




RESEARCH REPORT

The relationship between poststimulus pause, learner proficiency, and working memory in an Elicited Imitation Task

John M. Norris^{1*} , Shoko Sasayama² and Michelle Kim³

¹Educational Testing Service (ETS) Japan, Tokyo, Japan; ²Waseda University, Tokyo, Japan; ³Educational Testing Service (ETS), Princeton, NJ, USA

*Corresponding author. E-mail: norrisjm123@gmail.com

(Received 16 November 2021; Revised 07 June 2022; Accepted 16 June 2022)

Abstract

The Elicited Imitation Task (EIT) is a popular technique for efficiently measuring global proficiency in multiple languages, and accumulated evidence indicates high reliability and strong relationships with other proficiency measures. Nevertheless, several dimensions of EIT design remain open to investigation, including the assumption that a pause is required in between the aural stimulus and oral response, to ensure processing of the input and prevent so-called parroting. This study investigated the relationship between three poststimulus pause conditions, learners' proficiency and working memory, and their EIT scores as well as their perceptions of task difficulty, mental effort, focus, and interest. Findings indicated no differences in performances or perceptions between the 0-second pause, 2-second pause, and 5-second pause conditions, and a weak relationship between EIT performance and working memory. Across all conditions, the EIT distinguished consistently among proficiency levels, correlated strongly with a criterion proficiency measure, and produced remarkably reliable scores.

Designing Elicited Imitation Tasks to measure language proficiency

The Elicited Imitation Task (EIT), in which language learners hear a series of sentences and attempt to repeat them verbatim, has become a popular approach to the efficient estimation of second language knowledge and proficiency in a variety of languages (Yan et al., 2016). While EITs have been designed to assess distinct language constructs, such as implicit knowledge of specific grammatical rules or ability to distinguish grammatical and ungrammatical constructions (e.g., Erlam, 2006; Spada et al., 2015), the prevailing construct focus in recent years has been on more holistic notions of language proficiency (Wu et al., 2022). The basic design approach to EITs for measuring proficiency is to present initially shorter (e.g., a few syllables) and then increasingly longer (e.g., up to 30 or more syllables) sentences spoken aloud, followed by the opportunity for learners to reproduce orally what they heard. Responses are generally scored for precision of repetition at the word level, that is, the proportion of words

accurately repeated from the original sentence (rather than syllables repeated, or idea units, or similar). A key assumption of the EIT is that learners must process the spoken input for both form and meaning to be able to repeat it accurately; the longer the sentences, the more challenging the task of processing the input, and the more difficult the repetition (Davis & Norris, 2021).

EITs have been widely developed and investigated, across numerous languages, and for multiple assessment purposes. A particularly popular application of EITs has been to serve as an indicator of general proficiency, or global oral proficiency, in second language acquisition research (e.g., Bowden, 2016; Gaillard & Tremblay, 2016; Tracy-Ventura et al., 2014; Wu & Ortega, 2013). Recently, EITs have also begun to appear as operational components of commercial, large-scale language tests of speaking proficiency (e.g., Bernstein et al., 2010; Davis & Norris, 2021). The popularity of EITs can be attributed both to their ease of development and delivery, and to their consistently high psychometric qualities (see review of both of these aspects in Wu et al., 2022). Designing EITs is a relatively straightforward endeavor (following Ortega et al., 2002): (a) select a series of sentences that range in syllable length from quite short (usually starting around six syllables) to relatively long (anywhere from 19 to 30 syllables), gradually increasing the number of syllables with each subsequent sentence; (b) allow the sentences to vary naturally in terms of their syntactic complexity, that is, with longer sentences containing more complex structures such as subordinate or relative clauses; (c) avoid excessively low-frequency vocabulary words and specialist or jargon terminology; and (d) audio record the sentences in a native speaker voice at a normal rate of speech. Test-takers are then instructed to listen to each sentence and repeat as much or as exactly as they can, and their responses are audio recorded and subsequently rated (or automatically scored) for accuracy of the repetition.

Following these test design guidelines, resulting EITs have been shown repeatedly to possess high reliability and strong relationships with other measures of L2 proficiency (Wu et al., 2022). In their recent meta-analysis, Kostromitina and Plonsky (2021) found average reliabilities across EITs to be very high ($r = .93$), and they found that EIT scores correlated strongly with criterion measures of L2 proficiency (average adjusted $r = .75$), including, in particular, self-assessments (adjusted $r = .81$) and standardized proficiency assessments (adjusted $r = .74$). In a recent, large-scale (with more than 500 participants representing 10 different L1s) study of a new English-language EIT, Davis and Norris (2021) found that, as expected, EIT scores correlated the highest with other measures of speaking ability (with r values ranging from .78 to .84) while also maintaining moderate to strong relationships with measures of writing (.73), listening (.68), and reading (.57). They also found a similar strength of relationship ($r = .69$) between EIT scores and a C-test, the latter typically interpreted to measure global language proficiency in the written modality (see Norris, 2018).

A key design parameter in many EITs is the introduction of an interruption of some kind, typically either a grammaticality judgment or a silent pause, in between the stimulus sentence and the opportunity to repeat. The purpose of the interruption is to limit what has been referred to as a “parroting” phenomenon (i.e., immediate repetition may allow learners to remember and parrot much of what they heard, without necessarily processing it for meaning), and to decrease the potential that differences in working memory capacity will influence EIT success. Early research by McDade et al. (1982), which showed that after a pause of 3 seconds learners could only repeat sentences accurately if they had fully comprehended them, inspired Ortega et al. (2002) to recommend a 3-second silent pause as a critical EIT design feature. Following their lead, many EITs (e.g., Bowden, 2016; Kim et al., 2016; Tracy-Ventura et al., 2014;

Wu *et al.*, 2022) have similarly included a pause of 2 to 3 seconds followed by a brief (typically .5-second) tone sound, thereby delaying sentence repetition. According to Kostromitina and Plonsky (2021), studies of EITs that adopted a poststimulus pause found stronger correlations ($r = .76$) with criterion measures of proficiency than studies that did not include a pause ($r = .63$), leading them to recommend that the optimal design of an EIT should include a pause after the stimuli. Interestingly, the length of pause intervals has varied only minimally across EITs that have operationalized it (i.e., from 2 to 3 seconds), and direct comparison of the effects of different pause intervals has not been undertaken to date.

