animacy judgment tests) and the results were compared with the learners' utterance fluency during oral proficiency interviews. Among different oral fluency indices, filled pauses were correlated with the reaction time scores, meaning that cognitive fluency was observed in pause phenomena. However, in an investigation of cognitive versus utterance fluency, de Jong, Steinel, Florijn, Schoonen, and Hulstijn (2013) found that speech rate (mean syllable duration) was the best predictor of cognitive fluency ($\beta = .50$), while pause phenomena (silent pause duration) explained only 5% of the total variance of cognitive fluency.

Several studies examined the effect of practice on utterance fluency. De Jong and Perfetti (2011), for example, examined whether a fluency-focused oral activity (the 4/3/2 task) impacted the development of temporal aspects of spontaneous production. The learners who repeated the same speech multiple times, as opposed to giving different speeches, improved their fluency by reducing the length of pauses, increasing syllables per minute, and lengthening stretches of speech. In Suzuki and DeKeyser (2017a),

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

reaction time included "from the onset of the prompt word to the end of the utterance" (p. 174), suggesting that the scores indicated how fast the learners spoke—a similar index to speech rate. Having found improvement in the fluency measure, Suzuki and DeKeyser claimed that "[a] speed measure is a more sensitive index of procedural knowledge because it not only tests the accuracy of conjugation, but also how fast the morphological knowledge can be deployed" (p. 182). In sum, research has shown that repeated practice supports proceduralization (or automatization) both in accuracy and fluency (see also Lim & Godfroid, 2015). However, the practice and testing materials in the studies reviewed have been rather decontextualized (e.g., fill in the gap), leaving a question as to whether practice effects extend to a more communicative skill.

QUALITY OF PRACTICE

Proceduralization can be facilitated by repeated practice. As Anderson and Schunn (2000) put it: skill acquisition theory places "a premium on the practice which is required to learn permanently the components of the desired competence" (p. 2). However, two issues related to the quality of practice require further investigation. First, previous research considered correct responses only and ignored incorrect responses, possibly due to the nature of practice—an incorrect response in a paired-associate task may not contribute to retention rates of correct choices. However, in instructed L2 learning settings, learners who are developing procedural L2 knowledge produce incorrect utterances. During this process, learners may produce partially inaccurate utterances based on their interlanguage systems.

The second issue is related to the "meaningfulness" of practice. A concept of practice includes a wide array of "activities in the second language engaged systematically, deliberately, with the goal of developing knowledge of and skills in the second language" (DeKeyser, 2007b, p. 8). To that end, it is claimed that meaningful practice "is better at driving the brain processes that result in successful recall" (Anderson, 2005, p. 198). Nonetheless, much of cognitive psychology research has operationalized practice as memorization during paired-associate tasks. For instance, participants in Pavlik and Anderson (2005) memorized English translations of Japanese words. The task that the participants in Morgan-Short, Steinhauer, Sanz, and Ullman (2012) performed involved an artificial language, where participants followed instructions in the language during a computer board game. In L2 research as well, practice has been rather decontextualized. Although the task in de Jong and Perfetti (2011) was delivered as part of classroom activities, it was a monologue task in which students individually recorded their speeches in a computer lab, responding to general questions, such as "What do you think about pets?" The production tasks in Suzuki and DeKeyser (2017a) involved depicting pictures to the experimenter and describing short video clips to the computer.

A tenet of skill-learning theories is transfer-appropriate processing; a skill learned in a certain context is best transferred to performance in a similar context (Morris et al., 1977). DeKeyser (1997) showed that practice of comprehension led to proceduralization of comprehension skills, while production practice led to production skills. In a reading fluency study, Martin-Chang and Levy (2005, 2006) compared isolated-training involving computerized word-naming games and context-training providing

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

learners with the target words in a story. The results showed that learners who were given isolated training scored higher on the test in which individual words were presented on a computer screen, while learners in the context-training group, who were taught words in passages, outperformed the isolated-training group in the reading passage test. Hence, evidence suggests that practice effects are specific in regard to skills, tasks, and context, and if the goal of L2 instruction is to help learners develop communicative skills, practice should be contextualized. (We will detail our operationalization of contextualization in the "Methods" section.)

THE CURRENT STUDY

DeKeyser (2010) called for an investigation in which practice is operationalized in a "broader definition of systematic practice and fulfill[s] specific functions in the learning process that mere communicative input and interaction cannot" (p. 159). The current study answers this call by examining practice effects in an EFL classroom setting. While SLA researchers have long argued for the importance of meaningful, communicative practice to develop communicative skills (Ellis & Shintani, 2014; Lightbown, 2000; Loewen & Sato, 2017; Ortega, 2014), empirical research has examined practice that is not meaningful. Therefore, the current study examined the impact of contextualized practice. We operationally define *contextualized practice* as a type of practice that occurs in an intact classroom where learners engage in production that is transfer appropriate to authentic, communicative language use.

Additionally, limited research has examined practice effects on grammatical structures (e.g., Bird, 2010; Suzuki & DeKeyser, 2017a), or focused on *wh*-questions—the target structure in the current study. In terms of outcome measures, previous research generally focused on processing speed during vocabulary memorization tasks (e.g., Nakata, 2015; Schuetze, 2015; Sobel, Cepeda, & Kapler, 2011). Some studies examined utterance fluency (e.g., de Jong & Perfetti, 2011; Suzuki & DeKeyser, 2017a), but with decontextualized monologue tasks. The current study examined L2 production during interactive tasks with the classroom teacher and analyzed the learners' accuracy and fluency.

Furthermore, the contribution of both types of memory systems to improved L2 performance has not been identified. In other words, we do not know much yet about how declarative knowledge contributes to the development of procedural knowledge. We explored this issue by considering the scores of a declarative knowledge test as a predictor variable when gauging the impact of practice on procedural knowledge. Lastly, given that L2 learners unavoidably produce inaccurate structures, we explored the differential impact of correct and incorrect utterances in regression models. The research questions were as follows:

RQ1: Does contextualized practice facilitate L2 speakers' production of accurate and fluent *wh*-questions?

RQ2: What learner and practice factors predict changes in the accuracy or fluency of the L2 speakers' *wh*-questions?

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

METHODS

PARTICIPANTS

The initial participant pool consisted of 55 students enrolled in an undergraduate degree program in English education at a large private university in Santiago, Chile. However, students who missed more than two practice sessions ($n = 21$) were excluded from the study. Therefore, the final participant pool consisted of 34 students (16 males and 18 females) who ranged in age from 18 to 24, with a mean age of 21.3 ($SD = 2.0$). In terms of their prior English instruction, they had studied English in primary and secondary schools for a mean of 11.5 years ($SD = 3.0$), but few students reported using English for communication outside the classroom. Only five students reported 1 to 2 hours per week of conversation with their friends in English and three students reported that they used English at work. The students had never taken any standardized proficiency tests, but based on the university curriculum, they had recently completed classes at the B1 level on the Common European Framework of Reference (Council of Europe, 2001). The students were taking a required English class that followed general principles of communicative language teaching, and emphasized the development of students' communicative competence. Instructors regularly engaged in whole-class discussions. The course was theme based and targeted specific lexical and grammatical forms in each unit.

TARGET STRUCTURE

The target structure of the intervention was *wh*-questions, which have been widely studied in classroom-based SLA research to explore the developmental outcomes associated with corrective feedback (Adams, 2007; Lightbown & Spada, 1990), task complexity (Kim, 2012), and structural priming (McDonough & Chaikitmongkol, 2010), to test the effectiveness of form-focused instruction (Spada & Lightbown, 1999; White, Spada, Lightbown, & Ranta, 1991) and to explore learners' awareness of L1-L2 differences (Ammar, Lightbown, & Spada, 2010; Lightbown & Spada, 2000). In addition to allowing for comparisons with previous research, the focus on *wh*-questions was also motivated by pedagogical reasons, as questions complemented the objectives and materials of the participants' EFL integrated skills course.

