The Modern Language Journal

Investigating Distribution of Practice Effects for the Learning of Foreign Language Verb Morphology in the Young Learner Classroom

ROWENA E. KASPROWICZ,[1] EMMA MARSDEN,[2] and NICK SEPHTON[3]
[1] *University of Reading, Institute of Education, London Road Campus, 4 Redlands Road, Reading RG1 5EX, UK Email: r.kasprowicz@reading.ac.uk*
[2] *University of York, Department of Education, York YO10 5DD, UK Email: emma.marsden@york.ac.uk*
[3] *University of York, Digital Creativity Labs, York YO10 5GE, UK Email: nick.sephton@york.ac.uk*

Within limited-input language classrooms, understanding the effect of distribution of practice (spacing between practice) on learning is critical, yet evidence is conflicting and of limited relevance for young learners. For second language (L2) grammar learning, some studies reveal advantages for spacing of 7 days or more, but others for shorter spacing. Further, little is known about the role of cognitive individual differences (e.g., language analytic ability; LAA) in mediating practice distribution effects for L2 grammatical knowledge development and retention. To address this gap, this classroom-based study investigated whether distribution of practice and LAA moderated the effectiveness of explicit, input-based grammar instruction for young first language (L1) English learners of French (aged 8 to 11). The study revealed minimal differences between longer (7-day) versus shorter (3.5-day) spacing of practice for learning a French verb inflection subsystem, at either posttest or delayed posttest. Minimal group-level gains and substantial within-group variation in performance at posttests were observed. Accuracy of practice during training and LAA were significantly associated with posttest performance under both practice schedules. These findings indicated that within an ecologically valid classroom context, differences in distribution of practice had limited impact on learner performance on our tests; rather, individual learner differences were more critical in moderating learning. This highlights the importance of considering individual learner differences in the development of resources and the potential of digital tools for dynamically adapting instruction to suit individuals.

Keywords: distribution of practice; foreign language learning; game-based learning; grammar; lag effects; language analytic ability; young learners

The Modern Language Journal, 0, 0, (2019)
DOI: 10.1111/modl.12586
0026-7902/19/1–27 $1.50/0

This article has been awarded Open Data and Open Materials badges. All data and materials are publicly accessible via the IRIS repository at https://iris-database.org. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

ENGAGING IN EXTENSIVE, REPEATED, meaningful practice is an essential component of learning, facilitating the transition from initial reliance on declarative knowledge (e.g., explicit knowledge of a grammatical rule) to proceduralized and eventually automatized knowledge that can be accessed more efficiently under time pressured contexts such as spoken interaction (DeKeyser, 2007, 2015; Lightbown, 2008; Segalowitz, 2003). Evidence suggests that practice that draws attention to linguistic features can be particularly useful for learning forms that have low salience, low communicative value, or complex relationships between first (L1) and

second (L2) language (e.g., Doughty & Williams, 1998; R. Ellis, 2006; Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011; McManus & Marsden, 2017, 2018, 2019a, 2019b; VanPatten, 2015). However, whilst there has been extensive focus on the nature of practice required to facilitate L2 learning, an important remaining question concerns the amount and frequency of practice that is needed to maximize its effectiveness (DeKeyser, 2015; Rogers, 2017).

This question, although relevant to all learning and skill development, is particularly pertinent to the foreign language (FL) classroom, where class time is severely limited (Swanson & Mason, 2018) and there is little exposure outside of the classroom. For example, the Australian Curriculum recommends 350 hours across 7 years of schooling between Foundation (age 4–5) and Year 6 (age 11–12), approximately 1.25 hours per week (Australian Curriculum, Assessment and Reporting Authority, 2011, p. 28). Similarly, in the United Kingdom, children between the ages of 7 and 11 receive on average 30 to 60 minutes per week (Tinsley & Board, 2017). Teachers must, therefore, decide how to allocate this short time in order to maximize learning. For example, primary schools in the United Kingdom can either offer two shorter FL sessions per week or one longer session. Anecdotal evidence suggests considerable debate at local and national levels about such decisions, and yet there is little research demonstrating whether one approach is more beneficial than another.