A few studies have also investigated whether working memory capacity is related to performance on EITs that include the pause feature, in response to assertions that EIT performance relies on rote memorization and is therefore primarily a test of phonological short-term working memory (PSTM) rather than implicit language knowledge or language proficiency (e.g., Vinther, 2002). As Kim *et al.* (2016, p. 658) cogently explained, “If EITs primarily measure learners’ ability to hold stimuli in their PSTM, one would expect strong correlations between EIT performance and STM tests. If learners’ EIT performance requires the re-construction of knowledge that is beyond storing information, then a weak or nonsignificant correlation is expected.”

In Kim *et al.* (2016), a Korean EIT was investigated, consisting of the typical series of sentences that increased gradually in syllable length, and after each of which a 2-second pause plus a .5-second beep sound was inserted prior to repetition. PSTM was investigated with a forward digit span test ranging from three to nine digits. Findings indicated that EIT scores and the PSTM measure correlated only to a small and not statistically significant degree ($r = .30$), a pattern that has been replicated in the few other studies that have investigated this relationship (e.g., Okura & Lonsdale, 2012; Park *et al.*, 2020). Based on these patterns, Kostromitina and Plonsky (2021, p. 19) recommended including a poststimulus pause as part of the optimal design of EITs, rationalizing: “Overall, it seems that longer sentences and a pause before repetition may present additional cognitive load for the participants, which help limit their reliance on working memory and thereby serve to more clearly tap learner proficiency.” Extending this logic, it seems reasonable to hypothesize that EITs without any poststimulus pause would exhibit stronger relationships with measures of short-term working memory, while those with lengthier pauses would exhibit declining magnitudes of such relationships.

Another possible approach to investigating whether EIT design features have an effect on learner performance is to tap into learners’ own perceptions of the experience. To date, only a single study has incorporated learner perception data in examining factors that affect EIT performance. Wu *et al.* (2022) asked learners to rate various aspects of the EIT experience immediately following performance, including (among others) difficulty of the task, quality of their performance, and whether comprehending versus producing the sentences was more challenging. They found strong negative correlations between perceived task difficulty and EIT scores, and correspondingly strong positive correlations between perceived quality of performance and EIT scores. They also found that perceived challenges in comprehension were the best predictor of differences in EIT scores, among various possible EIT difficulty factors. This introduction of learner perception data as a lens on EIT performance opens new possibilities into understanding the potential role played by various design features. For example, under the assumption that variations in a poststimulus pause would lead to differences in EIT scores and differential relationships with working memory (attributable to differences in cognitive load, as noted by Kostromitina & Plonsky, 2021), tapping into

learner perceptions of cognitive load (Sweller, 1994)—based on indicators such as task difficulty, mental effort, and ability to focus—might shed new light on the cognitive reality of the EIT experience from the learner perspective, as it has begun to contribute in understanding performance on other types of cognitively complex language tasks (e.g., Sasayama & Norris, 2019). Similarly, gauging learner interest in the EIT experience might also serve to illuminate whether engagement in a relatively inauthentic language task is related to test-taker proficiency and performance (e.g., Purpura, 1997), a fundamental concern in adopting short-cut measures like the EIT as a proxy for global language proficiency.

To date, there has not been a direct comparison of EITs designed with or without a poststimulus pause, nor for pauses of differing lengths, in terms of possible effects on learner performance or relationship with working memory capacity. Findings from recent research into design features for other types of EITs have, surprisingly, indicated that previously assumed effects do not apply. For example, Erlam and Wei (2021) found that including a grammaticality “belief” judgment about the stimulus sentences did not have any effect on performance compared with a no-judgment condition. Additional research has suggested that EITs demonstrate quite stable psychometric properties across various possible moderating factors and design differences (e.g., Isbell & Son, 2021); that is, performance on EITs seems to rely primarily on learners’ abilities to process and repeat what they hear, regardless of other possible variations in how the tasks are designed, delivered, and scored. Whether the poststimulus pause design feature has any effect on learner performance or other qualities of EIT scores is the focus of this investigation.

The current study

The current study investigated the relationship between poststimulus pause, learners’ proficiency levels and working memory capacity, and their EIT performances as well as their perceptions of task difficulty, mental effort, ability to focus, and interest. The study was guided by the following research questions:

1. What is the relationship between the poststimulus pause length conditions and English learners’ performance on the EIT?
2. To what extent is the relationship between the poststimulus pause length conditions and English learner’s performance on the EIT moderated by their proficiency levels?
3. What is the reliability of the EIT, and does it vary by the poststimulus pause length conditions?
4. To what extent do the poststimulus pause length conditions affect English learners’ perceptions of task difficulty, mental effort, and focus?
5. To what extent is the relationship between the poststimulus pause length conditions and English learner’s performance on the EIT moderated by their working memory capacity?
6. Is there a relationship between English learners’ proficiency levels and their perceptions of interest in the EIT?

Methods

To address these questions, English language learners at distinct proficiency levels engaged in three versions of an EIT with different poststimulus pause lengths. In this

section, we describe the methodology used in the study, including (a) participants, (b) materials and instruments, (c) procedures, and (d) data scoring, coding, and analysis.

Participants

A total of 276 English language learners (174 female, 102 male) participated in the study. To include learners with a variety of proficiency levels, participants were recruited from English language programs within and outside of the United States. A total of 132 participants were studying English in the United States, while the others were studying English in Ecuador ($n = 119$), Mexico ($n = 1$), and Colombia ($n = 24$). Participants had various first languages, but the majority were native speakers of Spanish, Chinese (Mandarin or Cantonese), Japanese, or Korean. Their ages ranged from 18 to 52, with an average age of 22.55 years. Participants had studied English for 8.87 years ($SD = 5.65$) on average. For recruitment purposes, participants' proficiency levels were first estimated based on the level(s) of English language courses in which they were enrolled at their university. To gauge the proficiency level of English language courses offered at different institutions, a site coordinator at each institution was asked to provide both an estimated level of the course according to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2021) and an estimated TOEFL iBT score range for each course. According to these estimates, participants were divided into three proficiency groups (i.e., low, mid, high) prior to assignment to research forms. Table 1 presents the ranges of estimated CEFR levels (and corresponding TOEFL iBT scores based on Papageorgiou *et al.*, 2015), and sample sizes, for each of the low-, mid-, and high-proficiency groupings.