The *wh*-question form under investigation consisted of a question word followed by an obligatory auxiliary verb, subject, and lexical verb, which has been referred to as a *wh*-aux 2nd question (Pienemann & Johnston, 1987; Pienemann, Johnston, & Brindley, 1988). This type of *wh*-question was selected because L2-English learners often alternate between several variants: the targetlike form in which the obligatory auxiliary verb is supplied (*when was the baby born?*), an interlanguage form without an auxiliary verb (**when the baby born?*), or an interlanguage form with a misplaced auxiliary verb (**when the baby was born?*). Even when learners produce an auxiliary verb in the correct location, they may still make errors involving subject/auxiliary verb agreement (e.g., **where are the family going?*) or auxiliary/lexical verb agreement (e.g., **what goal does he has?*). In the current study, only *wh*-questions that require auxiliary verbs were considered because a missing auxiliary verb is not ungrammatical for some question types, such as *wh*-questions with the copula (*what color is the dog?*) or questions in which the *wh*-word functions as a subject (*which boy likes chocolate?*).

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

MATERIALS

Declarative knowledge test

The declarative knowledge test was created by the researchers to assess the participants' preexisting declarative knowledge of the target structure. The test included three item types that the participants were familiar with from their previous English classes: fill in the blank, correct the error, and scrambled sentences. The 16 fill-in-the-blank items provided a short dialogue consisting of a question-and-answer sequence in which one word in the question had been deleted, specifically the question word (four items), the auxiliary verb (eight items), or the lexical verb (four items). The error correction section also provided 16 short dialogues in the form of question and answer sequences. However, each question contained an error that involved the question word (four items), the auxiliary verb (six items), or the lexical verb (six items). Finally, the scrambled sentence section provided eight sentences in jumbled order, which included five *wh*-aux 2nd questions, two *wh*-subject questions, and one negative question. The 40-item test was pilot tested with students at the same university ($N = 21$) and instrument reliability was assessed using Cronbach's alpha, which was .75. Based on the item statistics, seven items that showed no variation were removed from the test. Therefore, the final test included 13 fill-in-the-blank items, 15 error correction items, and 5 scrambled sentences, for a total of 33 items. Reflecting our main focus, the test contained 24 *wh*-aux 2nd questions along with nine items that targeted other types of questions, including *wh*-subject and *wh*-copula questions. Although the test included these *wh*-questions that do not require an auxiliary verb as distracters, the students' performance on these items was not included in their declarative knowledge scores. Based on the participants' performance, the instrument reliability (Cronbach's alpha) of the declarative knowledge test was .87 for the 24 *wh*-aux 2nd items and .89 for all 33 items.

Contextualized practice activities

The current study operationalized contextualized practice by drawing on general educational as well as L2 psycholinguistic literature. Drawing on activity theory (Leont'ev, 1981), Van Oers (1998) argued that contextualization involves connecting a specific learning context and meaning making in that context (see also Lyster & Sato, 2013). In the current investigation, the learners engaged in practice that triggered their goal-oriented behavior. Specifically, their goal was to develop communicative skills. The practice activities were designed to help them achieve the meaning making (the goal) in the classroom context. The topics of practice were chosen based on the particular context, including the existing curriculum, so as to further facilitate learners' "continuity of thought" (Walz, 1989, p. 162) between their regular activities and those for the current experiment. For the same purpose, the activities were collaboratively developed with the teachers. Finally, the practice was done between students and their teachers, as opposed to the researchers, to increase the purposeful use of language in the classroom context.

The activities consisted of teacher-fronted, information-gap tasks that complemented three themes targeted in the students' classes during the intervention period: work and education, environmental issues, and physical and mental health. Six information-gap task types were developed, with multiple versions created to yield 15 tasks distributed

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

TABLE 1. Task versions by practice activity

Task type	Communicative goal	Versions	Practice activity
Biography	Determine the identity of a famous Chilean athlete, scientist or architect	Tomás González	1
		Mar cela Contreras	3
		Alejandro Aravena	5
Interview	Obtain information about the instructor's life	Travel, social media, love, goals	2
		Restaurants, inspirational people, moving abroad, relaxation	4
Picture difference	Discover differences between two pictures depicting similar scenes	Messy living room	1
		Space travel	3
		At the beach	5
Story completion	Reveal the narrative that links a series of pictures	Missing cat	2
		Lost and found	4
Truth or lie?	Decide if the instructors are telling the truth about their lives or lying	Music, holidays, housing bad habits	1
		Food, exercise, friendship, television	3
		Celebrations, sports,fashion, movies	5
W'hat happened?	Figure out what problems were encounteredin a neighborhood	Shopping & errands	2
		Work & travel	4

across five practice sessions (three tasks per session). A complete list of the task types, versions, and the distribution across the practice activities is provided in Table 1. Although there was variation in the communicative goal of the tasks (e.g., deciding if the instructor was telling the truth or lying vs. guessing the identity of a Chilean celebrity), they were all one-way tasks in which the instructor held the information and the students' role was to ask questions to obtain that information. All activities were led by the teacher in the classroom where the whole class interacted with the teacher concurrently, so as to approximate the experimental setting to real classroom interaction.

All tasks were presented through PowerPoint slides that showed images along with prompts in the form of *wh*-words (*what*, *when*, *where*, *how*, *why*, *how much/many*, *what* + *noun*, and *which* + *noun*). The number of *wh*-prompts varied from four to eight across the six task types, but each practice session presented a total of 20 *wh*-prompts with the same distribution of specific *wh*-words. The *wh*-prompts were timed to appear on the images one at a time, giving students 10 seconds to produce a question using the *wh*-prompt. An audio signal (i.e., a beep) was used to indicate that time had expired before the next *wh*-prompt appeared (see Supplementary Materials for sample tasks). The instructor provided answers after each *wh*-prompt or at the end of each task. Depending on the task, the teachers were provided with the answers. For instance, in the Biography task, where the learners' goal was to guess a Chilean celebrity, the teacher was given relevant information about the celebrity. However, in the Interview task, where the learners asked questions to know more about the teacher's personal life, the teacher answered the learners' questions spontaneously.

DESIGN

A within-group, repeated-measures design was used to determine whether the participants' accuracy and fluency of *wh*-questions changed across the five practice sessions.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

Accuracy was operationalized as grammatically correct production of the target structure, while fluency was operationalized as utterance fluency. We analyzed three indices of utterance fluency—speed, pause, and repair phenomena. For the second research question, an associational design was used to determine what factors were associated with changes in the participants' production of *wh*-questions. The outcome variable was the gain scores for accuracy and fluency from the first and final sessions. The predictor variables were (a) the scores of the declarative knowledge test; (b) the total number of correct *wh*-questions across the second, third, and fourth sessions; and (c) the total number of incorrect *wh*-questions across the second, third, and fourth sessions.

PROCEDURE

The activities were carried out in the participants' regularly scheduled classes over a five-week period (see Table 2). In the first week, the participants completed the consent and background information forms (15 min), along with the declarative knowledge test (20 min). To familiarize the students with the activity format and audio-recording equipment, a task with communicative activities similar to the intervention activities was implemented in week two (20 min). However, the familiarization task targeted *yes/no* questions instead of *wh*-questions. In weeks three to seven (i.e., Time 1 to Time 5), learners engaged in the contextualized practice activities. Each session took approximately 15 minutes to complete, and was administered at the beginning or end of a class period based on its fit with the lesson's other instructional activities. All students held a digital audio-recorder to individually record their language production during the practice tasks. It is important to stress that the teachers did not provide any feedback on the learners' erroneous utterances. While skill acquisition theory underscores the importance of feedback to avoid entrenching inaccurate knowledge structures, feedback was controlled in the current experiment so as to tease apart the effect of contextualized practice.

