Applied Psycholinguistics, page 1 of 30, 2016
doi:10.1017/S0142716416000084

Exploratory research on second language practice distribution: An Aptitude × Treatment interaction

YUICHI SUZUKI
Kanagawa University

ROBERT DEKEYSER
University of Maryland College Park

Received: June 7, 2015 Accepted for publication: January 30, 2016

ADDRESS FOR CORRESPONDENCE
Yuichi Suzuki, Faculty of Foreign Languages, Kanagawa University, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa 221–8686, Japan. E-mail: szky819@kanagawa-u.ac.jp

ABSTRACT
The current exploratory study aimed at investigating the role of cognitive aptitudes in determining the effect of practice distribution on second language learning. The study investigated to what extent language-analytic ability and working-memory capacity predicted the acquisition of grammar under two learning conditions that differ in the interval between the two training sessions. Learners of Japanese as a second language were trained on an element of Japanese morphosyntax under either distributed practice (7-day interval) or massed practice (1-day interval). The results revealed that language-analytic ability was only related to performance after distributed practice, whereas working-memory capacity was only related to performance after massed practice. These Aptitude × Treatment interaction findings can help establish the learning processes operating under distributed/massed practice conditions.

Many studies have investigated ways of enhancing the effectiveness of second language (L2) grammar learning. One line of investigation has explored this question by comparing different types of learning and feedback conditions, such as types of corrective feedback (Li, 2010), explicit and implicit treatments (Norris & Ortega, 2000; Spada & Tomita, 2010), and comprehension-based versus production-based instruction (Shintani, Li, & Ellis, 2013). A relatively unexplored area of research is the effect of distribution of practice in L2 grammar learning: whether different intervals between multiple practice sessions facilitate the retention of the targeted knowledge. A large body of literature in cognitive psychology suggests that the ratio of intersession interval (ISI; the amount of time between the practice sessions) to retention interval (RI; the amount of time between the end of practice and the testing time) influences the skill acquisition process (see Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Rohrer, 2015, for review in the psychology literature; see Serrano, 2012, for review in second language acquisition [SLA]). In the area of L2 grammar learning, a few empirical studies have investigated whether an optimal ratio of ISI to RI facilitates learning (Bird, 2010; Miles, 2014). These findings suggest that, even with the same amount of study time, L2 grammar learning can be enhanced by optimal scheduling of practice time. This topic is particularly important in foreign language learning classrooms where the amount of in-class practice is often limited.

As renewed attention has recently been given to cognitive aptitudes for L2 learning (Carroll, 1981; Granena, 2015; Li, 2015; Linck, Hughes, et al., 2013; Skehan, 2012, 2015), an emerging line of investigations has explored how individual differences in cognitive aptitudes moderate the effectiveness of different types of L2 instruction (e.g., Brooks, Kempe, & Sionov, 2006; Erlam, 2005; Goo, 2012; Robinson, 1997; Yilmaz, 2013). It has been shown that some L2 instructional treatments are more or less effective for particular individuals, depending on their aptitude profile (Doughty, 2013). Examining Aptitude × Treatment interactions (ATI) will ultimately inform us about how we can enhance the effectiveness of instruction by matching instruction to the learner's aptitudes (Cronbach & Snow, 1977). Furthermore, ATI can potentially unveil the underlying L2 learning processes if certain aptitudes are found to play a different (facilitative/inhibitory) role with different treatments (DeKeyser, 2012; Robinson, 2002).

Following this line of investigations, the current study aims to explore to what extent individual differences moderate the effectiveness of different levels of distribution of practice. It examines the role of individual differences in the two primary components of aptitudes, language-analytic ability (LAA) and working memory capacity (WMC), on L2 practice distribution. Japanese as L2 learners were trained on an element of Japanese morphosyntax, accompanied with vocabulary learning, in an explicit step-by-step manner, in either distributed practice (7-day interval) or massed practice (1-day interval). To the best of our knowledge, no prior research has examined the distributed/massed practice from an ATI perspective; the current study, exploratory in nature, is a first attempt to better understand this underresearched area.

In what follows, we will review the role of cognitive aptitudes in L2 grammar learning, with a particular focus on the two most researched aptitude components, LAA and WMC. We then point out the relevance of investigating a combination of LAA and WMC for distributed/massed learning and lay out the design of the current study.

LAA AND L2 GRAMMAR LEARNING

One of the most well-known aptitude test batteries is the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1959). Among the components identified in the MLAT, Carroll (1981) suggested that two have found to be especially important for L2 grammar learning: grammatical sensitivity and inductive learning ability. Grammatical sensitivity can be measured by a subtest in the MLAT (words in sentences), whereas the MLAT does not have a subtest that specifically targets the inductive learning ability. A more recent aptitude test, the LLAMA test (Meara, 2005), for instance, has the Llama_F subtest for measuring the inductive learning ability. In the current paper we consider both components to be measures of LAA, based on Skehan's (1998) theorization of aptitude. LAA is defined as the capacity "to infer rules of language and make linguistic generalizations or extrapolations" (Skehan, 1998, p. 204).

LAA has been found to play a role in L2 grammar learning (De Graaff, 1997; Erlam, 2005; Hwu & Sun, 2012; Li, 2013; Ranta, 2002; Robinson, 1997; Sheen, 2007; Shintani & Ellis, 2015; Trofimovich, Ammar, & Gatbonton, 2007; Wesche, 1981; Yilmaz, 2013). Robinson (1997) investigated how LAA moderated L2 grammar learning under four types of learning conditions: the implicit condition (participants were told to remember the stimuli and pay attention to the location of some of the words), the incidental condition, the rule search condition, and the instructed condition. LAA was correlated significantly with learning outcomes in all conditions except the incidental condition, suggesting that LAA plays an important role in a learning process where more effortful, formal rule learning is involved.[1]

A similar pattern of findings has been documented in L2 research on feedback. Sheen (2007) examined how LAA contributed to the L2 acquisition of English articles in two types of corrective feedback in writing: direct feedback and direct feedback plus metalinguistic explanation. LAA was found to be significantly related to learning gains after both types of feedback, but the relationship was stronger for the feedback that included metalinguistic information. This suggests that learners with high LAA benefit more from metalinguistic feedback to improve their accuracy. Consistent with Sheen's (2007) findings, other studies have found LAA to be particularly important in a learning situation where the learners engage in form-focused learning during feedback (Trofimovich et al., 2007; Yilmaz, 2013).[2]

Overall, LAA plays a significant role in form-focused L2 grammar learning. Because the current study involves explicit form-focused grammar learning, it is conceivable that LAA is related to the learning outcomes regardless of practice distribution. However, if the distribution of practice influences the degree of engagement in explicit form-focused learning in any way, then the contribution of LAA may change in distributed or massed practice conditions.

WMC AND L2 GRAMMAR LEARNING

With recent advancements in understanding various aspects of memory in psychology, the conceptualization of aptitudes has been extended to include WMC as well as other components (Linck, Hughes, et al., 2013; Miyake & Friedman, 1998; Skehan, 1998, 2002). All models of working memory (WM) see it as a limited-capacity system, but the models differ in how they define WMC or conceptualize its limitations (Baddeley, 2012; Conway, Jarrold, Kane, Miyake, & Towse, 2007; Engle, 2002). In the present paper, WM is operationalized as a control mechanism that regulates the operation of various cognitive processes such as shifting, updating, and inhibiting (Miyake & Friedman, 2012; Miyake et al., 2000). It stores and manipulates immediate information until this information is integrated into the cognitive process (Linck, Osthus, Koeth, & Bunting, 2013). Following this operationalization, WMC is typically assessed by complex tasks such as the operation-span (Ospan) task in which participants are asked to execute two tasks at the same time: recalling a list of alphabet letters (storage) while solving math problems (processing). In what follows, we will review research that examined the role of WMC in L2 learning (see Linck, Osthus, et al., 2013; Williams, 2012, for review). In these studies, WMC was assessed by a variant of complex tasks (e.g., Ospan, reading-span, and listening-span tasks), all of which taxed both storage and manipulation components.

Accumulating evidence suggests that the effect of WMC is associated with different types of instruction (Brooks et al., 2006; Goo, 2012; Kempe & Brooks, 2008; Mackey, Adams, Stafford, & Winke, 2010; Sagarra, 2007; Sanz, Lin, Lado, Stafford, & Bowden, 2014; Trofimovich et al., 2007; Yilmaz, 2013). Brooks et al. (2006) investigated the role of WMC in learning noun gender marking with a miniature version of Russian. They manipulated the type variability (i.e., the amount of vocabulary in inflection categories to be presented). Participants were randomly assigned to three groups: high variability (24 nouns), medium variability (12 nouns), and low variability (6 nouns). The results showed that learning outcomes were superior in the high-variability input condition *only* among high WMC participants. WMC seems to mediate the acquisition of morphology through the large set of items.