Materials and instruments

Elicited Imitation Task

Three forms of the EIT were developed in keeping with standard practices of EIT design, though with a few exceptions. Each form consisted of 10 sentence stimuli, ranging from 6 or 7 to 25 or 26 syllables, with a consecutive increase of several syllables per sentence. All sentences included in the three forms addressed a common theme. The participants were given a scenario in which they were hired as a campus tour guide and being trained to describe features related to a university campus. Within this scenario, participants were asked to repeat exactly what the trainer said. This design feature was adopted as part of a larger test development project within which it was considered essential for all test tasks to provide some degree of communicative context and purpose to enhance test-taker engagement (see Davis & Norris, 2021). Although the sentences and forms were thematically related, there was no repetition of lexical-semantic content across the items, hence no reason to believe that the common theme

Table 1. Low-, mid-, and high-proficiency groups

| Proficiency level | CEFR level | Approximate TOEFL iBT range | <i>N</i> |
|-------------------|------------|-----------------------------|----------|
| Low | A1–A2 | 41 or below | 71 |
| Mid | B1 | 42–71 | 94 |
| High | B2–C2 | 72–120 | 111 |

Note: TOEFL iBT scores were not collected for all participants in the study. The TOEFL iBT score range is provided as a frame of reference for interpreting approximate differences in participant proficiency levels.

Table 2. Sample Elicited Imitation Task stimuli from form 1, form 2, and form 3

| | Form 1 | Form 2 | Form 3 |
|-------------|---|---|---|
| Sentence 1 | Welcome to our campus. | The tour will take 1 hour. | Living in the dorm is fun. |
| Sentence 10 | It is my sincere wish that today you have developed a good sense of what studying here is like. | I hope to see you next year, when you get admitted and decide to enroll in our fantastic college. | For anyone who has questions, I'd be happy to stick around and follow up with you when we are done. |

would have any effect on the object of study in the current investigation, namely poststimulus pause length.

To the extent possible, the three forms of the EIT were designed to be equivalent not only in terms of the theme and the number of syllables but also in terms of the frequency and complexity of the vocabulary and grammar used in the stimuli (see Table 2 for sample stimuli). Vocabulary was limited to terms used to describe university campus life and physical surroundings, with jargon, excessively long words (more than four syllables), and technical or specialist terms (e.g., for particular disciplines) excluded. Grammatical, and especially syntactic, complexity was allowed to vary naturally, with longer sentences featuring compound and complex sentences including the use of subordinate clauses and embedding. To the extent possible, sentences in the same position on each form were designed to have similar syntactic complexity (e.g., for shorter stimuli, a simple sentence with a single independent clause was represented on each form; for medium-length stimuli, a compound sentence with two coordinated independent clauses was represented on each form; for longer stimuli, a complex sentence with one or more subordinate clauses was represented on each form). Note that all stimulus sentences were audio recorded by a single, female, native speaker of US English who was trained to maintain an even pace in speaking the sentences. Three unique forms (i.e., each with its own set of sentences) were necessary for this study to eliminate any practice effect as the participants repeated the task under each pause condition, as we elaborate next.

To investigate the potential role played by different poststimulus pause lengths, parallel versions of each of the three EITs were developed such that each came with different lengths of pause—0 seconds, 2 seconds, or 5 seconds—inserted after each stimulus. These three pause lengths were determined to: (a) represent the prevailing length of pause in most EITs developed to date (i.e., the 2-second pause); (b) contrast that with the elimination of a pause (i.e., the 0-second pause); and (c) extend the length to a salient but not exaggerated degree (i.e., the 5-second pause). The different conditions were operationalized through instructions that the participant should wait to repeat the sentence until a .5-second tone sound was played. For the 0-second condition, participants heard the sound immediately following the sentence, and for the other conditions the sound occurred after a 2-second or a 5-second pause. Consequently, nine versions (three forms by three pause conditions) were developed. Importantly, the order of versions experienced by participants was carefully counterbalanced to rule out any ordering, form difference, or practice effects. The order of the nine versions was also counterbalanced through systematic assignment across the three proficiency groups, so that participant proficiency levels would be equally represented. Prior to starting the EIT, participants were given instructions on how to complete the task, and they engaged in several practice items (with the corresponding pause condition) before starting each form.

Working memory test

To explore the role of working memory in learners' EIT performances, a forward digit span test (Olsthoorn *et al.*, 2014) was administered as a quick estimate of short-term working memory. In this test, participants were first shown a series of digits on a computer screen and then asked, on the following screen, to recall the sequence of digits in the order presented by clicking the corresponding numbers on a number pad on the screen. As the participant recalled the digits correctly, the number of digits increased by one until the digit sets of a particular length were recalled incorrectly twice. The participant's working memory score was determined by the number of digits in the final set that were recalled accurately, ranging from 3 to 12. Although using a single digit span test undoubtedly underrepresents the full construct of working memory, it was deemed sufficient for the purpose of estimating potential differences within the participant population and potential relationships with EIT performance under distinct pause conditions. This test was also easily administered in a self-access, computer-automated procedure, in keeping with the automated delivery of the full set of research procedures.

C-test

A C-test (Norris, 2018) was administered to all participants as a measure of their global L2 proficiency. This C-test followed the standard design of deleting the second half of every second word in coherent, paragraph-length texts. The C-test consisted of two texts, and each had 20 blanks for a total of 40 blanks. Participants were given 7 minutes to complete each text. Cronbach's alpha for the 40-item C-test was .92.

EIT perception and posttask questionnaires

After completing each form of the EIT, participants were given a short questionnaire that asked: (a) "How difficult was this version of the listen and repeat task for you?"; (b) "How much brainpower or effort did you use in doing this version of the listen and repeat task?"; and (c) "To what extent were you able to focus/concentrate on doing this version of the listen and repeat task?" Participants responded on a 7-point Likert scale, ranging from 1 (*Very easy, Very little, or Not at all focused*) to 7 (*Very difficult, A lot, or Highly focused*). After completing the third and final form of the EIT (and answering the three questions), in the EIT posttask questionnaire participants were asked whether they noticed any difference among the three sets. Lastly, they were asked "How interesting was this listen and repeat task for you?" and why it was or was not interesting for them. The EIT questionnaires were prepared in five languages, including English, Spanish, Chinese, Japanese, and Korean to make sure that the participants—especially those with low levels of English proficiency—were able to understand and answer the questions in their native language.