DATA CODING AND ANALYSIS

Declarative knowledge

The tests were scored by the researchers by giving one point to each correct item, with no half points for partially correct answers. For the fill-in-the-blank items, correct items

TABLE 2. Research activities by week

Week	Research activity
1	Consent & background information forms, declarative knowledge test
2	Familiarization task: *yes/no* questions
3	Contextualized practice 1: Biography, picture difference, truth or lie
4	Contextualized practice 2: Interview, story completion, what happened
5	Contextualized practice 3: Biography, picture difference, truth or lie
6	Contextualized practice 4: Interview, story completion, what happened
7	Contextualized practice 5: Biography, picture difference, truth or lie

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

were those for which a participant had supplied a question word, auxiliary verb, or lexical verb that resulted in a grammatically accurate question. For the error correction items, correct items had a change (insertion or deletion) that repaired the error and resulted in a grammatically accurate question. Finally, the scrambled question items were considered correct if the question contained all the words provided in a grammatically accurate word order without any additions or deletions. The participants' scores for the 24 *wh*-aux 2nd question items were summed and used in the subsequent analysis, but their scores for the nine distracter questions (*wh*-subject and *wh*-copula) were not analyzed further.

Accuracy

The audio-recordings of the practice sessions were transcribed and verified by research assistants. The transcripts were analyzed by research assistants for the occurrence of *wh*-questions, which were classified as fragments, *wh*-subject, *wh*-copula, and *wh*-aux 2nd. Definitions and examples of each type are provided in Table 3.

After classifying all the participants' questions, their *wh*-aux 2nd questions were further analyzed for accuracy. To be classified as accurate, a *wh*-aux 2nd question required the correct (a) question word, (b) word order (i.e., position of auxiliary verb), (c) subject/auxiliary verb number agreement, and (d) auxiliary verb/main verb agreement in tense and aspect. Errors involving forms unrelated to question formation, such as article use, plurals, or lexical choice (e.g., *blood banker* as opposed to *blood bank*), were not considered in the accuracy coding. Proportional scores were obtained by calculating the percentages of incorrect utterances out of total utterances in each session. A subset of the data (20%) was coded by a research assistant for interrater reliability, which was

TABLE 3. *Wh*-question types with examples

Question type	Definition	Examples
Fragments	*Wh*-question word with one or two lexical items	What subject? What kind?
Wh-subject	*Wh*-questions in which the *wh*-word acts as the subject, and auxiliary verbs are not required	How many people went with pets? What happened on your last birthday? *What make Alex feel hungry? *What inspire him to study English?
Wh-copula	*Wh*-questions with copular *be*	What is your goal in life? What is his motivation? *Why does she your favorite actress? *What is inspirational people for you?
Wh-aux 2nd	*Wh*-questions that require an auxiliary verb located between the *wh*-word(s) and the subject	Which sport does he like? How often does she practice this sport? Where did she donate blood? *Where did you took the cane? *When she go to the university? *Where the plane is going?

* INDICATES ungrammatical questions

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

calculated as Cohen's kappa for the classification of *wh*-questions (.85) and intraclass correlation coefficients for the accurate *wh*-aux 2nd questions (.94).

Fluency

Following previous practice research, we focused on fluency of correct productions of *wh*-aux 2nd questions. The resulted number of questions analyzed for fluency was 1,133 out of 1,973 (57.43%) (details are presented in the results section). First, fluency of individual questions was analyzed using Audacity. Then, the average scores of the fluency indices, based on the number of questions that each learner produced, were calculated for each learner at each practice session. Hence, the scores presented in the "Results" section indicate fluency per question.

Three fluency indices were used: speech rate (syllables per second), mean length of pauses (average length of pauses), and repair phenomena (summed frequency of false starts, repetitions, and self-corrections). In terms of the cutoff length of pauses, de Jong and Bosker (2013) compared different cutoff lengths (between 100 ms to 1,000 ms) in relation to vocabulary knowledge and perceived fluency. The results showed that a 250–300 ms cutoff was the "optimal threshold for measuring the number of pauses (per total or per speaking time) with respect to studies that aim to investigate L2 proficiency" (p. 20). Following this result, the silences shorter than 250 ms, so-called micropauses (Riggenbach, 1991), were not considered as pauses in the current study. Interrater reliability was calculated on a subset of the data (20%) which yielded acceptable agreement rates for all indices (speech rate: $r = .78$; mean length of pauses: $r = .75$; repairs: $r = .91$).

Statistical analyses

In the dataset, only 19 learners participated in all five sessions. As a result, 25 cases were missing, among a potential of 170 cases (34 participants tested five times). Accordingly, we conducted missing values analyses on the dataset, using the expectation-maximization algorithm (see Mbogning, Bleakley, & Lavielle, 2015).[2]

To understand the overall patterns of accuracy and fluency changes over the five sessions, descriptive statistics were first analyzed and visually inspected. Then, the data were submitted to two types of inferential statistics. First, repeated measures ANOVAs (RM ANOVAs) were chosen for examining the impact of contextualized practice on the participants' accuracy and fluency. The assumption of normality was met for error rate at five sessions (all above $p = .20$ according to Kolmogorov-Smirnov) as well as all fluency indices at five sessions (all above $p = .06$ according to Kolmogorov-Smirnov). However, Mauchly's tests showed that the dataset violated the assumption of sphericity (error rate: $p = .005$; speech rate: $p < .000$; mean length of pauses: $p = .004$; repair phenomena: $p < .000$). In the "Results" section, therefore, we report the results with Huynh-Feldt correction (sphericity not assumed). For the post-hoc pairwise comparisons, LSD was chosen given that some of the comparisons were irrelevant to the objective of the study (e.g., Time 2 vs. Time 5). The alpha level for all tests of significance was set at .05.

Second, to answer the second research question pertaining to the predictors of improved accuracy or fluency, hierarchical multiple regressions were chosen. In total, four

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

models were tested for the four indices of accuracy and fluency. In all models, the two predictor variables were the frequencies of correct and incorrect productions. The expectation-maximization algorithm was not applied to the raw frequencies data, meaning that the scores reflected actual production frequencies. The frequencies from Times 2, 3, and 4 were summed; Times 1 and 5 were not included so as to avoid violating the independence of observation for the inferential statistics. To explore the contribution of declarative knowledge to proceduralization, the scores of the declarative knowledge test were entered in the first block (Step 1: see Petrocelli, 2003). In the second block (Step 2), the frequencies of correct and incorrect productions were entered as the predictors. In so doing, we tested (a) the predictive value of declarative knowledge for proceduralization in the first models, and (b) we controlled the learners' preexisting declarative knowledge and teased apart the effect of the quality of practice on the changes in accuracy and fluency over time in the second models.

The assumptions of hierarchical multiple regressions were met (see Osborne & Waters, 2002). First, the sample size requirement based on two predictors was met according to Pedhazur (1982). Second, all the outcome variables were normally distributed: error rate (Kolmogorov-Smirnov, $p = .08$), speech rate (Kolmogorov-Smirnov, $p = .20$), mean length of pauses (Kolmogorov-Smirnov, $p = .09$), and repair phenomena (Kolmogorov-Smirnov, $p = .14$). Third, multicollinearity was not observed in the dataset: the correlation coefficient between the two predictors was .36 (see Supplementary Materials for the correlation matrix). Also, the variance inflation factors in all models ($VIF_{range} = .67–1.48$) supported the use of regression analyses (Heiberger & Holland, 2004). Finally, no homoscedasticity was observed in the scatterplots.

RESULTS

DATA OVERVIEW

Over the five practice sessions, the participants produced 2,851 questions in total. As can be seen in Table 4, learners produced comparable numbers of questions in each session, as shown by the individual averages ranging from 18.90 to 23.21. Also, judging from the standard deviations, ranging from 6.24 to 11.28, it can be said that all learners produced questions in every session to a certain degree. The minimal frequencies in each session were: Time 1 = 18; Time 2 = 17; Time 3 = 15; Time 4 = 16; and Time 5 = 15.