A more recent study investigated how WMC moderated the effects of grammar instruction prior to practice (Sanz et al., 2014). Participants were trained, through input-based practice and explicit feedback, on a morphological system in miniature Latin. They were assigned to either a group with a grammar lesson prior to practice or a group without any grammar lesson. The results showed that WMC played a role in practice without grammar instruction, but it did not moderate the training outcomes when practice was accompanied by grammar instruction. They concluded that providing prepractice grammar explanation reduced the effect of individual differences in WMC. In other words, WMC was related to the outcomes from learning without grammar explanation because practice without an advanced organizer (e.g., grammar explanation) made learning more demanding, taxing WM more.[3] This pattern seems to fit one of the general findings in ATI research in education that "aptitude becomes more important as the treatment puts more of a burden of information processing on the learners" (DeKeyser, 2013, p. 29). This suggests that the role of WMC will become more important if either distributed or massed learning conditions place more processing demands on the learning process.

If learning demands are different between distributed and massed practice in the current study, it may be possible to predict the relative contribution of WMC in the learning process. Among several theories that have been proposed to account for the underlying learning mechanisms in spacing effects, two theories appear to predict different learning demands placed on distributed practice and massed practice. Although the current study did not directly attempt to test the validity of these theories (for reviews on the theories, see Serrano, 2012; Toppino & Gerbier, 2014), these two accounts are briefly explained to motivate two competing hypotheses regarding how WMC can possibly moderate either distributed or massed practice: (a) the study-phase retrieval account and (b) the discriminative-contrast account.

In the study-phase retrieval account, the successful retrieval of the earlier-learned material at a later time plays a crucial role in better retention (e.g., Toppino

& Bloom, 2002). As long as people can successfully retrieve the previous item (Karpicke & Roediger, 2007, 2008), greater spacing leads to better retention. Conversely, a longer interval increases the probability of failure in retrieving the previous item. This suggests that learners with low WMC may be at risk of completely forgetting the previously learned items and failing to benefit from distributed practice. If this is a scenario, it is likely that distributed practice (i.e., a longer interval between training sessions) may be more sensitive to WMC than massed practice.

In contrast, the discriminative-contrast account offers a different scenario (e.g., Kornell & Bjork, 2008). In a nutshell, it claims that spacing the presentation of learning items helps learners to discriminate the critical features of the items. When time intervals are interspersed between the two study phases, it is more advantageous in discriminating between categories. In other words, massed practice may make it difficult to distinguish the elements, which places more demand on working memory. Because individuals with low WMC suffer from interference from prior similar items more than those with high WMC (Kane & Engle, 2000), WMC may be more crucial when items are presented en masse in a shorter period of time.

APTITUDE COMPLEXES AND L2 GRAMMAR LEARNING

The two aptitude components in the current study have often been examined separately in previous research, and only a few studies have examined the role of both LAA and WMC in the same study design. The idea of using a combination of aptitudes or an "aptitude complex" for different types of instruction has been explored in other fields such as instructional psychology (Cronbach & Snow, 1977). Aptitude complexes are a set of cognitive aptitudes that are assumed to be recruited together for different learning processes (Ackerman, 2003; Snow, 1987).

In SLA research, Robinson (2007) proposed an aptitude complex for explicit rule learning: metalinguistic rule rehearsal and memory for contingent text. These two "ability factors" are further broken down to "cognitive abilities." Metalinguistic rule rehearsal consists of grammatical sensitivity and rote memory, whereas memory for contingent text (the ability to remember and rehearse written information) is a combination of WM for text and speed of WM. According to Robinson, these cognitive abilities are relevant to learning processes in which learners are given a rule explanation with written examples, remember and rehearse the rule, and apply it for comprehension or production exercises.

Following Robinson's (2007) framework as a guide for exploring aptitude complexes for distributed/massed practice, the present study examines the two cognitive abilities, LAA and WMC, simultaneously. Because the learning processes in the current study involve an explicit mode of grammar learning (e.g., learning about a rule and applying it for comprehension and production activities), examining both LAA and WMC is of great relevance and importance for the study. The current study does not measure the same four cognitive abilities as proposed by Robinson above; our approach is exploratory and focuses on LAA (which subsumes grammatical sensitivity) and WMC (which is relevant for remembering and rehearsing grammatical rules for production).

Two empirical studies will be presented that investigated the role of LAA and WMC in L2 feedback treatments.[4] Both studies involved LAA and WMC as covariates as well as random assignment of participants to two feedback treatment groups. Their studies yielded different patterns of ATI, and they suggest potential ATI patterns that the present study may find by exploring LAA and WMC. Yilmaz (2013) investigated how LAA and WMC interacted with two feedback types in L2 grammar learning. Learners were randomly assigned to either an explicit-correction group or a recast group. The results showed that both LAA and WMC were only related to learning gains in the explicit-correction group and that learning gains in the recast group were not related to the aptitudes. Similarly, Li (2013) investigated how the two aptitude components influenced L2 grammar learning under two types of feedback (metalinguistic feedback and recast). A different pattern of the results from Yilmaz (2013) emerged; LAA was the only predictor of the performance after recast, whereas WMC was only related to the performance after metalinguistic feedback.

Findings from Li (2013) and Yilmaz (2013) suggest two different patterns of ATI. In Li's study, high-LAA learners benefit most from one treatment (i.e., recast), while high-WMC learners benefit most from the other treatment (i.e., metalinguistic feedback). In Yilmaz' study, one type of treatment (i.e., explicit correction) is associated with both aptitude components; the other type of treatment (i.e., recast) was not sensitive to either aptitude. The purpose of presenting two studies is not to discuss potential explanations for the divergent findings; rather, we argue that conceptualizing two aptitudes, LAA and WMC, as a set may be useful for distributed/massed L2 grammar learning.

In sum, the present study attempts to advance understanding of an underresearched problem (how an aptitude complex moderates the effects of practice distribution) because retention of skills over a certain time interval is presumably susceptible to individual differences in both LAA and WMC.

THE PRESENT STUDY

The current study aimed to investigate whether the effectiveness of distributed/massed practice[5] is moderated by individual differences in LAA and WMC. Forty beginner-level learners of Japanese as an L2 were trained on an element of Japanese morphosyntax (the present progressive form, *–te imasu*), accompanied by vocabulary learning, under either massed or distributed practice conditions. Their LAA was measured with the Llama_F (Meara, 2005), and their WMC was assessed with the automated Ospan task (Unsworth, Heitz, Schrock, & Engle, 2005). The massed practice group performed the same set of training tasks twice at a 1-day interval, whereas the distributed practice group repeated the task set at a 7-day interval.

The current study addressed the following primary research question: do LAA and WMC moderate the learning outcomes in the same way when L2 practice is distributed or massed? Given the novelty in investigating how individual differences mediate the effect of various levels of distribution of practice on L2 learning, the question was rather exploratory, without any commitment to specific predictions, but possible scenarios are delineated below as to how the role of LAA and WMC may play out with distributed and massed practice.

Because LAA is generally associated with explicit grammar learning (Skehan, 2012), the question was left open as to whether LAA would play a different role under two different distributions of practice. It was conceivable that LAA would facilitate L2 grammar learning regardless of practice distribution. Nevertheless, it was also possible that LAA might be more crucial in one of the learning conditions, if the different learning intervals changed the underlying L2 learning processes. Any of these patterns of results would offer an important insight into the learning processes that were facilitated or inhibited by the individual differences in LAA (DeKeyser, 2012; Robinson, 2002).

Based on the two accounts for spacing effects, two opposite predictions were put forth on the role of WMC in distributed/massed practice. The study-phase retrieval account (e.g., Toppino & Bloom, 2002) predicted that distributed practice was problematic for low-WMC learners because the longer interval between the training sessions may tax WMC more than the shorter interval. In contrast, the discriminative-contrast account (e.g., Kornell & Bjork, 2008) predicted that massed practice would draw more on WMC because it requires discrimination among the learning items and may place more demand on WMC.

METHOD

Participants

Forty beginner-level Japanese L2 learners participated in the study (25 females, 15 males). Their mean age was 21 ($SD = 2.89$). The first language of the participants was English, except for two individuals (their first languages were Nepali and Romanian), who were included in the study because they were highly proficient, using English for their undergraduate study. The first group of participants recruited were enrolled in third-semester Japanese courses during the study ($n = 29$), and another group of participants had taken Japanese courses for two semesters before but were not taking any Japanese courses at the time of study ($n = 11$). The analyses were conducted separately on the first group only and the two groups combined; similar patterns of results were obtained across the two analyses. Given that the correlational analyses between the aptitude tests and the outcome measures are more stable with a larger sample size, the results presented will be from the analyses conducted on the whole group ($n = 40$).