Background questionnaire

Participants filled out a background questionnaire to provide demographic information, including questions on gender, age, and duration of English language learning.

Procedures

Before participating in the experiment, participants completed an eligibility survey to provide information about their English proficiency levels, including English course(s) in which they were enrolled, standardized English language assessment scores if available, and self-assessment of their English proficiency levels (A1 through C2 on the CEFR). To elicit self-assessment of their CEFR levels, participants were presented with the CEFR global scale (i.e., a short description of what learners at each proficiency level should be able to do; Council of Europe, 2001) and were asked to choose the level that best described their English ability. For participants studying outside of the United States, the survey was fully translated into Spanish. Participant eligibility was determined based on whether they were currently enrolled in English-language courses for which proficiency levels had been determined (see previous description), whether they had access to a computer that met the study requirements (including audio recording capacity in particular), whether they could complete the study in one sitting in a quiet location without being disturbed, and whether they fit our sampling needs for different proficiency levels. A few candidates were deemed ineligible for one or more of these reasons.

Once participants' eligibility was confirmed, they were invited to participate in the study. Each participant took part in the study online at home, using their own computer. The participants were asked to complete all tasks alone in a quiet space without any aids. They accessed the study through a link to an ETS-proprietary online platform that then guided them automatically through all steps in the study, including management of time spent on each task. Once the study began, the entire process was controlled by the computer program; time available to complete each EIT task, the C-test, and the working memory test was controlled by the computer, whereas time allowed to complete questionnaires was left flexible. The study consisted of two parts; in Part 1, each participant completed three forms of the EIT with different pause length conditions (0, 2, or 5 seconds), the EIT perception questionnaire following each form, and the EIT posttask questionnaire (as well as two e-mail writing tasks not discussed in this article). Importantly, participants' sentence repetitions were automatically audio recorded and submitted using the internet to a master server at ETS as each EIT form was completed. After Part 1, the participants were given a mandatory 10-minute break before they were able to move on to Part 2. In Part 2, participants completed the short background questionnaire, the C-test, and the working memory test. On average, the whole session (Parts 1 and 2 combined) took 60 to 90 minutes.

Data scoring, coding, and analysis

Elicited Imitation Task

Prior to scoring, all audio-recorded repetitions were reviewed to make sure participants had followed instructions and attempted to record a legitimate response for each sentence. Four participants were found to have attempted to benefit from assistance of another person, and their data were eliminated. Another 17 participants encountered technical difficulties—attributable to internet connectivity challenges—such that one or more of their responses was not recorded. Their data were also eliminated from analysis, such that only participants with complete data sets were included.

Scoring rubrics were developed to assess how accurately the participants were able to repeat the stimuli verbatim. Based initially on the rubrics proposed by Ortega

et al. (2002), the scale and descriptors were modified somewhat to provide for a more precise estimation of repetition accuracy on all stimuli and to enhance rater reliability. These adjustments were made in conjunction with a large-scale test development project, where raters would also need to efficiently rate hundreds of responses at a time (see Papageorgiou et al., 2021). The rubrics in this study represented two aspects of performance: (a) content (i.e., to what extent the stimulus was repeated verbatim) and (b) intelligibility (i.e., to what extent the repetition could be understood by a listener). A six-point rating scale ranged from Score 5 (highest) to Score 0 (lowest). Descriptions of content and intelligibility were developed iteratively to characterize the degree of successful repetition across the six levels. A fully successful score 5 response was characterized as a response that was an exact repetition of the stimulus and fully intelligible. If the response included minor changes in words or grammar, or minor ambiguities related to intelligibility, but otherwise fully captured the meaning expressed in the stimulus, the response was awarded a score 4. A score of 3 was awarded for a response that was a full sentence containing most of the original content but for which the meaning of the prompt sentence was not captured entirely, either due to missing ideas or unintelligible components. A score of 2 indicated that a substantial portion of the original content was missing or highly inaccurate or highly unintelligible, and a score of 1 captured very little of the original prompt or was largely unintelligible. A score of 0 was given when the participant provided either no response, no English response, or if the content was unrelated to the prompt. Note that a version of the scoring rubric that is used for operational testing purposes is publicly available (Educational Testing Service, 2021).

Using these rubrics, the EIT responses were double scored by four raters. Raters first reviewed the rubrics with sample responses at each score band and then participated in a calibration session where they practiced scoring responses and discussed discrepancies. Each individual response was then scored by two raters. The scores between the two raters for each response were averaged, and each averaged score on the 10 stimuli was totaled to come to the final score (total possible $k = 50$). In rating responses, the raters were largely in agreement. On the 6-point scale, the ratings given by pairs of raters were either exactly the same (68%) or adjacent (i.e., within \pm one point) for 98% of the responses scored.

Working memory test

Each participant's verbal working memory capacity was determined by the maximum length of digits recalled correctly before having made two incorrect responses on the digits of a particular length. If participants incorrectly recalled a sequence of *three* digits (the first stimulus length presented on the test) two times, their working memory capacity was determined as NS (nonscorable). The possible working memory scores ranged from 3 to 12.

C-test

The 40-item C-test was automatically scored using an exact-response approach for each blank. A participant was given one point for each blank where they were able to enter the correct missing letters. Each text included a total of 20 blanks, with 40 being the maximum score possible.

EIT questionnaire

Participants' responses to the questions about difficulty, effort, and focus were analyzed by calculating average ratings for each question on each of the forms experienced (i.e., the 0-, 2-, or 5-second pause conditions). Their responses to the question about interest that was asked only once at the end of the EIT session were analyzed by calculating the average ratings overall and by proficiency level. Regarding the yes-no question about whether they noticed any difference among the three sets, the number of responses for each answer choice were tallied separately for each pause condition.