Across the data collection points, the learners produced more *wh*-aux-2nd questions—the target structure of the current study—than *wh*-subject and *wh*-copula questions. As the intervention activities aimed for, the production of *wh*-subject questions was the lowest, amounting to 3% (84/2,851) of the total questions (details can be found in Supplementary Materials). Notably, the learners produced *wh*-aux-2nd questions consistently over the five sessions (range = 12.73–16.26), with much less variation than *wh*-subject and *wh*-copula questions.

In the regression analyses, the cumulative frequencies of *wh*-aux-2nd questions from Times 2, 3, and 4 were entered as the predictor variables. In total, the learners produced 1,216 *wh*-aux-2nd questions. On average, they produced 21.62 ($SD = 14.58$) correct questions and 23.82 ($SD = 16.77$) incorrect questions. Both scores were normally distributed according to Kolmogorov-Smirnov (correct: $p = .20$; incorrect: $p = .19$).

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

TABLE 4. Descriptive statistics of question types

	Time 1		Time 2		Time 3		Time 4		Time 5		Total	
	f	*M(SD)*	*f*	*M(SD)*	*f*	*M(SD)*	*f*	*M(SD)*	*f*	*M(SD)*	*f*	*M(SD)*
wh-S	3	.09(.29)	32	1.03(.55)	7	.24(.51)	33	1.57(1.36)	9	.42(81)	84	.62(.89)
wh-C	343	10.39(6 .08)	107	3.46(1.88)	188	6.48(3.95)	50	2.38(1.66)	133	6.33(3.04)	821	6.08(4.79)
wh-A	420	12.73(4.66)	504	16.26(6.34)	398	13.72(7.86)	314	14.95(5.23)	310	14.76(5.28)	1946	42.24(25.56)
Total	766	23.21(8.83)	643	20.74(7.59)	593	20.45(11.28)	397	18.90(6.24)	452	21.52(7.65)	2851	

Note. f = the frequency of produced questions; *wh*-S *wh*-subject questions; *wh*-C = *wh*-copula questions; *wh*-A = *wh*-auxiliary 2nd questions. The means represent the number of questions produced by each learner.

ACCURACY AND FLUENCY CHANGES OVER TIME

Accuracy

The RM ANOVA of error rate showed a significant main effect: $F(3.59, 118.40) = 4.863$, $p = .002$, $\eta_p^2 = .128$. Table 5 reports the descriptive statistics of each session.[3] The pairwise comparisons for each interval revealed that the error rates were significantly reduced from Time 1 to Time 2 ($p = .002$, $d = .32$, 95% CI [.54, 15.26]) and from Time 4 to Time 5 ($p = .025$, $d = .32$, 95% CI [.86, 12.07]). In addition, the comparison between Time 1 and Time 5 reached significance ($p < .000$, $d = .72$, 95% CI [7.33, 22.68]). The changes in the other intervals were not significant: Time 2–Time 3 ($p = .994$, $d < .00$, 95% CI [–7.16, 7.22]); Time 3–Time 4 ($p = .832$, $d = .02$, 95% CI [–5.25, 6.49]). Figure 1 depicts the changes of error rate over time.

Fluency

First, the RM ANOVA of speech rate detected a significant main effect: $F(1.93, 63.71) = 18.108$, $p < .000$, $\eta_p^2 = .354$. The pairwise comparisons between each interval showed that the speech rates significantly increased from Time 1 to Time 4 at each interval: Time 1–Time 2 ($p = .048$, $d = .26$, 95% CI [–.316, –.001]); Time 2–Time 3 ($p = .025$, $d = .30$, 95% CI [–.394, –.028]); and Time 3–Time 4 ($p < .000$, $d = .33$, 95% CI [–.311, –.132]). The difference between Time 4 and Time 5 did not reach significance ($p = .067$, $d = .40$, 95% CI [–.019, .523]). Furthermore, the comparison between Times 1 and 5 showed a significant increase ($p < .000$, $d = 1.53$, 95% CI [–1.117, –.570]). Figure 2 depicts the changes in speech rate over time.

The RM ANOVA of mean length of pauses also yielded a significant main effect: $F(3.34, 110.27) = 18.108$, $p = .005$, $\eta_p^2 = .114$. The analyses of the intervals showed that the mean length of pauses significantly decreased from Time 4 to Time 5 ($p = .001$, $d = .59$, 95% CI [.047, .160]). The differences in the other intervals were not significant: Time 1–Time 2 ($p = .932$, $d < .00$, 95% CI [–.098, .106]); Time 2–Time 3 ($p = .204$, $d = .16$, 95% CI [–.023, .103]); or Time 3–Time 4 ($p = .750$, $d < .00$, 95% CI [–.064, .046]).

TABLE 5. Changes in error rates and fluency indices over five sessions

	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 3 *M* (*SD*)	Time 4 *M* (*SD*)	Time 5 *M* (*SD*)
Error rate	49.34 (23.51)	41.44 (26.90)	41.41 (25.11)	40.79 (22.21)	34.33 (17.64)
Speech rate	2.45 (.49)	2.61 (.71)	2.82 (.70)	3.04 (.65)	3.30 (.60)
MLP	.46 (.23)	.46 (.26)	.42 (21)	.42 (.18)	.32 (.19)
Repairs	.28 (.20)	.25 (.16)	.27 (.14)	.25 (.13)	.26 (.14)

Note. MLP = mean length of pauses. The fluency indices—speech rate, MLP, and repairs—show scores per question.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

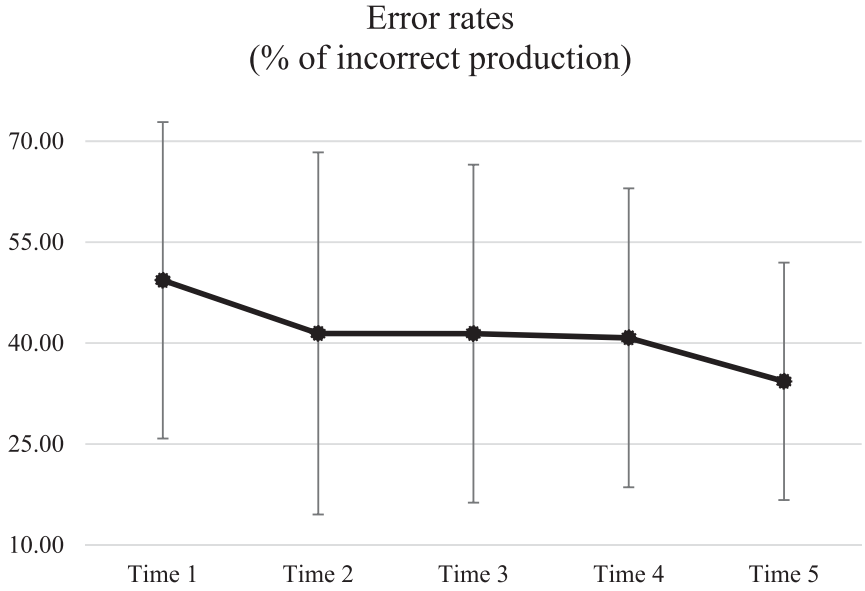


FIGURE 1. Error rates and their standard deviations at five sessions.

However, the comparison between Time 1 and Time 5 reached significance: $p = .005$, $d = .66$, 95% CI [.045, .233].[4] Figure 3 visually describes the changes in mean length of pauses over time.