Target structure

The present study targeted a morphosyntactic structure in Japanese, the *–te* form of the verb, which is used for expressing the present progressive, as in *–te imasu* (e.g., *Taro wa ki o nobotte imasu*; Taro-subject tree-object is climbing). The target structure was introduced in the second semester of Japanese, and all the participants had learned about the *–te* form before the study, but most of them had not mastered it completely. The focus of the study is on six categories of regular verbs, which involve allomorphic stem changes (Table 1). Three verbs from each **Table 1**

Table 1. *Conjugation of the te-form*

Stem	Transformation Rule	Uninflected Form	*Te*-Form
–*r*	Q	*nobor-u* (to climb)	*nobot-te*
–vowel	Q	*hiro-u* (to pick up)	*hirot-te*
–*m*	n	*tatam-u* (to fold)	*tatan-de*
–*b*	n	*musub-u* (to tie)	*musun-de*
–*k*	i	*migak-u*(to polish)	*migai-te*
–*g*	i	*sosog-u* (to pour)	*sosoi-de*

<table>
<tr><td></td><td>Time 1</td><td></td><td>Time 2</td><td></td><td>Time 3</td><td></td><td>Time 4</td></tr>
<tr><td>Massed</td><td rowspan="2">Pre-test
+
Training
Session 1</td><td>→
1 day</td><td rowspan="2">Training
Session 2
+
Post-test</td><td rowspan="2">→
7 days</td><td rowspan="2">Post-test
+
Ospan
task</td><td rowspan="2">→
21 days</td><td rowspan="2">Post-test
+
LLAMA F</td></tr>
<tr><td>Distributed</td><td>→
7 day</td></tr>
</table>

Figure 1. Research design.

of six categories were used for the training session (see Appendix A for all 18 verbs). All the verbs were action verbs and were unknown to participants as shown by the pretest scores.

The uninflected form of these verbs is converted to the –*te* form with either the /te/ or the /de/ allomorph. When the stem ending of a consonant verb is /r/ or /w/, it turns into /Q/ or reduplication of the initial consonant in the –*te* form (e.g., *nobotte* for *noboru* or *hirotte* for *hirou*); when the ending is /m/ or /b/, it turns into /n/ (e.g., *tatande* for *tatamu* or *musunde* for *musubu*); and when it is /k/ or /g/, it turns into /i/ (e.g., *migaite* for *migaku* or *sosoide* for *sosogu*; Vance, 1987).

Research design

The current study involved a between-subject factor (practice distribution) with two covariates, LAA and WMC, as within-subject factors. As shown in Figure 1, **Figure 1** there were four individual sessions for each participant. The first author met the participants in a quiet laboratory for the pretests and training session 1 (Time 1), training session 2 followed by posttests (Time 2), posttests and the Ospan task (Time 3), and posttests and the Llama_F test (Time 4). Participants were randomly assigned to a massed practice group (massed, $n = 18$) or a distributed practice group (distributed, $n = 22$). The pre- and posttests were administered from Time 1 to Time 4 to measure knowledge of the target grammatical structure (see Pre- and Posttests below).

The ISIs (1- vs. 7-day) and RIs (7- vs. 28-day) were determined based on Rohrer and Pashler's (2007) optimal ratio of ISI and RI. Rohrer and Pashler (2007) found **Q1** that the optimal timing for relearning (i.e., ISI) depends on how far removed delayed testing is from the end of practice (i.e., RI). They suggested that the optimal ISI is approximately 10% to 30% of the RI. The current study set the ISIs and RIs such that the ISI–RI ratios fell within the optimal range of 10% to 30% only in one of the groups at Times 3 and 4. Specifically, an ISI–RI ratio within the optimal range was used for the massed practice group at Time 3 (14%), and for the distributed practice group at Time 4 (25%; see Suzuki & DeKeyser, 2015, for details). However, the focus of the study was not on the effectiveness of different ISIs on the outcomes at different RIs. Rather, the primary aim was to examine the role of the aptitudes on the two learning conditions that were different in terms of the ISIs, and by extension to what extent the effect of aptitudes on outcome performance, if observed, would persist in the delayed posttests (i.e., Times 3 and 4).

Training sessions

Each training session consisted of computerized tasks in which learners practiced the use of the present progressive. Participants completed the three-stage practice in an explicit step-by-step manner: vocabulary learning, explicit grammatical explanations, and comprehension and production practice. The set of training tasks took about 45–50 min. All the tasks below were conducted in the same way in Training Sessions 1 and 2.

Vocabulary learning. Eighteen verbs (with object nouns) were learned on the computer screen (see Appendix A). The participants were presented with a picture that represented action verbs and were asked to say the uninflected form of a verb phrase within 5 s (e.g., *batto o huru*; swing the bat). After these 5 s, they were presented with the Japanese phrase both aurally and visually, in blue letters, along with the written equivalent of the English translation in black. The vocabulary remained on the screen for 5 s, and the next picture appeared automatically. They repeated the set seven times.

Grammatical explanation. After the vocabulary training, participants were presented with a sheet of paper that contained the grammatical explanations about the target forms and the conjugation chart for all six categories (see Appendix B). This sheet was available for reference during the entire practice session. In order to make sure that participants knew the correct conjugation forms, they were given a worksheet in which they transformed the uninflected form of the 18 verbs (accompanied by the pictures from the vocabulary learning task) into the present progressive *–te* form.

Comprehension and production practice. After the explicit explanation, participants completed the comprehension and production practice by using present progressive sentences one on one with the experimenter. The cards, which had the same pictures as the ones used during the vocabulary training session, were laid out on the table. The experimenter read aloud the sentence that described the action in present progressive form in one of the pictures (e.g., *batto o hutte imasu; someone is swinging the bat*), and the participant's task was to pick up the corresponding card as soon as possible. Because Japanese is a pro-drop language, the sentence did not mention the subject of the sentence, which is natural. This task was repeated twice. For the practice of producing the sentences, the roles were reversed from those in the comprehension practice. Participants were asked to describe the picture to the experimenter on their own, so that he could pick up the picture that participants described. When participants could not describe the picture, the experimenter described the card for them. This task was also repeated twice. Feedback in the form of recasting was given if participants produced an incorrect form of the verb.[6] Finally, participants performed a narrative task, describing what a person in a video was doing with the present progressive form. The participants were told to describe the action using the *–te* form while the silent video clip was played for 10 s. There were 18 video clips describing all the practiced action verbs (no vocabulary list was given). After each video clip, the correct sentence was presented both aurally and visually on the screen for 4 s. As in the case of the two previous tasks, the video narrative task was performed twice to ensure that participants received enough practice on the sentences.

Pre- and posttests

Two types of oral tasks were employed as the pretest and posttests at Time 1 through Time 4: the rule application test, which measured the rule of the *–te* form conjugations by using nonce verbs, and the picture sentence completion test, which measured the vocabulary and the *–te* form conjugations. The rule application test was always administered before the picture sentence completion task because the first test assesses narrower knowledge (rules) than the second (rule + vocabulary). No feedback was given throughout the testing phase. The tests were computerized and administered with the DMDX software (Forster & Forster, 2003); the responses were audiorecorded.

At Time 1, the rule application test was administered before the training session to measure the preexisting knowledge of the *–te* form rules, and the sentence completion test was also conducted to check that participants did not know any of the verb phrases that would be practiced in the training. None of them correctly described any pictures or knew the verb phrases to be practiced (see Results). This ensured that no participants had seen the *–te* form of those verbs prior to the study. The posttests were administered at three later time points in order to assess the durability of the treatment effects.

The outcome tests assessed two aspects of *–te* form use: accuracy and speed. Accuracy indicates to what extent participants can use the *–te* form appropriately in oral production. Speed was measured by the time it took to utter the target construction from the prompt to the end of the utterance.

Rule application test. The purpose of the rule application test was to assess to what extent the participants could use the *–te* form rules correctly. Eighteen nonce verbs were created based on the practiced verbs, keeping the same initial phoneme and number of moras; different nonce verbs were used in each test to avoid practice effects (e.g., the practiced verb, *yaburu*, was converted to nonce verbs such as *yomaru*). Participants were required to convert the sentence with an uninflected verb into one with present progressive by using a nonce verb. After a fixation cross appeared in the center of the screen, an uninflected form of a nonce verb (e.g., *yomaru*) was presented both in the written and oral modality. Immediately after the end of the word, participants were asked to convert the word into the present progressive form as quickly as possible (e.g., *yomatte imasu*). They were given 10 s to complete their response. Participants performed a practice session with five unknown pseudoverbs in order to become (re)familiarized with the format of the test every time they took it. It took approximately 3 min to complete the test.

Picture sentence completion test. Whereas the rule application test only targeted the rules for the *–te* form, the picture sentence completion test assessed to what extent the participants could use the correct *–te* form of the verbs that they practiced. In the test, participants were presented with a picture in which someone was performing an action, immediately followed by an auditory stimulus, which was the subject of a sentence to be completed (i.e., *otokonohito-ga* [man], *onnanohito-ga* [woman], *otokonoko-ga* [boy], or *onnanoko-ga* [girl]). Their task was to complete the sentence by describing what the person was doing (e.g., *ki o nobotte imasu*; [he/she] is climbing the tree). They were given a maximum of 15 s to complete their responses. The 18 verb phrases used were the ones that participants practiced during the training. Four practice items with two basic verbs (i.e., eat a hamburger and watch TV) preceded the actual 18 items. At the pretest, participants were shown pictures to check what each picture meant before they took the picture sentence completion test. It took approximately 5 min to complete the test.

Individual difference measures

LAA was measured by the Llama_F, which is one of subtests of the LLAMA aptitude test (Meara, 2005); WMC was measured by the automated Ospan task from Unsworth, Heitz, Schrock, and Engle (2005).