Statistical analyses

Statistical analyses were conducted on complete data sets available for the measures implicated in each research question. Of the 276 participants who completed all steps in the study, 21 were removed for the reasons mentioned previously, and an additional nine participants received nonscorable as their verbal working memory scores. Thus, a total of 255 participants' EIT responses were analyzed for most research questions, and a total of 246 participants' responses were analyzed for examining the relationship with working memory.

Statistical analyses focused on comparing mean values for experimental conditions and groups on the different measures, both descriptively and inferentially. All data were checked for their distributional properties and found to meet the minimal expectations for conducting analyses within the ANOVA family of statistics. An experiment-wise alpha level was set at $p < .05$ for all statistical tests, and Bonferroni adjustments were made where multiple tests of the same data set were involved. Cronbach's alpha statistics were calculated to investigate the reliability of each version of the EIT, and Pearson product-moment correlation coefficients were calculated to examine the relationships among different measures. All statistical analyses were conducted in SPSS v. 27.

Results

To answer the first research question, the potential effect of poststimulus pause length differences was analyzed by comparing average EIT scores across the three conditions (i.e., pauses of 0, 2, and 5 seconds). Table 3 provides descriptive statistics of the EIT scores for the three proficiency groups and overall across the three pause conditions. Comparing the overall scores on the EIT ($k = 50$), it is apparent that pause length had no discernible effect, with less than one point of difference across the three conditions. Similar patterns were observed within each of the proficiency groups. By contrast, the proficiency groups differed substantially, by approximately 6 score points between each level, and this difference was consistent across the three pause conditions.

To confirm this initial observation, a repeated-measures analysis of variance was conducted for EIT scores, with pause condition serving as the repeated factor. The analysis indicated no statistically significant difference across mean EIT scores in the three conditions (Wilks's Lambda $F = 2.18 (2, 252), p = .121, \eta_p^2 = .017$) and no statistically significant interaction between pause length and proficiency group (Wilks's Lambda $F = .725(4, 502), p = .575, \eta_p^2 = .006$). However, a follow-up univariate analysis indicated a statistically significant difference between the three proficiency groups ($F = 61.18 (2, 252), p = .000, \eta_p^2 = .327$). Figure 1 shows the clear and consistent differences in mean EIT scores between the three proficiency groups, with

Table 3. Descriptive statistics for EIT scores by pause condition and proficiency group

| | | Average EIT Score | | |
|-------------|--------------|-------------------|---------------|------------|
| | | Mean | SD | N |
| 0 sec pause | High | 36.114 | 6.3497 | 105 |
| | Mid | 29.546 | 7.6183 | 87 |
| | Low | 23.246 | 10.1401 | 63 |
| | Total | 30.694 | 9.3602 | 255 |
| 2 sec pause | High | 36.190 | 6.0385 | 105 |
| | Mid | 30.195 | 7.4195 | 87 |
| | Low | 23.214 | 9.9599 | 63 |
| | Total | 30.939 | 9.1746 | 255 |
| 5 sec pause | High | 36.200 | 5.9297 | 105 |
| | Mid | 30.402 | 8.4061 | 87 |
| | Low | 24.254 | 9.7761 | 63 |
| | Total | 31.271 | 9.1741 | 255 |

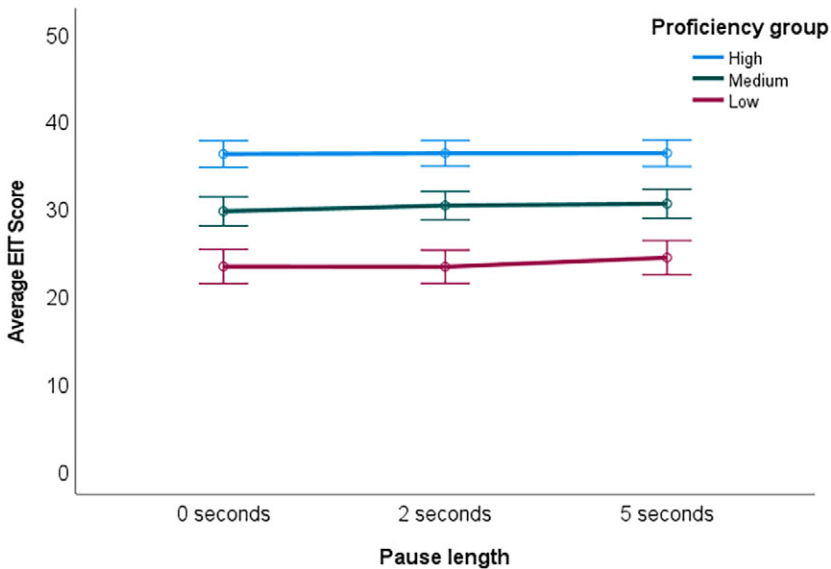


Figure 1. Average EIT scores by pause condition and proficiency group.

nonoverlapping 95% confidence intervals between the three groups on all conditions. It is also apparent that pause length had no discernable effect on average EIT performances.

The question of whether pause length affected the reliability of the EIT was answered by calculating Cronbach’s alpha reliability estimates for each of the nine versions (three forms by three pause conditions). Recall that participants were assigned systematically to each version, such that the three proficiency groups were equivalently represented on each. The three forms of the EIT were all found to produce remarkably reliable scores, and reliability estimates exhibited almost no variation in relation to pause length condition (see Table 4). Note that similar, high reliability estimates are common for

Table 4. Cronbach’s alpha reliability estimates for EIT forms and pause conditions

| | Pause length | | |
|--------|--------------|-------|-------|
| | 0 sec | 2 sec | 5 sec |
| Form 1 | .94 | .91 | .92 |
| Form 2 | .93 | .93 | .93 |
| Form 3 | .92 | .93 | .92 |

EIT tests (e.g., Kostromitina & Plonsky, 2021), and even shorter tests (e.g., with as few as 5–7 sentences) have demonstrated very good reliability (see Davis & Norris, 2021).