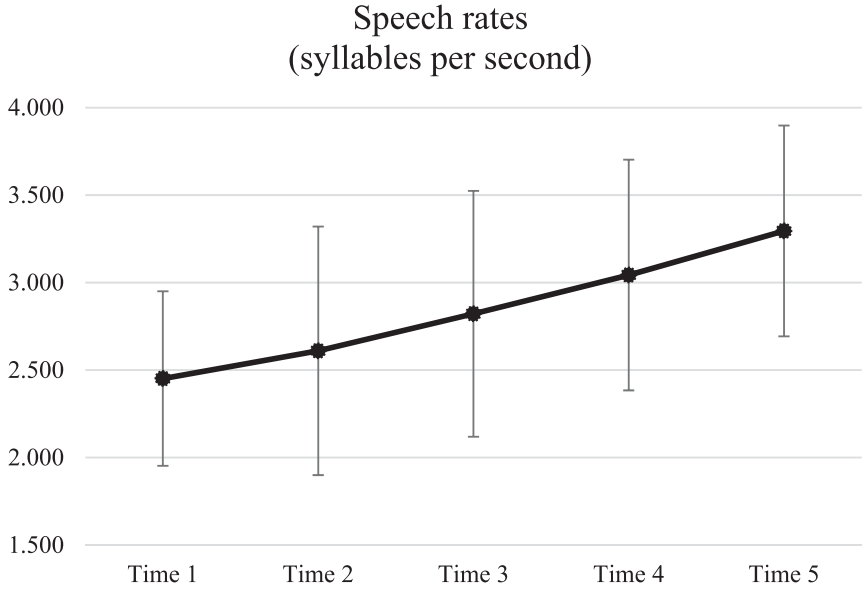


FIGURE 2. Speech rates and their standard deviations at five sessions.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

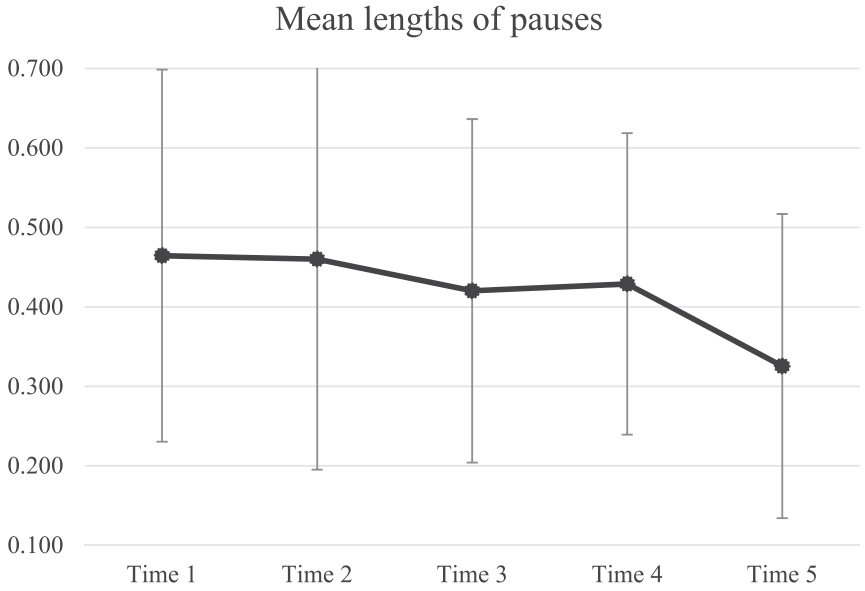


FIGURE 3. Mean lengths of pauses and their standard deviations at five sessions.

Finally, the RM ANOVA of repair phenomena did not show a significant effect: $F(3.20, 105.72) = .351$, $p = .801$, $\eta_p^2 = .011$. Figure 4 depicts the changes in the number of repair phenomena over time.

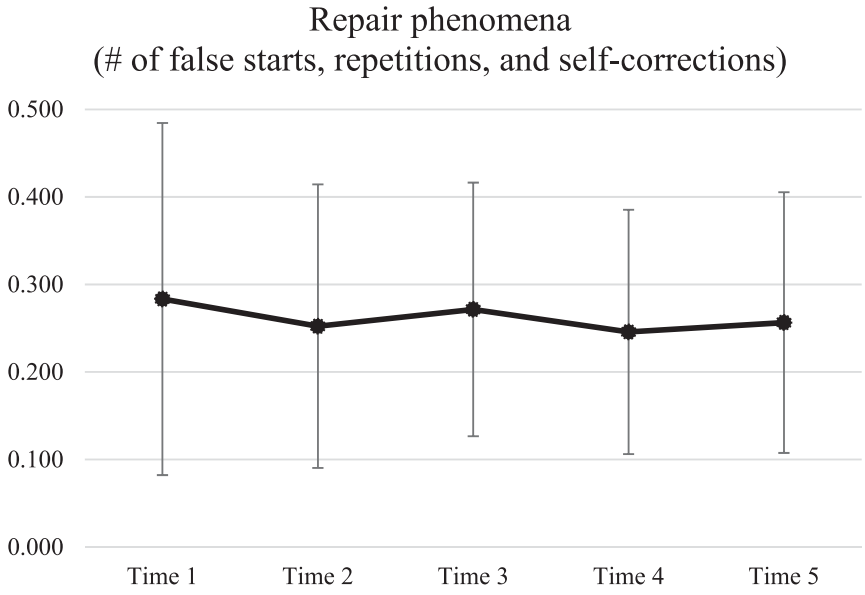


FIGURE 4. Repair phenomena and their standard deviations at five sessions.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

PREDICTORS OF PRACTICE EFFECT

Preexisting declarative knowledge

The grammar test administered prior to the intervention included 24 items related to *wh*-aux-2nd questions. The learners scored the 24 questions correctly with a mean of 59% (*SD* = 25.12). The data was normally distributed as evidenced by the result of Kolmogorov-Smirnov (p = .08).

Accuracy

The first regression model for accuracy, where the preexisting declarative knowledge was the predictor, yielded a nonsignificant result: $F(1, 32) = .232$, $p = .634$, $R^2 = .007$. However, the second model, which tested the effect of contextualized practice over and above learners' preexisting declarative knowledge, was significant: $F(2, 30) = 4.558$, $p = .019$, accounting for 24% (R^2) of the total variance in the change over time ($\Delta R^2 = .231$). As shown in Table 6, the grammar test scores did not predict the practice effect in the second model either ($p = .299$; $\beta = –.195$). However, both correct ($p = .007$) and incorrect ($p = .045$) productions were significant predictors. The negative correlation coefficients indicate that the more correct questions a learner produced, the more accurate his or her production became over time. More precisely, the standardized beta coefficient (β) indicates that for every unit increase in correct production, a decrease of .56 *SD* units can be expected in the change of error rate. On the contrary, production of incorrect questions showed positive predictive power ($\beta = .39$), meaning that the more incorrect questions a learner produced, the less accurate she or he became at the end of intervention.

Fluency

The difference scores (i.e., Time 5 minus Time 1) of the following fluency indices were entered as the outcome variables in each model: (a) speech rate, (b) mean length of pauses, and (c) repair phenomena (see Supplementary Materials for the correlation matrix).

The first model for speech rate, where the preexisting declarative knowledge was the predictor, yielded a nonsignificant result: $F(1, 32) = .472$, $p = .497$, $R^2 = .015$. The second model, after entering correct and incorrect productions, was not significant either: $F(2, 30) = 1.264$, $p = .304$, accounting for 11% (R^2) of the total variance in the change over time ($\Delta R^2 = .098$). Table 7 shows the details of the second model.

TABLE 6. Regression results for predictors of change of error rates

Outcome	Predictors	β	sr^2	t	p
Error rate	Grammar test	–.195	–.190	–1.057	.290
	Correct production	–.563	–.469	–2.909	.007
	Incorrect production	.388	.357	2.094	.045

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

TABLE 7. Regression results for predictors of change of fluency indices

Outcomes	Predictors	β	sr^2	t	p
Speech rate	Grammar test	–.044	–.041	–.224	.824
	Correct production	.295	.251	1.420	.166
	Incorrect production	–.329	–.291	–1.665	.106
Mean length of pauses	Grammar test	.190	.166	–.922	.364
	Correct production	–.071	–.060	–.328	.745
	Incorrect production	–.084	.074	.408	.686
Repair phenomena	Grammar test	.134	.117	.648	.522
	Correct production	–.025	–.021	–.113	.911
	Incorrect production	–.142	.124	.685	.498

Neither of the two models for mean length of pauses was significant; The first model: $F(1, 32) = .730$, $p = .399$, $R^2 = .022$; and the second model: $F(2, 30) = 0.293$, $p = .830$, with a minimal change from the first model ($\Delta R^2 = .006$).