Llama_F. In the Llama_F, participants were required to induce rules of grammar by looking at pictures and word sequences that described them; in this way, Llama_F can measure LAA independently from participants' first language (for descriptions of the grammar rules in Llama_F, see Jackson, 2014). In the learning phase, participants were given 5 min to learn a new language by seeing sentences matched with pictures. In the testing phase, the program displayed a picture and two sentences, one grammatical and the other ungrammatical. Their task was to choose the grammatical sentence. The test consisted of 20 items.

Ospan tasks. In the Ospan task, for each item, participants solved a math problem, indicating whether the solution for an equation was correct or incorrect. After each math problem, they were presented with a letter of the alphabet and asked to remember it. After each set of math problems and letters, they were asked to select the letters in the presented order. Successful performance on this task requires temporary updating of incoming information consecutively. There were 15 trials in total: 3 trials for 5 sets with five different sizes each (3–7). The total number of sets was 75.

Scoring and analysis

For all the pre- and posttests, five trained independent raters, native speakers of Japanese, scored and analyzed the speech data, using the sound analysis software Praat. The raters were trained until their coding matched those of the first author, using 15% of data. The two aptitude tests were automatically scored by the software to administer them. See below for reliability data for each test.

Rule application test. Accuracy of the utterances was coded by listening to each item. Minor pronunciation errors were ignored in the scoring, and the accuracy of utterances with repair was determined based on the last utterance. The reliability of the accuracy scores in the rule application test across time was calculated with Cronbach α; and the indices were 0.941 at Time 1, 0.890 at Time 2, 0.892 at Time 3, and 0.916 at Time 4.

The participants' utterances were also analyzed for speed or response time (RT), measured from the onset of a word to the end of the utterance. In order to exclude responses resulting from different processes than normal responses, three data cleaning procedures were conducted before computing RT for each participant. First, the RTs of responses that contained incorrect utterances were excluded from the analysis. Second, the RT was not calculated for responses with repairs, rephrasing, and/or false starts, because it was impossible to determine whether they were due to a lack of linguistic knowledge or other random sources (e.g., slip of the tongue). Third, we defined outliers as RTs below 500 ms and RTs higher than 3 *SD* above the grand mean for each participant. These cutoff values were determined after inspection of the data (for a similar approach, see De Jong, Steinel. Florijn, Schoonen, & Hulstijn, 2013). The number of outliers identified ranged from 3.7% to 8.7% of the data sets across the tests. In total, the percentages of valid temporal measures retained were 29.7%, 83.2%, 81%, and 80% for Time 1, Time 2, Time 3, and Time 4, respectively. Given the small percentage of valid responses for Time 1, temporal measures of the rule application test at Time 1 were not included in the subsequent analyses. Note that four participants in the distributed practice group were excluded from RT analyses only, due to the low number of valid (correct) responses.

Picture sentence completion test. As in the rule application test, responses in the picture sentence completion test were analyzed in terms of accuracy and speed. Accuracy was scored based on the accuracy of the *–te* form, and minor pronunciation mistakes in vocabulary were ignored (e.g., *sukutte imasu* for *tsukutte imasu*). The reliability of the accuracy scores was calculated with Cronbach α; and the indices were 0.801 at Time 2, 0.862 at Time 3, and 0.875 at Time 4.

The RT was calculated from the onset of a picture to the end of the utterance. The percentage of outliers ranged between 2.3% and 6.2% of the data sets across the tests. The percentages of valid RT data retained were 80%, 62.8%, and 57.4% for Time 2, Time 3, and Time 4. Because the accuracy was 0 for Time 1, no valid RT could be measured at Time 1.

Table 2. *Accuracy and speed on the rule application test by group from Times 1 to 4*

	Accuracy				Speed			
	Massed ($N = 18$)		Distributed ($N = 22$)		Massed ($N = 18$)		Distributed ($N = 18$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Time 1	5.06	5.42	5.91	5.55	—	—	—	—
Time 2	16.67	1.75	14.59	4.50	3635	777	3704	817
Time 3	15.89	3.03	14.05	4.73	3655	886	3604	593
Time 4	15.00	4.03	14.36	4.95	3725	1115	3438	560

Llama_F. Because the Llama_F program does not automatically record the individual item responses, the beep sounds indicting the correct/incorrect responses were audiorecorded using the freeware software Audacity (http://audacity.sourceforge.net). Each test item was then scored correct or incorrect. One item was excluded due to its negative item-total correlation ($r = -.355$); the possible maximum score was therefore 19. The internal consistency for Llama_F estimated with Cronbach α was slightly below the acceptable range ($\alpha = 0.53$) in the present study. This somewhat low reliability estimate for the Llama_F scores might attenuate results of the correlation analyses.

Ospan task. The Ospan task was scored as the sum of all correctly recalled letters in correct positions. In other words, no credit was given unless the set of the letters in a trial was recalled in the right order. If an individual correctly recalled three letters in a set size of five, for example, the score was zero. In order to make sure that the Ospan task was performed appropriately, only participants with high accuracy rates in the math problems are usually recommended for inclusion. The average accuracy rate was 94.29% (*SD* = 3.93%, range = 83%–100%). In Unsworth et al. (2005), an 85% accuracy criterion (i.e., a maximum of 12 errors out of the 75 operations) was set for all participants. Two participants scored slightly below the criterion (84% and 83%), but they were retained in the subsequent analyses because the accuracy was still high and it could avoid losing valid data points from the relatively small sample size in the study. Reliability indexed by Cronbach α was satisfactory ($\alpha = 0.79$).

RESULTS

Descriptive statistics

Pre- and posttests. Because the focus of the current study was not on the differences between the outcomes of the two groups, the results for the two outcome tests are compared only briefly between the two groups (see Suzuki & DeKeyser, 2015, for details). Table 2 presents the accuracy and speed scores on the rule application test. For accuracy, the massed practice group seems to outperform the distributed **Table 2**

Table 3. *Accuracy and speed on the sentence completion test by group from Times 1 to 4*

	Accuracy				Speed			
	Massed ($N = 18$)		Distributed ($N = 22$)		Massed ($N = 18$)		Distributed ($N = 18$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Time 1	0.00	0.00	0.00	0.00	—	—	—	—
Time 2	16.89	1.78	15.23	3.02	4027	711	4441	720
Time 3	14.17	3.19	12.18	4.72	4865	994	4924	1002
Time 4	11.78	4.45	11.82	4.73	4748	940	5360	1182

practice at the descriptive level, but there were no large differences on either accuracy or speed between the two groups across time. For speed, no systematic differences were found at any time. Of interest to the primary correlational analyses in the study, the standard deviations of accuracy scores were slightly smaller in the massed practice group at Time 2, but there were some variations within each group. The RT variations were also observed on the speed measures in both groups. Note that the two groups did not differ on the mean scores for the rule application test at Time 1, t (38) = –0.49, p = .63, which partially accounts for the equal proficiency between the groups. **Table 3**

Table 3 presents the accuracy and speed scores on the sentence completion test. In general, advantages were found for the massed practice group, at least descriptively except for the accuracy scores at Time 4; the only significant difference was found on the speed measure at Time 4. As in the rule application test, the standard deviation for accuracy was smaller for the massed practice group at Time 1. There seems, however, to be enough variation in the outcome measures across the two groups to allow us to examine the correlations with the aptitude components.

Llama_F and Ospan task. The mean score for Llama_F was 16.20 (SD = 2.23, range: 10–19). The mean Ospan score was 50.78 (SD = 16.03, range = 10–75). The Pearson product-moment correlation coefficient between the two was negligible for the massed group (r = .107, n = 18, p = .674), for the distributed group (r = –.021, n = 22, p = .925), and for both groups combined (r = –.011, n = 40, p = .947). These results suggest that the LAA and WMC components were relatively independent, and no difference in their relationship was found between the groups.

Relationship of outcome test scores with LAA and WMC

The primary research question is to what extent the individual differences in LAA and WMC moderate the learning gains in the two groups. In order to investigate the role of LAA and WMC, a series of Pearson r correlation coefficients were computed between the gain scores from the pretest to the posttests and the scores on the Llama_F and Ospan task across groups. We considered using analyses of

Table 4. *Pearson correlations among scores on the rule application test, Llama_F, and Ospan*

	Ospan		Llama_F	
	Massed	Distributed	Massed	Distributed
Accuracy				
Time 2	–.199 (.430)	–.089 (.692)	.438 (.069)	.358 (.102)
Time 3	–.362 (.140)	–.208 (.352)	.234 (.350)	.169 (.451)
Time 4	–.216 (.388)	–.340 (.122)	.281 (.259)	.104 (.646)
Speed				
Time 2	–.047 (.854)	.343 (.138)	.147 (.560)	–.250 (.289)
Time 3	–.108 (.669)	–.152 (.523)	.186 (.459)	–.195 (.409)
Time 4	–.023 (.929)	–.061 (.799)	.202 (.421)	–.426 (.061)

Note: Accuracy scores in the rule application tests were gain scores from the pretest at Time 1. The values in parentheses are two-tailed p values.

covariance with LAA and WMC as the covariates, but correlation analyses were chosen due to the small number of participants in the study (see, e.g., Shintani & Ellis, 2015, for a similar approach). In the rule application tests, the gain scores for accuracy were calculated by subtracting the score at Time 1 from the subsequent posttests (i.e., Times 2–4); for the speed measures, the RTs at Time 2 through Time 4 were used as gain scores, because no valid RTs were computed at Time 1. Because no participants knew the vocabulary tested on the sentence completion task at Time 1, the gain scores were equivalent to the scores on those tests at Times 2, 3, and 4.