The question of how practice should be distributed to facilitate learning and retention of knowledge has received extensive attention within cognitive psychology (Cepeda et al., 2006), yet only a handful of studies have addressed this question in relation to L2 grammatical knowledge development (Bird, 2010; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017a). Such studies have yielded conflicting results, in part due to methodological differences (e.g., length and nature of instruction, nature of tests), and have focussed exclusively on adult learner populations. Additionally, whilst increasing attention has been paid to the role of individual cognitive differences (e.g., language analytic ability [LAA], working memory) in moderating the effectiveness of a given type of L2 practice (e.g., Li, 2015), little is known about their role in mediating learning under different practice schedules (Suzuki & DeKeyser, 2017b).

The present study therefore aimed to contribute to this area of research by investigating the impact of a) practice distribution, and b) LAA on L2 grammar learning by young learners in the primary school classroom, an underresearched population and context.

Another purpose of the current study is to explore the potential of digital language-learning tools to enable learners to engage in practice and "offer a [still largely] unexploited opportunity to schedule study sessions in ways that optimize long-term retention" (Rohrer & Pashler, 2007, p. 186). Additionally, such tools provide a rich source of data for improving our understanding in this area, in situ in classrooms, without compromising control over experimental design and internal validity, and can therefore enable more robust investigation of causal relationships between training and testing performance under different conditions.

LITERATURE REVIEW

Skill Acquisition and Practice Distribution

Practice plays a critical role in skill acquisition theories of learning (DeKeyser, 2015). Deliberate, intensive practice enables learners to move from an initial reliance on declarative, explicit knowledge to the development of procedural knowledge, which may in turn become automatized given sufficient practice opportunities. Such theories posit that these processes apply to the development of a wide range of skills, including L2 learning. Optimizing not only the nature but also the sequence and spacing of practice is therefore critical for efficient learning.

Studies from cognitive psychology have consistently demonstrated that temporally spacing practice sessions leads to better learning and retention than massing practice into a single session (for a review, see Cepeda et al., 2006); the so-called *spacing effect*. Of even greater relevance to the instructed classroom context, where instruction tends to be interspersed over days or weeks, is the question of whether varying the spacing (i.e., amount of time) between practice sessions also affects how well and for how long learnt information is remembered; the so-called *lag effect*. The time between practice sessions is known as the intersession interval (ISI). A comprehensive comparison of multiple ISIs by Cepeda et al. (2008) revealed an interdependence between the optimal ISI and the amount of time between the final practice session and the testing time, which is known as the retention interval (RI). They demonstrated that for the learning of trivia facts, as the RI increased, the optimal ISI also increased. For example, for a RI of 7 days, the optimal ISI was 3 days,

whereas for a RI of 35 days, the optimal ISI was 8 days. The optimal spacing of practice sessions, therefore, seems to be dependent upon when the learnt knowledge will be needed (e.g., in testing or use).

Numerous theoretical accounts have been proposed to explain the findings that (a) spacing practice is beneficial for learning, and (b) provision of longer spacing between practice sessions leads to better knowledge retention (Toppino & Bloom, 2002). Study-phase retrieval (Toppino & Bloom, 2002) and reminding (Benjamin & Tullis, 2010) accounts propose that successful retrieval of a previously learnt item at a later time point will serve to strengthen the representation of that item, particularly when successful reminding or retrieval occurs after a "high degree of forgetting or a low amount of reminding" (Benjamin & Tullis, 2010, p. 239).