Similar to the results of the other fluency indices, neither of the two models for repair phenomena was significant; the first model: $F(1, 32) = 0.317$, $p = .577$, $R^2 = .010$; and the second model: $F(2, 30) = 0.276$, $p = .842$, accounting for 3% (R^2) of the total variance in the change over time ($\Delta R^2 = .017$).

SUMMARY OF THE RESULTS

The results of the grammar test showed that the learners had not fully developed the declarative knowledge of the target structure ($M_{\%} = 59$). With this amount of declarative knowledge, the learners engaged in five sessions of contextualized practice. The RM ANOVAs showed that error rates of oral production incrementally decreased over time, especially between Time 1 and Time 2 and between Time 4 and Time 5. The difference between Time 1 and Time 5 was also significant. Similar patterns were observed in the changes of speech rate: incremental increases over time and a significant difference between Time 1 and Time 5. Although the decrease of mean length of pauses was gradual, the comparison between Time 1 and Time 5 reached significance. However, repair phenomena were not affected by the contextualized practice. In terms of predictors of the practice effect, preexisting grammar knowledge did not predict changes in either accuracy or fluency. The second model for accuracy showed that correct productions predicted the decrease of error rate. In contrast, the production of incorrect questions negatively affected the change in accuracy. Neither correct nor incorrect productions predicted changes in any indices of fluency.

DISCUSSION

CONTEXTUALIZED PRACTICE AND PROCEDURALIZATION

The first research question asked about the impact of contextualized practice operationalized as L2 production during communicative activities in the classroom. The

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

results of error rate showed an initial significant decline (Time 1–Time 2) followed by gradual changes (Time 2–Time 4). This pattern of changes conforms to the expected learning curves resembling the power function of practice. By engaging in contextualized practice, the learners were able to become more accurate in their spontaneous production over time; in other words, the practice may have helped develop the procedural memory of English *wh*-questions. In the current study, feedback was not provided during the practice sessions (cf. Suzuki & DeKeyser, 2017a). It may be the case that during the proceduralization, learners became more reliant on procedural memory that, in turn, freed up their short-term memory for accessing their declarative knowledge. Differently put, the practice did not aid the learners to learn a novel structure; rather, it helped them develop procedural representations of a target structure for which the learners already possessed declarative knowledge. In this situation, practice alone was sufficiently powerful to facilitate proceduralization.

One may argue that the initial decline was due to the testing effect. However, we note that not only did the learners participate in a training session with the identical activity but also the testing effect does not account for the significant difference between Time 4 and Time 5. We speculate that the significant change between Time 4 and Time 5 is an indication of partial proceduralization (DeKeyser, 2010). As can be seen in the error rate in Time 5 (34.33%), there was much more developmental space left at the end of the intervention. Accordingly, we believe that the learners were not yet at the automatization stage at Time 5 (see Hulstijn, Van Gelderen, & Schoonen, 2009; Segalowitz & Segalowitz, 1993). This interpretation suggests a possibility that the changes observed in the current study may represent faster access to the declarative system rather than deployment of the relevant knowledge in the procedural system.

The current study focused on three fluency indices: speech rate, mean length of pauses, and repair phenomena. The result of speech rate differed from that of accuracy in that, first, there was no significant change from Time 1 to Time 2, and second, the learners consistently became faster over the five sessions. The mean length of pauses also decreased but only when Time 1 and Time 5 were compared. While the power function was not observed in those indices, the overall changes support previous fluency studies. In de Jong and Perfetti (2011), for example, the group who repeated the same speech involving the same lexical items exhibited more signs of proceduralization than the group that recited different speeches. Those learners increased speech rate (phonation/time ratio) and decreased mean length of pauses. The current study showed that the learners who engaged in repeated practice of the same grammatical structure (*wh*-questions) but with different lexical items became faster with shorter pauses during communicative interaction.

The differential findings of the three fluency indices add to the discussion pertaining to cognitive versus utterance fluency. If we consider that the learners were in the process of proceduralization of English *wh*-questions, only speech rate and pause phenomena were susceptible to the effect of contextualized practice. Also, given the consistent increase in speech rate, it can be argued that speech rate best represents the change in the underlying cognitive mechanism (de Jong et al., 2013). However, repair phenomena were largely unrelated to the other measurements. In Freed's (1995) study that compared the fluency of L2 learners who studied at home and those who studied abroad showed that the study abroad learners produced faster speech; however, those learners produced more repair

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

markers, such as false starts, repetitions, and self-corrections. Freed (1995) concluded that the study-abroad learners "monitor their speech and self-correct along the way" (p. 141) while expressing complex structures. The results of the current study seem to support this claim in that although the learners became more accurate and faster, their repair attempts were not related to those changes.

In sum, contextualized practice yielded positive changes in both accuracy and fluency, except for repair phenomena. The effect sizes of the comparisons between Time 1 and Time 5 of error rate ($d = .72$), speech rate ($d = 1.53$), and mean length of pauses ($d = .66$) were all medium to large (Plonsky & Oswald, 2014). Regardless of the exact knowledge structure that the learners developed, it is encouraging that their communicative production became more accurate and faster, and their access to relevant memory systems became more efficient and effortless. From a practical perspective, such changes may be an L2 learning goal for both learners and teachers (DeKeyser, 2017).

CONTRIBUTION OF DECLARATIVE KNOWLEDGE

The second research question was related to learner (preexisting declarative knowledge) and practice (correct vs. incorrect productions) factors as potential predictors of the effect of contextualized practice. The results support skill-learning theories in that (a) declarative knowledge assists proceduralization at least at the initial stage of the process and (b) repeated practice has a robust effect in allowing learners to increasingly rely on procedural memory over time. Skill-learning theories, including those in cognitive psychology (Schneider & Fisk, 1983), neuroimaging (Ullman, 2001), and psycholinguistic (DeKeyser, 1997) research, explain that declarative and procedural systems are developed concurrently while the shifts in reliance on the two systems occur as a person becomes more accurate and faster during a given task. It was predicted that the development of skilled performance may be supported by declarative knowledge to develop and/or restructure the representations in the procedural system.

Interestingly, the scores of the declarative knowledge test, administered prior to engaging in contextualized practice, did not predict the extent of the practice effect on accuracy or fluency changes. This result indicates that having declarative knowledge of a grammatical structure may not be related to the development of the procedural system of that structure when practice is considered as the cause of the changes. Accordingly, it could be said that contextualized practice alone facilitates a positive change in accuracy, on the one hand. On the other hand, the result seems to challenge skill acquisition theory in that learners may not need an explicit understanding of a grammatical structure to benefit from contextualized practice. However, in the current study, all learners possessed some declarative knowledge of the target structure. Hence, it is premature to argue that practice alone is sufficient to develop procedural memory of a grammatical structure. What the results suggest is, instead, that the amount of declarative knowledge was not related to the extent to which each learner benefited from practice.

To further examine the contribution of declarative knowledge to proceduralization, we ran additional correlational analyses (Pearson, two-tailed) between the grammar test scores and error rate in Time 1, which yielded a significantly negative correlation ($r = –.36$, $p = .036$). At Time 5, the relationship became nonsignificant ($r = –.16$, $p = .348$). This means that before the intervention, the more declarative knowledge a learner

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

possessed, the more accurate his or her oral production was. However, as the learners engaged in extensive practice, the initial declarative knowledge became less relevant to their oral production. It may be the case that knowing grammatical rules are helpful for engaging in the targetlike behavior at an earlier stage of proceduralization, yet practice effects forge such a relationship. Similar patterns were observed in speech rate (Time 1: $r = .46$, $p = .035$; Time 5: $r = –.15$, $p = .403$) and mean length of pauses (Time 1: $r = –.44$, $p = .010$; Time 5: $r = –.15$, $p = .398$). Those results suggest that at the beginning of the intervention, the learners who possessed more declarative knowledge produced faster utterances with shorter pauses. Such patterns disappeared after the intervention, which indicates the robust effects of contextualized practice on the development of procedural memory. Also, the results imply that had the learners had no declarative knowledge of the target structure, the proceduralization might not have been observed as they would not have any knowledge source to monitor their production while engaging in contextualized practice.