A total of 48 coefficients were computed among LAA, WMC, and learning outcomes under the distributed and massed learning conditions. This usually requires the Bonferroni correction for p values in order to lower the chances of obtaining false-positive results (Type I errors); however, no correction procedure was conducted due to the exploratory nature of the study (Bender & Lange, 2001). The significance values for p ($<$.05) should be interpreted only as preliminary findings. In other words, the main purpose was not to establish definitive conclusions but to seek potential correlational patterns to help guide the directions of future research. Based on the guidelines for L2 research[7] proposed by Plonsky and Oswald (2014), the coefficient r was interpreted as small-weak ($\approx$.25), medium-moderate ($\approx$.40), or large-strong ($\approx$.60). To supplement the interpretations of r and p values, confidence intervals for r were also reported for correlation coefficients with substantial effects (small-large effect sizes).

The results of correlational analyses for the rule application test are presented in Table 4 (see Appendices C and D for scatterplots). None of the 24 correlation **Table 4** coefficients produced a large effect, and no consistent correlational patterns were observed across time or condition. In terms of accuracy, moderate positive correlations were only found with Llama_F, for both of the groups, only at Time 2 (r = .438 and .358, p = .069 and .102, 95% confidence interval [CI] = –0.176, 0.787 and 0.013, 0.688). The correlations of the Ospan scores with the accuracy

Table 5. *Pearson correlations among scores on the sentence completion test, Llama_F, and Ospan*

	Ospan		Llama_F	
	Massed	Distributed	Massed	Distributed
Accuracy				
Time 2	.416 (.086)	–.095 (.674)	.072 (.777)	.658 (.001)
Time 3	.555 (.017)	.072 (.752)	–.065 (.798)	.719 (.000)
Time 4	.408 (.093)	.149 (.509)	.207 (.410)	.693 (.000)
Speed				
Time 2	–.606 (.008)	.116 (.607)	–.039 (.876)	–.688 (.000)
Time 3	–.473 (.047)	.215 (.376)	.000 (.999)	–.296 (.218)
Time 4	–.345 (.160)	.237 (.328)	.235 (.349)	–.437 (.061)

Note: The values in parentheses are two-tailed *p* values.

scores were not significant but negative overall; and the coefficients were small to medium for the massed and distributed practice conditions, respectively, at Time 3 (r = –.362 and –.208, p = .140 and .352, 95% CI = –0.702, 0.132 and –0.519, 0.173) and at Time 4 (r = –.216 and –.340, p = .388 and 122, 95% CI = –0.636, 0.340 and –0.597, –0.077). For the speed measures, a moderate negative relationship was found between the Llama_F score and the speed score at Time 4 only (r = –.426, p =.061, 95% CI = –0.705, 0.024). This means that improvement in the efficiency to apply the *–te* form rules from Time 1 to Time 4 (i.e., faster RT at Time 4 than at Time 1) was associated with LAA. In sum, no clear and consistent correlational patterns were found for the rule application test.

For the sentence completion test, intriguing patterns emerged between the two groups as shown in Table 5. The Ospan scores were consistently related to the learning gains among the massed practice group only: depending on the testing time, the strength of the relationships were weak to strong in absolute values among the massed group (.345 $< r <$ .606), whereas none of the correlation coefficients in the distributed practice group were significant or corresponded to more than small effect sizes (.072 $< r <$ .237, $p >$ .05). For the accuracy measures, in particular, there was a moderate to strong, significant positive correlation between the Ospan scores and the learning gains at Time 3 (r = .555, p = .017, 95% CI = 0.119, 0.850), and somewhat smaller, moderate nonsignificant coefficients were observed at Time 2 (r = .416, p = .086, 95% CI = 0.042, 0.775) and at Time 4 (r = .408, p = .093, 95% CI = –0.008, 0.734). For the speed measures, the coefficients with the Ospan scores became gradually smaller at later time points; medium to strong negative correlations were found at Times 2 and 3 (r = –.606 and –.473, p = .008 and .047, 95% CI = –0.820, –0.327 and –0.762, –0.146), followed by a weaker relationship at Time 4 (r = –.345, p = .160, 95% CI = –0.630, –0.015). In sum, only half of the six tests indicated significant correlations between the Ospan scores and the outcome measures in the massed practice group, but the overall pattern is consistent. In addition, the CIs for the correlation coefficients were wide; therefore, no definitive conclusions should be drawn.

In contrast, the opposite correlational patterns were observed for LAA. All the correlation coefficients between the outcome measures and the Llama_F scores were consistently higher in the distributed group than in the massed group. Depending on the testing time, the strength of associations with Llama_F in the distributed practice group was weak to strong in absolute values ($.296 < r < .719$); none of the coefficients in the massed group indicated more than a small effect size ($0 < r < .235$, $p > .05$). In particular, Llama_F scores were significantly correlated with accuracy measures in the distributed practice group consistently across three times with large effect sizes: $r = .658$, $p = .001$, 95% CI = 0.081, 0.937 at Time 2; $r = .719$, $p < .001$, 95% CI = 0.437, 0.898 at Time 3; and $r = .693$, $p < .001$, 95% CI = 0.348, 0.884 at Time 4. A similar pattern was also observed for the speed measure in the distributed practice group; the strong negative correlation was found at Time 2 ($r = –.688$, $p < .001$, 95% CI = –0.892, –0.397), and weaker but small to moderate negative correlations were observed at Time 3 ($r = –.296$, $p = .218$, 95% CI = –0.787, 0.216) and Time 4 ($r = –.437$, $p = .061$, 95% CI = –0.750, –0.147). Scatter plots for the sentence completion test are provided in Appendices E and F. In sum, four of the six tests indicated significant correlations between the Llama_F scores and the outcome measures in the distributed practice group, but the CIs were large due to the small sample size. Again, we should not draw any definitive conclusions from the present findings, but these tentative findings are discussed to inform and guide future ATI research.

DISCUSSION

The present exploratory study attempted to investigate how individual differences in LAA and WMC moderate the effectiveness of distributed and massed practice. When outcomes were measured by the rule application test (rules only), no consistent correlations were observed between the aptitudes and the learning gains. An intriguing asymmetrical pattern, however, emerged for the sentence completion test (lexicon + rules): WMC was more related to the effectiveness of massed practice (i.e., 1-day ISI), whereas LAA exclusively contributed to the effectiveness of distributed practice (i.e., 7-day ISI).

No consistent correlations were observed between the aptitudes and outcomes for the rule application test probably because learning "only" *–te* form rules might place less demand on learning than learning to integrate new vocabulary with the *–te* form rules (the sentence completion test). Because the participants had acquired some of the *–te* form rules (see the pretest scores on the rule application test at Time 1), learning and reviewing rules may have been less taxing for learners. It is conceivable that learning both new verbs and the corresponding *–te* form rules placed more demands on the aptitudes. This is consistent with the previous research on the ATI in which aptitudes play a more important role when the learning burden is higher (Brooks et al., 2006; DeKeyser, 2013; Sanz et al., 2014). Note that learners tended to have more difficulty in retrieving vocabulary rather than in applying the *–te* form rules based on the high accuracy scores on the rule-application test at Times 2, 3, and 4 (16.67, 15.89, and 15.00 out of 18, respectively). Because the picture sentence completion test not only requires using the *–te* form rules but also lexical retrieval, the role of lexical learning should be acknowledged as well as grammar learning.

Although no moderate or strong correlations were detected for the rule-application test, a weak but somewhat consistent overall negative pattern appears to exist between the Ospan scores and the accuracy scores at Times 3 and 4. Because these relationships were not significant and weak ($-.208 < r < -.362$, $p > .05$), this finding should be interpreted with caution. The negative relations may suggest that learners with high WMC engaged more in memorizing conjugated verb forms, which might have made it more difficult for them to induce the rules and apply them to the nonce verbs. The following discussion will focus on the findings on the sentence completion test, which assessed both lexical and grammatical knowledge.

WMC is associated with learning processes under massed practice

With regard to the role of WMC on the distributed/massed practice, two competing predictions were put forth. The findings in regard to these two predictions should be interpreted with two important notes. First, because the current study is exploratory in nature, the predictions were not intended for testing the theories. Instead, they facilitate interpretations and help establish directions for future research. Second, the current study did not make differential predictions for the two types of outcome measures, accuracy and RTs, because little is known how two measures relate to the two accounts.

The study-phase retrieval account suggested that WMC would be correlated with the performance after the distributed practice in which learners needed to retain the information for a longer period of time. The discriminative-contrast account, in contrast, predicted that WMC would be correlated with the outcomes in the massed practice group because higher WMC would be needed in order to prevent interference from similar learning materials (i.e., inflected verbs; Kane & Engle, 2000). The results in the present study supported the latter prediction: WMC was *exclusively* related to the outcomes in the massed practice group.