Turning to participant perceptions, the potential effect of pause length was also investigated by comparing participant ratings of task difficulty, mental effort, and focus required to complete the test under each pause condition. A multivariate repeated-measures analysis of variance was conducted for the EIT perception ratings of participants in each of the three proficiency groups, with pause length serving as the repeated factor and difficulty, mental effort, and focus serving as the perception measures. The analysis indicated no statistically significant effect of pause condition on the three perception measures (Wilks’s Lambda $F = 1.066 (6, 247), p = .384, \eta_p^2 = .025$) and no statistically significant interaction between pause length and proficiency group (Wilks’s Lambda $F = 1.128 (12, 494), p = .335, \eta_p^2 = .027$). However, as with the EIT score data, the analysis did indicate a statistically significant between-subjects effect for proficiency group (Wilks’s Lambda $F = 9.518 (6, 500), p = .001, \eta_p^2 = .103$). Tables 5–7 show the descriptive statistics for each perception measure. The effect of participants’ proficiency is clear in the consistently increasing perception ratings by group for each of the three measures: the EIT was perceived to be more difficult, to require more mental effort, and to call for greater focus the lower the proficiency level. Pause length, by contrast, had almost no effect on the three perception measures.

A follow-up question on the final questionnaire asked whether learners noticed any differences between the three EIT forms they completed. Of the 255 participants, 174 of

Table 5. Means (SDs) for participant ratings of EIT difficulty

| | Pause length | | |
|--------|--------------|-------------|-------------|
| | 0 sec | 2 sec | 5 sec |
| High | 4.30 (1.44) | 4.32 (1.67) | 4.38 (1.60) |
| Medium | 4.98 (1.29) | 4.94 (1.37) | 4.94 (1.34) |
| Low | 5.79 (1.31) | 5.70 (1.20) | 5.63 (1.30) |
| Total | 4.90 (1.48) | 4.87 (1.56) | 4.88 (1.52) |

Table 6. Means (SDs) for participant ratings of EIT mental effort

| | Pause length | | |
|--------|--------------|-------------|-------------|
| | 0 sec | 2 sec | 5 sec |
| High | 4.64 (1.54) | 4.66 (1.57) | 4.63 (1.58) |
| Medium | 5.29 (1.33) | 5.36 (1.29) | 5.36 (1.30) |
| Low | 5.68 (1.26) | 5.63 (1.17) | 5.60 (1.23) |
| Total | 5.12 (1.46) | 5.14 (1.45) | 5.12 (1.46) |

Table 7. Means (SDs) for participant ratings of EIT focus

| | Pause length | | |
|--------|--------------|-------------|-------------|
| | 0 sec | 2 sec | 5 sec |
| High | 4.78 (1.62) | 4.81 (1.42) | 4.78 (1.43) |
| Medium | 4.97 (1.51) | 4.98 (1.49) | 5.10 (1.45) |
| Low | 5.70 (1.20) | 5.14 (1.34) | 5.46 (1.26) |
| Total | 5.07 (1.53) | 4.95 (1.43) | 5.06 (1.41) |

Table 8. Descriptive statistics for C-test and working memory scores

| | Proficiency group | Mean | SD | N |
|--------------|-------------------|--------------|--------------|------------|
| C-test score | High | 31.15 | 6.443 | 103 |
| | Mid | 25.99 | 8.218 | 84 |
| | Low | 20.54 | 7.587 | 59 |
| | Total | 26.84 | 8.449 | 246 |
| WM score | High | 7.62 | 1.991 | 103 |
| | Mid | 6.82 | 1.857 | 84 |
| | Low | 7.14 | 2.338 | 59 |
| | Total | 7.23 | 2.058 | 246 |

Note: C-test scores are from a total $k = 40$; WM scores are the number of digits accurately recalled.

them said that they detected a difference between the three forms of the EIT. Interestingly, however, only 62 (24% of the entire participant population) identified the difference to be the poststimulus pause length. Others attributed the difference to the sentence stimuli per se, ranging from the length, complexity (vocabulary, structure), or pronunciation of the sentence stimuli, to the speed of presentation or the topic covered; however, there did not seem to be any pattern of differences systematically attributed to a given form of the EIT.

To investigate the relationship between EIT scores and other measures of English proficiency and working memory, scores on the C-test and working memory test were first calculated and compared across the three proficiency groups. Table 8 shows an evident increase in C-test scores from low- to mid- to high-proficiency groups, while working memory test scores differed minimally between the three groups. Bonferroni-adjusted univariate tests for the two measures indicated statistically significant differences between the three proficiency groups on average C-test scores ($F = 39.74$ (2, 243), $p = .000$, $\eta_p^2 = .260$) and no statistically significant differences between the three proficiency groups on working memory scores ($F = 3.657$ (2, 243), $p = .027$, $\eta_p^2 = .029$).

Table 9. Pearson correlations between EIT performance, proficiency, working memory, and interest

| | C-test | Working Memory | EIT Interest |
|-----------------------|--------|----------------|--------------|
| Working Memory | .130 | | |
| EIT Interest | -.127 | -.157 | |
| 0-sec pause EIT score | .609 | .147 | -.123 |
| 2-sec pause EIT score | .600 | .206 | -.138 |
| 5-sec pause EIT score | .628 | .239 | -.083 |

Note: All correlations statistically significant, $p < .05$.

Table 10. Participant interest in the EIT

| Proficiency group | Interest Rating | | |
|-------------------|-----------------|------|-----|
| | Mean | SD | N |
| High | 4.39 | 1.93 | 105 |
| Mid | 4.94 | 1.82 | 87 |
| Low | 5.24 | 1.70 | 63 |
| Total | 4.79 | 1.87 | 255 |

To examine the relationship between these external measures and EIT scores under the three pause length conditions, Pearson correlations were calculated between EIT scores in each condition and scores on the C-test and working memory test. Table 9 shows that EIT scores correlated strongly with the C-test, and very little difference was observed in the magnitude of these correlations between the three pause conditions. In terms of working memory, while the correlations overall were positive but weak, there were small but noticeable differences in magnitude between the three pause conditions. As pause length increased, the relationship between EIT scores and working memory scores also increased, with the strongest correlation observed for the 5-second pause condition.