Encoding variability accounts (Benjamin & Tullis, 2010) posit that it is not only the fact of having multiple retrieval opportunities that is important, but also the nature of the retrieval that occurs. Such accounts propose that environmental and contextual differences between practice sessions will result in each occurrence of a learning item being encoded differently, resulting in multiple effective retrieval routes. Similarly, the notion of transfer-appropriate processing suggests that providing multiple, varied practice opportunities will enable the learner to generate "richer, more contextualized representations of the learned material" (Lightbown, 2008, p. 38). An item encountered in a range of contexts is likely to have multiple associations, which will facilitate retrieval across different contexts. Such proposals tie into the concept of providing "desirable difficulty" (Bjork & Bjork, 2014, p. 58) in practice activities and sessions, in order to bring about deeper processing of target items and subsequently better learning. Bjork and Bjork propose that creating situations in which the learner has to work harder to retrieve information from long-term memory (e.g., through distributing practice sessions, varying practice contexts, and introducing contextual interference) will ultimately result in better long-term retention.

These accounts provide complementary interpretations for the general finding that allowing spacing between practice sessions improves learning and retention of target items and further that the amount of time between sessions should be balanced to create effortful retrieval, whilst limiting the likelihood of unsuccessful retrieval or complete forgetting. The question remains, however, as to the relevance of lag effects for L2 grammar learning.

Distribution of Practice Effects for L2 Grammar Learning

The investigation of lag effects (i.e., comparisons of two or more practice distributions varying in length) has been the focus of a small but growing number of SLA studies. Studies have begun to explore lag effects for L2 grammar learning (e.g., Bird, 2010; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017a) and vocabulary learning (e.g., Nakata, 2015; Serrano & Huang, 2018), as well as general L2 proficiency in intensive versus extensive instructional programmes (e.g., Collins & White, 2011; Serrano & Muñoz, 2007).

The four studies most relevant to the current study (on L2 grammar learning in FL contexts) have yielded conflicting results. See Appendix A for a detailed tabular overview of their designs and findings. Bird (2010) observed superior learning under a 14-day ISI than a 3-day ISI condition for L1 Malay learners of the L2 English tense and aspect system, when measured on a written error-correction task at delayed posttest (60-day RI). Similarly demonstrating benefits for spacing that is longer than 2–3 days, Rogers (2015) observed for L1 Arabic learners of L2 English cleft syntactic structures that a 7-day ISI led to superior performance at delayed posttest (42-day RI) than a 2.25-day ISI on a written grammaticality judgement test.

In contrast, Suzuki and DeKeyser (2017a) and Suzuki (2017) found some benefits for spacing that was shorter than 7 days. Suzuki and DeKeyser examined the learning of L2 Japanese present progressive verb morphology under a 1-day and 7-day ISI. They observed an advantage for the shorter ISI, in terms of response speed on an oral picture-description task at delayed posttest (28-day RI). Extending these findings, Suzuki (2017) observed superior gains in accuracy on an oral production task for a 3.3-day group compared to a 7-day group at delayed posttest (28-day RI) for the learning of simple and complex morphology within an artificial language system.

As Appendix A illustrates, there was substantial variation between the studies, which may to some extent account for the difference in findings. The studies utilized different interventions (varying in type and amount), outcome measures, and language features (though Suzuki, 2017, was a conceptual replication of Suzuki & DeKeyser, 2017a). Differences in treatment and task

complexity (Donovan & Radosevich, 1999) and the type of knowledge trained and elicited (Suzuki & DeKeyser, 2017a) may have contributed to the contradictory findings. In addition, each study utilized a slightly different set of ISIs and RIs, which, as described previously, can impact test results (Cepeda et al., 2008). Further, the participants in Bird (2010) and Rogers (2015) were identified as intermediate-level learners but as beginners in Suzuki & DeKeyser (2017a) and Suzuki (2017). Practice distribution effects may manifest differently at different proficiencies, with, say, shorter spacing being more helpful among beginner learners or, more generally, lag effects being more difficult to observe at lower proficiencies. It is also important to note that the larger ISI conditions in these studies distributed practice sessions over a longer period of time than the shorter ISI conditions; for example, four sessions over 4 weeks (7-day ISI) versus four sessions over 2 weeks (3.3-day ISI) in Suzuki (2017). In the FL classroom, however, teachers are unable to extend overall teaching time; therefore, a more relevant question concerns how practice can be optimally distributed within the specified curriculum time.