The findings suggest that L2 learners with lower WMC probably suffered from interference from similar vocabulary and grammar rules that are presented in a short period of time. Interference, rather than activation decay over time, may be responsible for the poor performance by learners with low WMC in the massed practice group. In particular, the updating function of WM, tapped by the Ospan task, may be responsible for this effect (Miyake & Friedman, 2012; Miyake et al., 2000). The updating function "requires ignoring irrelevant incoming information and also suppressing no longer relevant information" (Miyake et al., 2000, p. 89). This updating or inhibition ability might have played a crucial role in discriminating similar vocabulary and grammar rules in massed practice.

Conversely, distributed practice may involve a similar process of updating memory during the interval between the training sessions. A relatively longer, 7-day interval reduced the interference among similar morphological markers, and the effectiveness of distributed practice became less sensitive to individual differences in the updating function. The time interval for updating measured in the Ospan task is much shorter (i.e., seconds) than the actual interval between the two learning sessions for morphological markings (i.e., days). The updating function should be taxed more intensively when the interval is shorter, as was the case in the Ospan task. However, the learning complexity involved in L2 grammar learning may be as demanding, or possibly more demanding, than retaining alphabet letters in the Ospan task. While the higher demands associated with learning relatively complex morphological structures involved the updating ability more heavily in the massed practice condition, distributed practice lessened the burden on learning.

The current findings did not support the predictions aligned more with the study-phase retrieval account. It may be attributed to the specific ISIs (i.e., 1 day and 7 days) employed in the current study. The interval between the training sessions in the distributed practice (i.e., 7 days) might have been short enough for L2 learners to retain the memory of the verbs and inflections. If distributed practice involved a longer interval, WMC may play a more important role in the distributed practice. Furthermore, because the current study employed the Ospan task, which targets the updating function in WMC, it might have favored the discriminative-contrast account. The results would probably be different if different memory components were assessed.

LAA is associated with learning processes under distributed practice

LAA has been found to be a predictor of L2 grammar learning, such as in the area of effectiveness of corrective feedback (Sheen, 2007; Trofimovich et al., 2007; Yilmaz, 2013) and inductive learning (Erlam, 2005). One could predict that LAA would be related to the learning outcomes regardless of the practice distribution, because both learning conditions involved the exact same step-by-step explicit grammar practice. The novel finding of the present study is that LAA played a particularly important role in distributed practice. It may be assumed that learners with higher LAA were able to understand the rules of the *-te* form better; therefore, a deeper understanding of language structure might have allowed learners to benefit more from the spacing effects. Better understanding of the rules helped them retain the rule even after a 1-week interval. Having said that, the reason why LAA did not play a systematic role in the massed practice condition is not entirely clear;[8] further research is needed to examine underlying learning processes in distributed and massed practice.

Conclusions

The current study set out to investigate the role of individual differences in the effectiveness of distributed/massed L2 grammar practice. The preliminary findings demonstrated that LAA and WMC differentially influenced the effects of distributed and massed practice. More specifically, the role of WMC was more important in massed practice, whereas LAA was related to the effectiveness of distributed practice. The current findings should be interpreted cautiously. First, because the current study focused on L2 grammar learning and had a relatively small sample size, larger scale experiments should be conducted in different domains of L2 learning. Second, no control groups were employed that correspond to 1-day ISI and 7-day ISI groups. Third, the notion of distributed/massed practice was limited. In some circumstances, for instance, learners may determine learning intervals for specific items (e.g., delaying the review of remembered items), but the current study uniformly gave the same practice materials at given intervals during the two sessions. Thus, external validity should be further examined by operationalizing distribution of practice differently (e.g., different ISIs or different number of training sessions). Fourth, the current study investigated the acquisition of Japanese morphosyntactic structure, but the training session integrated vocabulary and grammar learning. It is not possible to tease them apart. It is worth asking to what extent vocabulary and grammar learning are differently influenced by distributed/massed practice. Fifth, this study used single tests to measure the two cognitive aptitudes (Llama_F and Ospan task). Different kinds of measurements as well as ours were used in previous research, and the present findings should be further attested using different measures for aptitudes. Finally, the current findings cannot be directly translated to program-level foreign language learning (Serrano, 2011; Serrano & Muñoz, 2007) because the study involved only two 1-hr training sessions in the laboratory environment.

In the meantime, the current study adds to the growing body of ATI research in SLA; its unique contribution consists of revealing differential roles of LAA and WMC under the distributed and massed practice conditions, shedding light on different underlying L2 learning processes. Future research is needed to replicate our findings and nuance their implications.

APPENDIX A

List of verb phrases

Category	Verb	English	Verb Phrase	Translations
R	*noboru*	climb	*ki o noboru*	to climb the tree
	yaburu	tear apart	*kami o yaburu*	to tear the paper
	huru	swing	*batto o furu*	to swing the bat
U	*hirou*	pick up	*gomi o hirou*	to pick up trash
	nuu	sew	*nuno o nuu*	to sew a piece of cloth
	sukuu	scoop	*tsuchi o sukuu*	to scoop soil
M	*tatamu*	fold	*fuku o tatamu*	to fold clothes
	momu	massage	*kata o momu*	to massage shoulder
	tsutsumu	wrap	*kyandi o tsutsumu*	to wrap candy
B	*hakobu*	carry	*pasokon o hakobu*	to carry the laptop
	tobu	jump	*roopu o tobu*	to jump rope
	musubu	tie	*himo o musubu*	to tie the string
K	*muku*	peel	*banana o muku*	to peel the banana
	migaku	polish	*kutsu o migaku*	to polish the shoes
	kudaku	smash	*kukki o kudaku*	to break the cookie into pieces
G	*sosogu*	pour	*mizu o sosogu*	to pour the water
	togu	sharpen	*houchou o togu*	to sharpen the knife
	aogu	fan	*sticchi o aogu*	to fan Stitch

APPENDIX B

Explicit information

How do you say 'I am reading' in Japanese?

yonde*(te-form of *yomu*) + *imasu

The -te form of a verb or Japanese verbal gerund works as a linker of two sentences, and generally, corresponding to *–ing* in English. In order to express an action in progress (e.g., I am doing), *"imasu (to exist, to be)"* is attached to the *-te* form, referring to the present moment.

For example,

1. *Hon o* ***yonde*** *imasu* (I am reading a book).

trans Book-*o* ***reading*** *be*

2. *Sushi o* ***tabete*** *imasu* (I am eating Sushi).

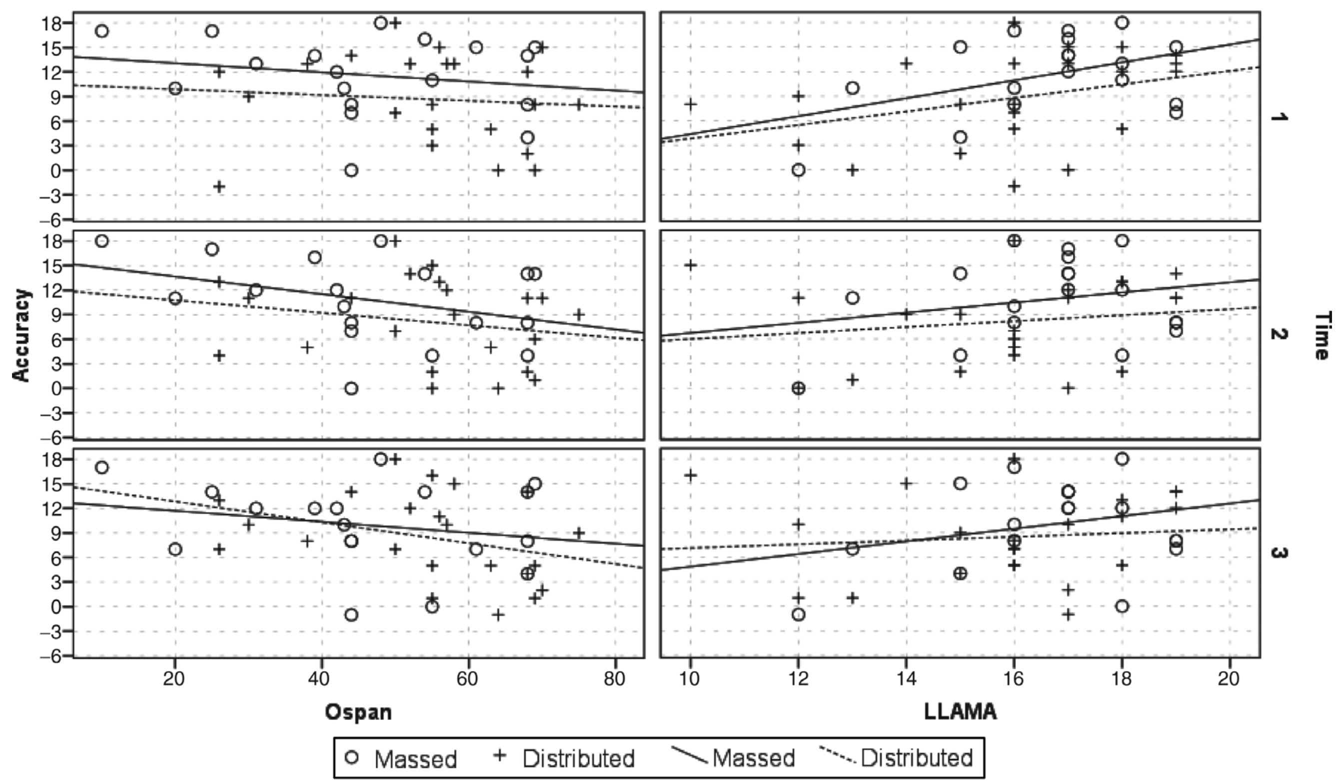
trans Sushi-o ***eating*** *be*

Conjugation of the *te*-form

Category	direct-style	*te*-form	English
r, u ➔ tte	tomaru	tomatte	To stop
	kau	katte	To buy
m, b ➔ nde	yomu	yonde	To read
	yobu	yonde	To call
k ➔ ite	aruku	aruite	To walk
g ➔ ide	nugu	nuide	To take off