From a pedagogical perspective, it can be said that teaching grammatical rules explicitly may help learners engage in targetlike production at the beginning. However, the amount of declarative knowledge that learners possess prior to engaging in practice may be unrelated to the individuals' rates of proceduralization. Therefore, promoting contextualized practice after the initial instruction would be recommended, without continuing explicit instruction. In other words, it may be the case that repeated contextualized practice may be sufficient to develop accurate procedural knowledge of a particular grammatical structure when learners have some declarative knowledge of that structure. However, the quality of practice seems to matter in the rate of proceduralization, which we discuss in the following section. In addition, we caution that we did not test the learners' declarative knowledge after the intervention; it is possible that declarative knowledge also developed due to the contextualized practice, which may have supported the development of procedural memory.

CORRECT VERSUS INCORRECT PRACTICE

The second regression model included the frequencies of correct and incorrect productions as the predictors. The results showed that correct productions predicted the change in accuracy. This may not be a surprising finding; the more a learner produces correct utterances, the more accurate she or he becomes, confirming the findings in the studies in which only correct responses were analyzed to investigate proceduralization (e.g., Rodgers, 2011; Suzuki & DeKeyser, 2017a). The current study extended the research agenda by investigating the impact of incorrect production. This was because L2 learners constantly produce inaccurate utterances and research has not discovered how those utterances impact L2 learning. Somewhat unexpectedly, the regression analysis showed that incorrect production negatively predicted the change of the accuracy scores.

At a glance, the results suggest that learners are better off not producing incorrect utterances. However, a careful interpretation is in order because the current study investigated practice of a complex linguistic structure in a classroom context. Despite its negative predictive power, we encourage teachers to promote contextualized L2 production as much as possible for the following reasons. First, the frequency of correct questions increased along with that of incorrect questions ($r = .361$; $p = .036$), meaning

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

that students were producing more of both forms. Besides, in the classroom, the teacher cannot force students to produce correct utterances only, especially during communicative activities. Second, as the output hypothesis states, L2 learners test their linguistic hypotheses based on their interlanguage systems by producing the L2, which helps them notice what they can or cannot produce (see Loewen & Sato, 2018). This cognitive process may have contributed to the proceduralization found in the current study, even though the incorrect productions were not related to the gain scores. Finally, the predictive power of correct production (β = .56) was higher than that of incorrect production (β = .39); hence, statistically speaking, promoting communicative usage of language would ultimately facilitate proceduralization of grammatical knowledge.

The findings related to incorrect production pinpoint the importance of feedback—another pillar of skill acquisition theory along with practice. Anderson and Schunn (2000) cautioned that without feedback during practice, "the student might wind up entrenching the wrong knowledge structures" (p. 16) and feedback at propitious moments during repeated practice promotes meaningful learning. To this end, L2 research of corrective feedback has abundantly documented that when a learner is given feedback, she or he may produce the correct version of the error, in the forms of successful uptake or modified output (Lyster, Saito, & Sato, 2013). Such evidence has been reported in the development of question formation as well (Lightbown & Spada, 1990). Had the teachers provided corrective feedback in the current study, the frequency of correct productions might have increased and that of incorrect productions decreased, which would have increased the overall effectiveness of the contextualized practice. Indeed, in Sato and Lyster (2012), EFL learners who engaged in extensive paired communicative activities improved their utterance fluency, while only those who received feedback improved accuracy. Hence, we recommend that the teacher increase opportunities whereby students engage in contextualized practice while providing corrective feedback at appropriate moments.

In terms of the changes in fluency, none of the predictors were significant in any model. Given the findings that speech rate and mean length of pauses improved over time, we can conclude that the correctness of production does not matter as far as fluency development is concerned. Simply speaking, by engaging in repeated practice, the learners became faster in executing a particular performance—asking *wh*-questions. Pedagogically, if the instructional goal is to improve fluency, the teacher does not have to be concerned about the learners' grammatical knowledge or correctness of their production. Clearly, however, an increase in fluency should not be the sole goal of L2 instruction.

CONCLUSIONS AND LIMITATIONS

The current study focused on three issues related to practice effects. First, while skill learning researchers have long argued for the importance of *meaningful* practice, often the experimental tasks have been mechanical and decontextualized. The current study operationalized "contextualized practice" and delivered it as classroom activities. Second, to investigate the practice effect on real language use, we examined changes in accuracy and fluency during communicative L2 production. Third, drawing on declarative/procedural models, the contribution of learners' declarative knowledge to

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

proceduralization was tested. While the exact nature of knowledge structures that contextualized practice (re-)structured is unknown, we conclude that contextualized practice supported the development of procedural knowledge of *wh*-questions, at least partially. At the same time, those learners may have become faster in accessing declarative knowledge of the structure.

The current study comes with ample ecological validity. The intervention activities were collaboratively developed with the teachers. The resulting activities align with L2 teachers' understanding of contextualized practice in general (see Korkmaz & Korkmaz, 2013). In addition, the intervention was delivered by the classroom teachers as part of regular classes. The drawbacks of increasing the study's ecological validity included the large attrition rate and the wide standard deviations (see Sato & Loewen, 2019, for pros and cons of classroom research). Nonetheless, given the positive results, we recommend that teachers implement activities that elicit repeated contextualized practice of a specific linguistic form, be it morphosyntactic, phonological, or pragmatic.

The limitations of the current study include, first, the lack of a control group. Without a comparison, we remain cautious to claim that the observed changes were solely due to the intervention. Second, although the frequencies of correct/incorrect productions were entered in the regression models, each learner engaged in different amounts of practice, which might have mediated the overall impact of repeated practice. Third, while the practice and outcome measures addressed the desired goal of L2 instruction (i.e., communicative production), the current study did not measure changes in comprehension. Given the evidence related to skill specificity (DeKeyser, 1997), it is possible that the learners developed productive skills only. Similarly, the grammar test might not have been an ideal measure of declarative knowledge. Future research may benefit by administering more sophisticated tests of declarative and procedural knowledge prior to and after practice. To answer a theoretical question as to whether particular L2 knowledge has been automatized—as opposed to a practical question as to whether L2 production has become faster and more accurate—a measurement that taps into the change in the underlying cognitive structure (e.g., brain scanning) would be useful. Automatization has been operationalized with a statistical technique (Segalowitz & Segalowitz, 1993), but there is a criticism of such an approach (Hulstijn et al., 2009). Fourth, we did not consider learners' individual aptitude found to moderate practice effects (see Suzuki & DeKeyser, 2017b). Future research can consider those weaknesses while considering the goal of contributing to our understanding of the role of contextualized practice in L2 learning and teaching.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263119000159

NOTES

[1]The idea of proceduralization/automatization relates to the interface positions in SLA research—whether explicit knowledge can become implicit knowledge. This theoretical debate is beyond the scope of this article, but in general we concur with Ullman and Lovelett (2018) arguing that "the explicit/implicit dichotomy is not

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

isomorphic to the declarative/procedural distinction…. Thus, it is difficult to directly compare the SLA and DP model positions" (p. 45).

[2]We considered that the absentees, resulting in the missing values in the dataset, were a random factor (missing completely at random) (see Myers, 2011). A total of 14.7% of missing values was considered acceptable for the values to be imputed (Roth, 1994).

[3]The application of missing data analyses did not alter the outlook of the dataset. The analyses with 19 participants who participated in all five sessions yielded significant differences in the same pairwise comparisons both of accuracy and fluency.