One other set of findings rounds out the current study. Participants were asked to rate their interest in completing the EIT on a scale from 1 (not at all interesting) to 7 (very interesting). As shown in Table 10, overall, participants rated the EIT above the mid-point of the scale, indicating moderate interest. However, interest ratings increased noticeably with a decrease in proficiency level, with the high and low proficiency groups differing by almost a full point on the scale, and a univariate analysis of variance indicated this difference to be statistically significant ($F = 4.643$ (2, 252), $p = .010$, $\eta_p^2 = .036$). Similarly, as shown in Table 9, small negative correlations were found between interest in the EIT and both C-Test scores and working memory test scores. Overall interest in the EIT was also weakly negatively correlated with performance under each of the pause conditions.

Discussion

The primary finding of the current study is that providing a pause of 2 or 5 seconds in between the EIT stimulus and response did not have any discernable effect on EIT performance compared with providing no pause. This lack of effect was also observed consistently across three distinct proficiency levels of English learners. Furthermore, the presence or absence of poststimulus pause had no effect on reliability estimates for the three EIT forms, nor on participants' perceptions of EIT difficulty or the mental effort and focus required. This finding is surprising in light of the prevailing wisdom in EIT design, and it runs counter to the meta-analytic indications of psychometric favorability in EITs that feature a poststimulus pause.

There are several possible explanations for the lack of effect of the pause conditions. In general, effects of degraded short-term memory are presumed to be initially detectable somewhere within the 15–30 second window (Atkinson & Shiffrin, 1971; Corkin, 2013), indicating the possibility that a pause of 2 or 5 seconds might not have been enough to challenge the memory capacity of participants and influence their repetition performance. The fact that the different pause lengths were only registered by

Table 11. Average performance ratings for EIT stimuli of differing lengths

| | Pause length | | |
|----------------------------|--------------|-------|-------|
| | 0 sec | 2 sec | 5 sec |
| Item #1 (6–7 syllables) | 4.6 | 4.5 | 4.5 |
| Item #10 (25–26 syllables) | 2.2 | 2.2 | 2.2 |

a quarter of the participants supports this interpretation to some extent, though only for some of the learners. What is most surprising, though, is that there did not seem to be any “parrotting” advantage (at all) for performance in the 0-second pause condition. It might be speculated that a parrotting advantage would be most apparent for longer-stimulus sentences in the 0-second condition, given that the shorter stimuli would be easier to recall regardless of pause or no pause prior to response. However, the data do not support that interpretation either. As shown in Table 11, the longest stimuli (item #10 sentences in each set) resulted in nearly identical performance ratings on all three pause conditions, as did the shortest stimuli (item #1 sentences).

Interestingly, there did seem to be a small degree of relationship between working memory and EIT performance overall, but contrary to expectations, that relationship increased slightly as pause length increased. Thus, participants with higher working memory scores did perform somewhat better than participants with lower working memory scores in all conditions, and most noticeably in the 5-second pause condition. This pattern runs counter to the hypothesized relationship between working memory and EIT performance. That is, the poststimulus pause has been introduced into EIT design precisely to mediate or eliminate the effect of working memory (McDade *et al.*, 1982; Ortega *et al.*, 2002; Park *et al.*, 2020), and the expectation, as expressed by Kim *et al.* (2016), would be that the correlation would decrease as pauses are introduced, especially in the 5-second pause condition. The observation that the strength of relationship increased raises some doubts regarding the presumed role played by the poststimulus pause, suggesting that: (a) working memory played some small role in EIT performance regardless of whether there was a pause or not, and (b) learners in the zero-pause condition were likely engaging in the processing and reconstruction of stimulus sentences rather than relying on working memory to parrot what they heard. However, it is important to highlight that the amount of variability in scores accounted for by the relationship with working memory was very small, even for the 5-second pause condition ($R^2 = .057$). Therefore, working memory did not seem to be influencing EIT performance to any substantial degree, a pattern replicated across multiple studies to date (Kim *et al.*, 2016; Park *et al.*, 2020), and the current study suggests that to be the case regardless of whether there is a poststimulus pause or not.

What did account for performance patterns on the EIT most clearly was the combination of learner proficiency differences and EIT design. In this study, three 10-item EIT forms were developed following the design principle of beginning with a short stimulus (6–7 syllables) and increasing the length by two syllables for each subsequent stimulus (up to 25–26 syllables). These three EIT forms, represented in each of the three pause conditions, produced remarkably and equivalently reliable scores, with Cronbach’s alpha estimates ranging from .91 to .94. The different versions of the EIT also clearly and comparably distinguished among three levels of learner proficiency. Within each pause condition, the EIT consistently separated low-, medium-, and high-proficiency learners by six score points between each level,

providing considerable criterion-related validity evidence of the EIT's capacity to distinguish proficiency differences. Additional concurrent validity evidence was provided by the substantial relationship between EIT scores (again, in each pause condition) and another proficiency measure, the C-test, with correlations ranging from .600 to .628. These robust psychometric patterns further reinforce what has been observed repeatedly for EITs: when constructed following basic design parameters (e.g., Ortega et al., 2002), they prove to produce remarkably reliable scores that are also persistently strongly related to other measures of global speaking proficiency (see also Isbell & Son, 2021). For further validity evidence related to the EIT design and scoring approach adopted in the current study, see Davis and Norris (2021).

A final interesting dimension to the relationship between learner proficiency differences and the EIT design has to do with how learners perceived the EIT. Consistently, the lower the learner proficiency level was, the higher were their ratings of EIT difficulty and required mental effort and focus, and—perhaps surprisingly—the higher their ratings of interest in performing the EIT. These patterns, too, were unrelated to whether or not there was a poststimulus pause. That the test overall was perceived to be more challenging by the lower-proficiency learners provides additional validity support for the EIT design, reflecting the reality that longer stimulus sentences are more difficult for learners who have less capacity to process the full input for both form and meaning, and then to reproduce it. That fundamental reality, indicated in performance patterns as well as learner perceptions (equivalently across all pause conditions), highlights the capacity of the EIT to tap into something deeper and more indicative of language proficiency than mere parroting. That this increased challenge (i.e., indicated by higher perceived difficulty and lower scores among the low-proficiency participants) was also associated with increased interest suggests that, contrary to common perceptions about language tests, presenting learners with challenging test tasks is not necessarily a bad or demotivating thing to do. Indeed, the overall positive interest in the EIT (coupled with generally high ratings of difficulty, effort, and focus) expressed by participants in this study, and the increasing interest levels for lower proficiency learners, provides at least some additional evidence of learner “buy-in” to the design of the EIT as a means for eliciting L2 speaking ability.