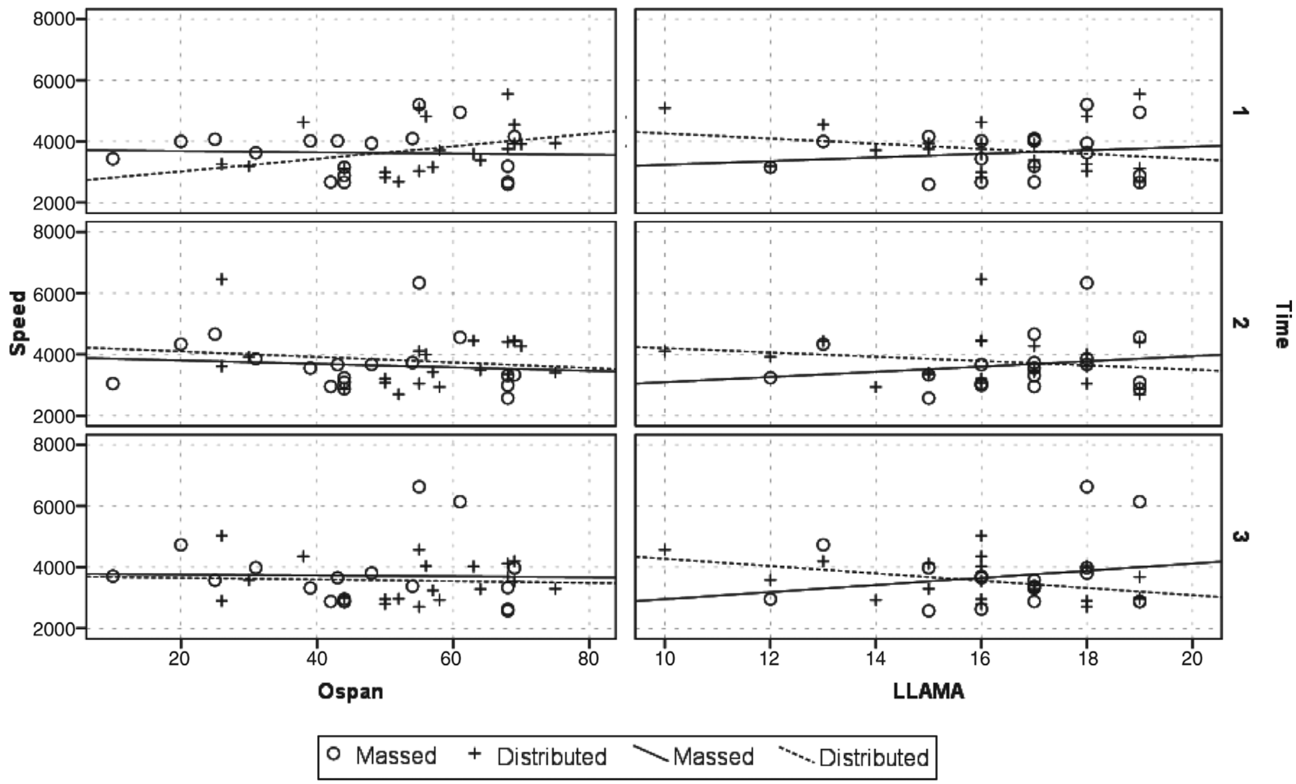
APPENDIX C

Scatter plots for accuracy, Ospan, and LLAMA scores on the rule-application test



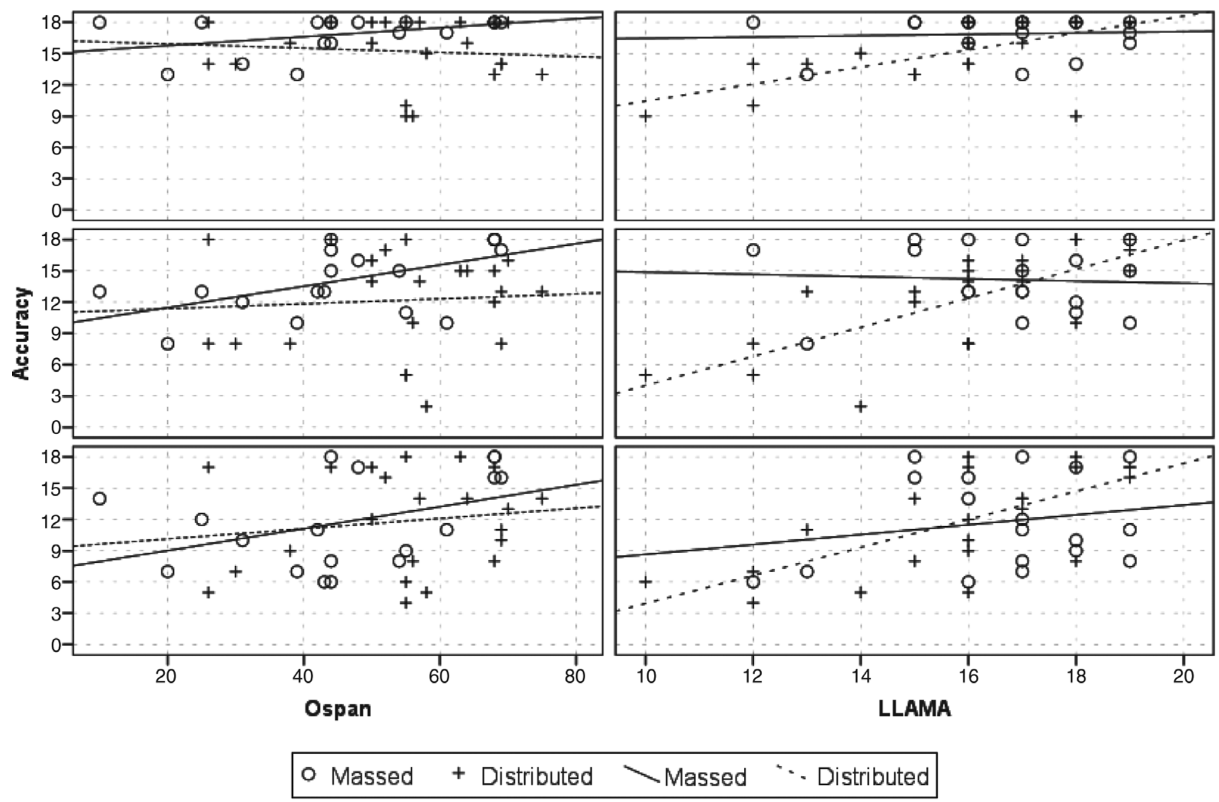
APPENDIX D

Scatter plots for speed, Ospan, and LLAMA scores on the rule-application test



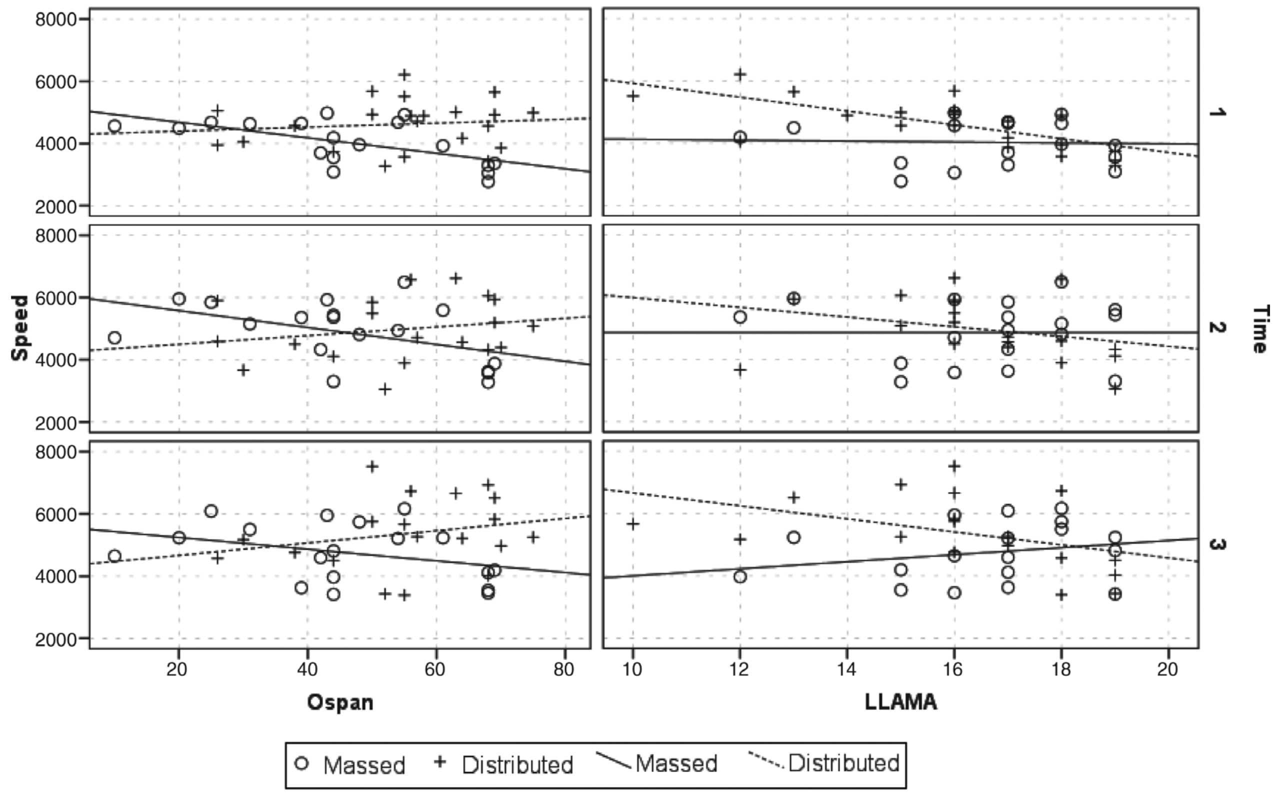
APPENDIX E

Scatter plots for accuracy, Ospan, and LLAMA scores on the sentence completion test



APPENDIX F

Scatter plots for speed, Ospan, and LLAMA scores on the sentence completion test



ACKNOWLEDGMENTS

This study was supported by the PhD program of SLA at the University of Maryland. We thank Dr. Steve Ross and Dr. Cathy Doughty for their insightful and constructive advice on this project. We also thank Yumiko Inagaki, Aya Shibasaki, Michiko Terazawa, Buthainah Al Thowani, and Assma Al Thowani for data coding. Many thanks to Ayami Suzuki, Taeko Hosaka, Kota Kato, and Shitaro Akiyama for material development. We are also grateful to Dr. Jaehyon Paik for sharing his manuscript; Shota Momma for his advice on test development; and Ms. Makiko Inoue, Ms. Kumiko Akikawa, and Mr. Carlos Pimentel for participant recruitment. Finally, we express our utmost gratitude to Kei Harata for his help throughout the project.

NOTES

1. The participants in the implicit learning condition were not instructed to focus on form, but the postquestionnaire on awareness showed that many of them searched for rules in the input, suggesting the involvement of conscious rule learning.
2. In Trofimovich et al. (2007), the instructional treatment, recast, was assumed to have induced form-focused learning to a large extent because learners were always asked to indicate whether their utterance and the recast were different or not.
3. Having grammar explanation may not always mean that learning burden is low. If grammar explanation consists of a large amount of complex information, for instance, the burden on the learner may increase.
4. The studies on feedback treatment were chosen here because they are among the few that tested the combination of LAA and WMC. Note, however, that explicit grammar instruction and learning through feedback cannot be equated with each other.
5. Because the training session consisted of vocabulary practice, grammar explanation, and practice, distributed/massed "instruction" might well be used. Distributed/massed "practice" was used throughout the paper because most of the training materials focused on practice vocabulary and grammar with a brief explanation.
6. The number of recasts was not recorded for each group, and it may differ between the groups.
7. A more general benchmark for interpreting effect sizes for *r* is given by Cohen (1988): small (.1), medium (.3), and large (.5), which is slightly lower than what was found for the L2 research domain in Plonsky and Oswald (2014). Given the dearth of research on L2 distributed/massed practice, it is hard to decide the most suitable benchmark for the present study, but we chose Plonsky and Oswald's benchmark because it was based on L2 research.
8. Another possible and more speculative interpretation is that L2 learners in the distributed practice group might have engaged more in rule-based learning, whereas item-based learning was more involved in the massed practice group (Skehan, 1998). Because more analytic learning was involved under this distributed learning condition, LAA played a more prominent role. In contrast, participants in the massed practice group encountered the learning items in clusters, and because the verb phrases may have been harder to analyze, they needed to rely more on exemplar-based learning in

which they remembered the *-te* form of the verb as a chunk. This learning process in the massed practice condition presumably taxed WMC more than in the distributed practice, which was indicated by the correlation between WMC and the learning gains in the massed practice.

REFERENCES

Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist, 38*, 85–93. doi:10.1207/s15326985ep3802_3

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29. doi:10.1146/annurev-psych-120710-100422

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—When and how? *Journal of Clinical Epidemiology, 54*, 343–349.

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

Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics, 27*, 185–209.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369–378.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Rowley, MA: Newbury House.

Carroll, J. B., & Sapon, S. M. (1959). *Modern Language Aptitude Test: MLAT*. New York: Psychological Corporation.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N. (2007). Variation in working memory: An introduction. In A. R. A. Conway, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 3–17). Oxford: Oxford University Press.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

De Graaff, R. (1997). The eXperanto experiment. *Studies in Second Language Acquisition, 19*, 249–276.

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

DeKeyser, R. M. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning, 62*, 189–200.

DeKeyser, R. M. (2013). *Aptitude*. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 27–31). New York: Routledge.

Doughty, C. J. (2013). Optimizing post-critical-period language learning. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 153–175). Philadelphia, PA: John Benjamins.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23.