[4]We examined the articulation rate (syllable per second after removing pauses) as well. The results showed the identical statistical patterns to those of speech rate.

REFERENCES

Adams, R. (2007). Do second language learners benefit from interacting with each other? In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 29–51). Oxford, UK: Oxford University Press.

Ammar, A., Lightbown, P., & Spada, N. (2010). Awareness of L1/L2 differences: Does it matter? *Language Awareness*, *19*, 129–146.

Anderson, J. (1985). *Cognitive psychology and its implications* (2nd ed.). New York, NY: Freeman.

Anderson, J. (2005). *Cognitive psychology and its implications* (6th ed.). New York, NY: Worth Publishers.

Anderson, J., & Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. *5*, pp. 1–33). Mahwah, NJ: Lawrence Erlbaum.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*, 635–650.

Chein, J. M., & Schneider, W. (2005). Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. *Cognitive Brain Research*, *25*, 607–623.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

de Jong, N., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. Paper presented at the 6th Workshop on Disfluency in Spontaneous Speech.

de Jong, N., & Perfetti, C. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*, 533–568.

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

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*, 195–221.

DeKeyser, R. (2007a). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Lawrence Erlbaum.

DeKeyser, R. (2010). Practice for second language learning: Don't throw out the baby with the bathwater. *International Journal of English Studies*, *10*, 155–165.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). New York, NY: Routledge.

DeKeyser, R. (Ed.) (2007b). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge, UK: Cambridge University Press.

Derwing, T. (2017). L2 fluency development. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 246–259). New York, NY: Routledge.

Eichenbaum, H. (2012). *The cognitive neuroscience of memory: An introduction* (2nd ed.). Oxford, UK: Oxford University Press.

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition*. London, UK: Routledge.

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*, 384–412.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam, The Netherlands: John Benjamins.

Gade, M., Druey, M. D., Souza, A. S., & Oberauer, K. (2014). Interference within and between declarative and procedural representations in working memory. *Journal of Memory and Language*, *76*, 174–194.

Healy, A., & Bourne, L. (Eds.). (1995). *Learning and memory of knowledge and skills: Durability and specificity*. Thousand Oaks, CA: Sage.

Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: An intermediate course with examples in S-PLUS, R, and SAS*. New York, NY: Springer.

Hulstijn, J., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*, 555–582.

Jiang, N. (2012). *Conducting reaction time research in second language studies*. New York, NY: Routledge.

Kim, Y. (2012). Task complexity, learning opportunities, and Korean EFL learners' question development. *Studies in Second Language Acquisition*, *34*, 627–658.

Korkmaz, S., & Korkmaz, Ş. Ç. (2013). Contextualization or de-contextualization: Student teachers' perceptions about teaching a language in context. *Procedia-Social and Behavioral Sciences*, *93*, 895–899.

Langley, P., Laird, J., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Leont'ev, A. N. (1981). The problem of activity in psychology. In J. V. Wertsch (Ed.), *The concept of activity in Soviet psychology* (pp. 37–71). Armonk, NY: M. E. Sharpe.

Lightbown, P. (2000). Classroom SLA research and second language teaching. *Applied Linguistics*, *21*, 431–462.

Lightbown, P., & Spada, N. (1990). Focus-on-form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition*, *12*, 429–448.

Lightbown, P., & Spada, N. (2000). Do they know what they're doing? L2 learners' awareness of L1 influence. *Language Awareness*, *9*, 198–217.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*, 1247–1282.

Loewen, S., & Sato, M. (2017). Instructed second language acquisition (ISLA): An overview. In S. Loewen & M. Sato (Eds.), *The handbook of instructed second language acquisition* (pp. 1–12). New York, NY: Routledge.

Loewen, S., & Sato, M. (2018). State-of-the-arts article: Interaction and instructed second language acquisition. *Language Teaching*, *51*, 285–329.

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. P. García Mayo, J. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). Amsterdam, The Netherlands: John Benjamins.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*, 1–40.

Martin-Chang, S., & Levy, B. (2005). Fluency transfer: Differential gains in reading speed and accuracy following isolated word and context training. *Reading and Writing*, *18*, 343–376.

Martin-Chang, S., & Levy, B. (2006). Word reading fluency: A transfer appropriate processing account of fluency transfer. *Reading and Writing*, *19*, 517–542.

Mbogning, C., Bleakley, K., & Lavielle, M. (2015). Joint modelling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the stochastic approximation expectation–maximization algorithm. *Journal of Statistical Computation and Simulation*, *85*, 1512–1528.

McDonough, K., & Chaikitmongkol, W. (2010). Collaborative syntactic priming activities and EFL learners' production of *wh*-questions. *The Canadian Modern Language Review*, *66*, 817–841.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience*, *24*, 933–947.

Morgan-Short, K., & Ullman, M. (2012). The neurocognition of second language. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 282–300). New York, NY: Routledge.

Morris, C., Bransford, J., & Franks, J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, *5*, 297–310.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*, 677–711.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Ortega, L. (2014). *Understanding second language acquisition* (2nd ed.). New York, NY: Routledge.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*, *8*, 1–5.

Pavlik, P., & Anderson, J. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586.

Pedhazur, E. (1982). *Multiple regression in behavioral research*. Fort Worth, TX: Holt, Rinehard and Winston.

Petrocelli, J. V. (2003). Hierarchical multiple regression in counseling research: Common problems and possible remedies. *Measurement and Evaluation in Counseling and Development*, *36*, 9–22.

Pienemann, M., & Johnston, M. (1987). Factors influencing the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp. 45–141). Adelaide, Australia: National Curriculum Resource Center, Adult Migrant Education Program.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition*, *10*, 217–243.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*, 423–441.

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *International Review of Applied Linguistics in Language Teaching*, *49*, 295–319.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*, 537–560.

Sato, M., & Loewen, S. (2019). Methodological strengths, challenges, and joys of classroom-based quasi-experimental research: Metacognitive instruction and corrective feedback. In R. DeKeyser & G. Prieto Botana (Eds.), *Doing SLA research with implications for the classroom: Reconciling methodological demands and pedagogical applicability* (pp. 31–54). Amsterdam, The Netherlands: John Benjamins.

Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development: Monitoring, practice, and proceduralization. *Studies in Second Language Acquisition*, *34*, 591–262.

Schneider, W., & Fisk, A. (1983). Attentional theory and mechanisms for skilled performance. In R. Magill (Ed.), *Memory and control of action* (pp. 119–143). New York, NY: North-Holland Publishing Company.

Schneider, W., Dumais, S., & Shiffrin, R. (1984). Automatic and controlled processing and attention. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 1–27). London, UK: Academic Press.

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, *19*, 28–42.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. London, UK: Routledge.

Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173–199.

Segalowitz, N., & Segalowitz, S. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*, 369–385.

Serrano, R., Stengers, H., & Housen, A. (2015). Acquisition of formulaic sequences in intensive and regular EFL programmes. *Language Teaching Research*, *19*, 89–106.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*, 763–767.

Spada, N., & Lightbown, P. M. (1999). Instruction, first language influence, and developmental readiness in second language acquisition. *The Modern Language Journal*, *83*, 1–22.

Suzuki, Y., & DeKeyser, R. (2017a). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*, 166–188.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press

Suzuki, Y., & DeKeyser, R. (2017b). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*, *38*, 27–56.

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *Psychology of Learning and Motivation*, *60*, 113–189.

Ullman, M. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition*, *4*, 105–122.

Ullman, M., & 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*, 39–65.

Van Oers, B. (1998). From context to contextualizing. *Learning and Instruction*, *8*, 473–488.

Walz, J. (1989). Context and contextualized language practice in foreign language teaching. *The Modern Language Journal*, *73*, 160–168.

White, L., Spada, N., Lightbown, P., & Ranta, L. (1991). Input enhancement and L2 question formation. *Applied Linguistics*, *12*, 416–432.

https://doi.org/10.1017/S0272263119000159 Published online by Cambridge University Press