Limitations

Several factors limit generalizations based on the current study. First, although of sufficient size and variability in English L2 proficiency, the participant sample was one of convenience. Several idiosyncratic characteristics were not controlled for, such as first language, age, language learning experiences, and reason for participating, and these may have had an effect on how the participants engaged in the various research activities and resulting patterns. Second, the forward digit span test adopted in the study provides a limited representation of short-term working memory, and alternative measures might have enabled more robust interpretations about this dimension of EIT performance. Third, to eliminate a practice effect, the three forms of EIT differed in terms of the sentence stimuli presented. Although general design parameters were followed to create maximally parallel forms, it is possible that the forms were differentially difficult, leading to different performances by the participants. At the same time, the presentation of the forms was also carefully counterbalanced, both across the pause-length conditions and across the proficiency groupings, thereby reducing any effects attributable to differences in the difficulty of the forms. Fourth, most critically,

individual participants completed the entire experiment in an unsupervised, self-access format. While efforts were made to ensure engagement and completion of all steps in the study (e.g., instructions at the beginning of the study, checking of completion, elimination of noncompliant participants), there was no way to control participants' actual efforts to pay attention, try their best, or provide honest/accurate answers to questionnaires.

Conclusion

Despite limitations, the findings in this study clearly indicated no differences in learner performances or perceptions among 0-, 2-, or 5-second pause conditions inserted between EIT stimulus sentences and responses. While a small positive relationship was identified between short-term working memory and EIT performance, that relationship was found to be weakest for the 0-second pause condition and strongest for the 5-second pause condition, contrary to expectations. Rather than working memory effects or poststimulus pause conditions, the relationship between learner proficiency differences and the fundamental EIT design seems to have been the overriding factor determining patterns of test scores in the current study. These findings support the use of a well-designed EIT as a robust and reliable indicator of second language speaking proficiency differences, and they also call into question the assumption of a parroting effect in EIT performance, suggesting that a poststimulus pause may be unnecessary.

Acknowledgments. We would like to acknowledge the important contributions made to this study by the site coordinators and participating English learners, as well as by our ETS colleagues Larry Davis, Michael Ecker, Jeremy Lee, Jung Aa Moon, Renka Ohta, and Michael Suhan.

References

- Atkinson, R. C., & Shiffrin, R. M. (1971). The control processes of short-term memory. *Institute for Mathematical Studies in the Social Sciences*, Stanford University.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27, 355–377.
- Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish elicited imitation task. *Studies in Second Language Acquisition*, 38, 647–675.
- Corkin, S. (2013). *Permanent present tense: The unforgettable life of the amnesic patient, HM* (Vol. 1000). Basic Books.
- Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching and assessment*. Cambridge University Press.
- Davis, L., & Norris, J. M. (2021). *Developing an innovative Elicited Imitation Task for efficient English proficiency assessment* (TOEFL Research Report RR-96). Educational Testing Service.
- Educational Testing Service. (2021). TOEFL Essentials Listen & Repeat Scoring Guide. <https://www.ets.org/s/toefl-essentials/rsc/pdf/speaking-rubric.pdf>
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27, 464–491.
- Erlam, R., & Wei, L. (2021). The importance of increased processing demands in the design of Elicited Imitation tests. *Language Teaching Research*. Advance online publication. <https://doi.org/10.1177/13621688211026032>
- Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, 66, 419–447.
- Isbell, D. R., & Son, Y. A. (2021). Measurement properties of a standardized Elicited Imitation Test: An integrative data analysis. *Studies in Second Language Acquisition*. Advance online publication. <http://doi.org/10.1017/S0272263121000383>

- Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an Elicited Imitation Test for SLA research. *The Modern Language Journal*, 100, 655–673.
- Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*. Advance online publication. <http://doi.org/10.1017/S0272263121000395>
- McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders*, 47, 19–24.
- Norris, J. M. (2018). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7–33). Peter Lang.
- Okura, E., & Lonsdale, D. (2012). *Working memory's meager involvement in sentence repetition tests*. In Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 2132–2137). Lawrence Erlbaum.
- Olsthoorn, N. M., Andringa, S., & Hulstijn, J. H. (2014). Visual and auditory digit-span performance in native and non-native speakers. *International Journal of Bilingualism*, 18, 663–673.
- Ortega, L., Iwashita, N., Norris, J.M. and Rabie, S. (2002, October) *An investigation of elicited imitation tasks in crosslinguistic SLA research*. Paper presented at the Second Language Research Forum, Toronto, Canada.
- Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL Essentials test 2021* (Research Memorandum No. RM-21-03). Educational Testing Service.
- Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Educational Testing Service.
- Park, H.I., Solon, M., Henderson, C., & Dehghan-Chleshtori, M. (2020). The roles of working memory and oral language abilities in Elicited Imitation performance. *The Modern Language Journal*, 104, 133–151. <https://doi.org/10.1111/modl.12618>
- Purpura, J. E. (1997). An analysis of the relationships between test-takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, 47, 289–325.
- Sasayama, S., & Norris, J. M. (2019). Unravelling cognitive task complexity: Learning from learners' perspectives on task characteristics and second language performance. In Z. Wen & M. Ahmadian (Eds.), *Researching second language task performance and pedagogy: Essays in honor of Peter Skehan* (pp. 95–132). John Benjamins.
- Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, 65, 723–751.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4, 295–312.
- Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014). “Repeat as much as you can”: Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Proficiency assessment issues in SLA research: Measures and practices* (pp. 143–166). Multilingual Matters.
- Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, 12, 54–73.
- Wu, S., Tio, Y., & Ortega, L. (2022). Elicited imitation as a measure of L2 proficiency: New insights from a comparison of two L2 English parallel forms. *Studies in Second Language Acquisition*, 44, 271–300.
- Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, 46, 680–704.
- Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, 33, 497–528.

Cite this article: Norris, J. M., Sasayama, S., and Kim, M. (2023). The relationship between poststimulus pause, learner proficiency, and working memory in an Elicited Imitation Task. *Studies in Second Language Acquisition*, 45: 1370–1387. <https://doi.org/10.1017/S0272263122000274>