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9*, 147–172.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, and Computers, 35*, 116–124.

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition, 34*, 445–474.

Granena, G. (2015). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics*. Advance online publication. doi:10.1017/S0142716415000120

Hwu, F., & Sun, S. (2012). The aptitude-treatment interaction effects on the learning of grammar rules. *System, 40*, 505–521.

Jackson, D. O. (2014). Learner differences in metalinguistic awareness: Exploring the influence of cognitive abilities and language experience. In R. T. Miller, C. M. Eddington, A. Henery, N. Marcos Miguel, A. Tseng, A. Tuninetti, et al. (Eds.), *Selected proceedings of the 2012 Second Language Research Forum: Building bridges between disciplines* (pp. 211–226). Somerville, MA: Cascadilla Proceedings Project.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 336–358.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.

Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning, 58*, 703–746.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"? *Psychological Science, 19*, 585–592.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*, 309–365.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal, 97*, 634–654.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics, 36*, 385–408. doi:10.1093/applin/amu054

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., et al. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63*, 530–566. doi:10.1111/lang.12011

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2013). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*, 861–883.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*, 501–533.

Meara, P. M. (2005). *LLAMA language aptitude tests*. Swansea: Lognostics.

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System, 42*, 412–428.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Erlbaum.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*, 8–14.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.

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

Ranta, L. (2002). The role of learners' language analytic ability in the communicative classroom. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 159–179). Amsterdam: John Benjamins.

Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult SLA. *Language Learning, 47*, 45–99.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211–266). Amsterdam: John Benjamins.

Robinson, P. (2007). Aptitudes, abilities, contexts, and practice. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 256–286). New York: Cambridge University Press.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*. Advance online publication. doi:10.1007/s10648-015-9332-4

Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 229–248). Oxford: Oxford University Press.

Sanz, C., Lin, H.-J., Lado, B., Stafford, C. A., & Bowden, H. W. (2014). One size fits all? Learning conditions and working memory capacity in ab initio language development. *Applied Linguistics*. Advance online publication.

Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning, 61*, 117–145.

Serrano, R. (2012). Is time concentration good or bad for learning? Reflecting on the results from the cognitive psychology and from the SLA literature. In C. Munoz (Ed.), *Intensive exposure experiences in second language learning* (pp. 3–22). Clevedon: Multilingual Matters.

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

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *Tesol Quarterly, 41*, 255–283.

Shintani, N., & Ellis, R. (2015). Does language analytical ability mediate the effect of written feedback on grammatical accuracy in second language writing? *System, 49*, 110–119.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning, 63*, 296–329.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–94). Philadelphia, PA: John Benjamins.

Skehan, P. (2012). Language aptitude. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 381–395). New York: Routledge.

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: A critical overview. *Applied Linguistics, 36*, 367–384. doi:10.1093/applin/amu072

Snow, R. E. (1987). Aptitude complexes. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning and instruction* (pp. 13–59). Hillsdale, NJ: Erlbaum.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*, 263–308.

Suzuki, Y., & DeKeyser, R. M. (2015). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*. Advance online publication. doi:10.1177/1362168815617334

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 437–444.

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

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 171–195). Oxford: Oxford University Press.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505. doi:10.3758/BF03192720

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany, NY: State University of New York Press.

Wesche, M. (1981). Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 119–154). Rowley, MA: Newbury House.

Williams, J. N. (2012). Working memory and SLA. In S. M. Gasss & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 427–441). New York: Routledge.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics, 34*, 344–368.