The conflicting findings highlight that lag effects for L2 grammar learning may be influenced by a number of factors, including the amount and nature of training, the nature and modality of testing tasks, the nature of knowledge, and individual learner characteristics. Further research is needed to paint a clearer picture of the role of lag effects for different types of learners engaging in different types of L2 grammar practice.

Distribution of Practice and Child L2 Learning

Critically, it is also important to note that the four studies mentioned previously were conducted with similar learner populations (i.e., adult, university-based learners) and Cepeda et al. (2006) noted that 85% of the studies in their meta-analysis were conducted with adults. Whilst there is emerging evidence that younger learners can benefit from focussed, explicit practice in particular language features (e.g., Kasprowicz & Marsden, 2018; Lichtman, 2016), explicit learning by younger learners tends to be slower than for older, more cognitively mature learners. Further, younger learners' cognitive abilities (e.g., LAA, working memory) are still developing, which may affect the extent to which they are able to store, access, retain, and recall target knowledge over distributed practice schedules. An as yet underexplored question therefore relates to the role of lag effects for L2 learning by younger learners.

A small number of studies have found advantages for distributed practice over massed practice for language learning among children (e.g., Fishman, Keller, & Atkinson, 1968; Lotfolahi & Salehi, 2016). Additionally, some research (e.g., Collins et al., 1999; Collins & White, 2011) has investigated intensive (5-month) versus more distributed (10-month) language programmes, but as this research was at the programme level and outcomes measures were wide ranging, the findings are less relevant to the rationale for the current study.

In sum, there is a limited amount of research into lag effects with children, particularly studies investigating longer time periods (i.e., ISIs of days or weeks for learning at RIs of weeks or months; Cepeda et al., 2006). Further research is needed to investigate interactions between practice distribution and specific aspects of L2 learning (e.g., grammatical knowledge development), on a range of measures, for young learners. Another issue that has been neglected to date is the potential influence of individual differences on lag effects. We now turn to one such difference, a component of aptitude: LAA.

Language Analytic Ability

LAA can be defined as "the capacity to infer rules of language and make linguistic generalizations or extrapolations" (Skehan, 1998, p. 204). LAA can be further broken down into two subcomponents: grammatical sensitivity (the ability to recognize the grammatical function of words) and inductive learning ability (the ability to infer the grammatical rules governing a set of language; Carroll, 1990; Roehr, 2008), both key to identifying and extrapolating linguistic patterns (Skehan, 2002). Given the emphasis on pattern recognition and application, LAA is thought to be particularly relevant to explicit language learning (DeKeyser, 2012; Robinson, 1997; Roehr, 2008; Skehan, 2002). We would also add deductive language-learning ability to previous models of LAA, the ability to understand a rule and apply it consistently where appropriate. This is likely to be particularly relevant to learning under instruction, where rules are frequently given before practice.

Language Analytic Ability and Lag Effects

To the best of our knowledge, only one study (Suzuki & DeKeyser, 2017b) has investigated the

relationship between components of aptitude and learning under different practice distributions. Suzuki and DeKeyser investigated whether LAA and working memory capacity moderated learning under shorter (1-day) and longer (7-day) ISIs. The results of Suzuki and DeKeyser's study indicated a clear interaction between aptitude and treatment for their adult L1 Japanese learners of L2 English, with LAA correlating positively with learning under the longer practice distribution and working memory with learning under the shorter practice distribution. The reasons why LAA played a role in the more distributed practice but not in the less distributed practice, whilst both involved the same grammar instruction, are not clear. As discussed previously, it is hypothesized that distributed practice benefits learning if an individual can recall previously learnt information at the time the practice occurs (Benjamin & Tullis, 2010; Toppino & Bloom, 2002). It may be, then, that higher LAA enables learners to establish more robust or accurate initial knowledge of the target structure, which can then be recalled more successfully in later sessions, thereby allowing learners with high LAA to benefit to a greater extent from longer spacing. Nevertheless, as acknowledged by Suzuki and DeKeyser (2017b), given the limited research in this area such a conclusion is tentative.

Language Analytic Ability and L2 Grammar Learning by Young Learners

Whilst numerous studies (e.g., Erlam, 2005; Li, 2015; Ranta, 2002; Robinson, 1997) have demonstrated that LAA does indeed influence L2 grammar learning by adolescents and adults, the role of LAA for younger learners has received much less attention. DeKeyser (2012) proposed children may rely less on more analytical components of aptitude such as LAA, due to the different learning processes involved in child versus adult learning, with children relying on more implicit learning and older learners on their developing explicit learning abilities (see also Doughty, 2003). Indeed, some findings suggest a much smaller or nonexistent role for LAA among young learners compared to older learners (e.g., DeKeyser, 2000; Harley & Hart, 1998), whereas others have observed that LAA can be predictive of L2 performance by young learners in both immersion classrooms (Ranta, 2002) and naturalistic contexts (Abrahamsson & Hyltenstam, 2008). However, these studies with younger learners have tended to be within naturalistic or immersion settings with sufficiently large amounts of input to facilitate implicit learning processes, leading to a greater reliance on memory-based components of aptitude than on analytical abilities (DeKeyser, 2012). In contrast, younger learners' ability to draw on implicit learning mechanisms is restricted by the severely limited exposure of FL classrooms (DeKeyser, 2000; Muñoz, 2008). However, there is evidence that with explicit instruction, younger learners can begin to learn explicitly (Kasprowicz & Marsden, 2018; Lichtman, 2016). Therefore, young learners' developing analytical abilities may play a role in such settings.

Only a handful of studies have investigated LAA and L2 grammar learning by young learners within the instructed FL context (e.g., Hanan, 2015; Kiss & Nikolov, 2005; Muñoz, 2014; Tellier & Roehr–Brackin, 2013, 2017). Tellier and Roehr–Brackin (2013) investigated the relationship between aptitude, measured by the Modern Language Aptitude Test-Elementary (MLAT-E), and L2 French learning by L1-English children aged 8–9. LAA was a significant predictor of L2 proficiency and correlated significantly with gains in grammar knowledge, as well as listening and reading abilities. Similarly, Kiss and Nikolov (2005) found that aptitude, including language analysis and grammatical sensitivity, was the strongest predictor of L2 English proficiency for young L1 Hungarian learners (aged 11 to 12), explaining over 20% of the variation in scores. These studies provide some evidence that LAA can indeed relate to L2 learning for young instructed FL learners.

The current study sought to expand on this research by not only investigating whether LAA moderated L2 grammar learning by young learners but also whether there was a differential effect depending on practice distribution.

RESEARCH QUESTIONS

The aim of this study is to explore the effects of longer versus shorter spacing of practice sessions in L2 grammar (inflectional verb morphology) learning in a hitherto underresearched learner population (young, beginner learners) in an ecologically valid FL classroom, and also to investigate whether LAA moderated learning success under either practice distribution. An additional contribution of our study is that, unlike in much instructed SLA research where performance during practice is not documented or reported, our digital tool enabled recording of learners' accuracy during the training. This allowed us to identify, rather than assume, any causal relationship between training and posttest

performance under either practice distribution. To this end, the following research questions were addressed:

RQ1. To what extent do shorter (3.5-day) and longer (7-day) spaced practice schedules influence development of verb inflections in young L1-English learners of L2 French?
  (a) To what extent does accuracy during input-based training moderate learning outcomes under 3.5-day and 7-day spacing schedules?
  (b) To what extent does LAA moderate learning outcomes under 3.5-day and 7-day spacing schedules?

METHOD

Participants

One hundred and thirteen beginner-level L1-English learners of L2 French (aged 8–11) from eight classes across seven primary schools participated in the study (60 boys, 53 girls). Six of the schools were part of one school alliance and were invited to participate following a presentation at a FL teacher-training session. The seventh school was located in the same region. The children had been learning French in school for a minimum of one academic year prior to the study and had minimal access to the language outside of the classroom.

Prior to the current study, French instruction tended to consist of weekly 40- to 60-minute lessons, focussing on learning of key vocabulary, development of comprehension and production skills, and some word-level grammar instruction (e.g., definite and indefinite articles, gender, pronouns, adjective agreement). There is no set scheme of work for UK primary school FL teaching; therefore, the exact content of language lessons in each class varied. Due to the large variation in FL teaching provision across UK primary schools (Tinsley & Board, 2017), this was impossible to avoid. To account for its potentially confounding effect, class was included as a random variable in the analysis (see *Analysis* section). Additionally, a 4-week preexperimental phase was included. All classes received four lessons introducing the core vocabulary utilized in the main experimental instruction. Each class teacher completed the activities with their class. The researcher observed one preexperimental lesson per class to ensure the materials were delivered consistently across classes.

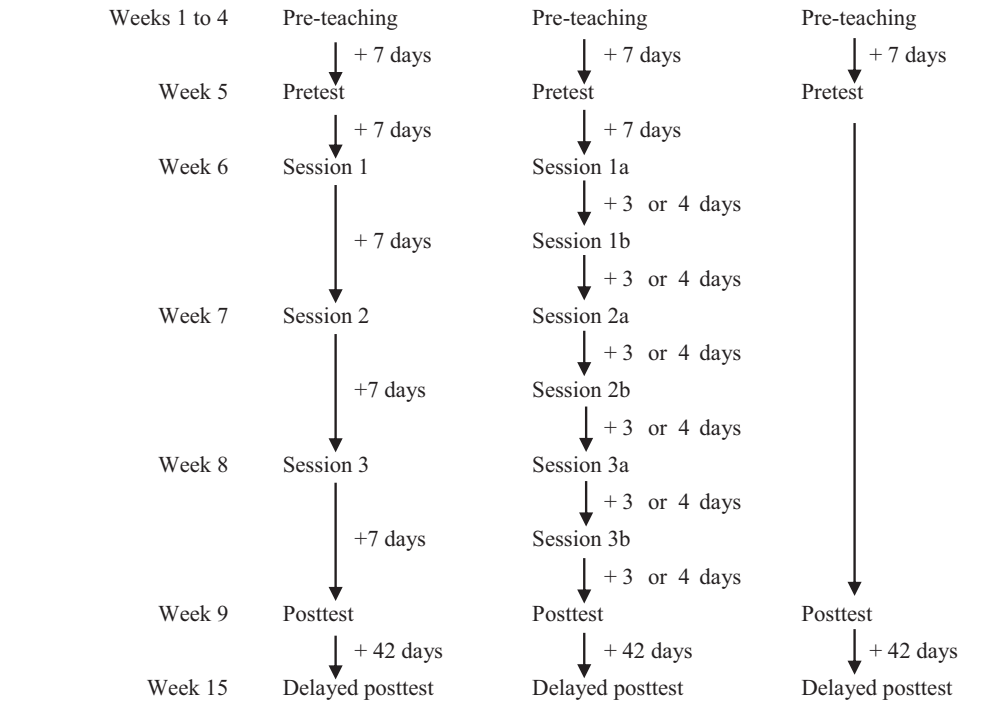
Intact classes were assigned to the experimental conditions (7-day, 3.5-day, and control group). Random assignment within classes was not possible, due to practical constraints: (a) It was not possible to have learners within the same class completing the training activities at different days or times, and (b) having control and treatment participants within the same class would have increased the likelihood of control participants being exposed to the treatment. The 7-day group included two mixed Year 5/6 classes (ages 9–11) and one Year 5 class (ages 9–10). The 3.5-day group included two Year 5 classes and one Year 4 class (ages 8–9). The control group included one mixed Year 4/5/6 class (ages 8–11) and one mixed Year 5/6 class.

Procedure

A quasi-experimental design was employed. The control group completed the tests only and reverted to their normal French lessons between pre- and posttests. The treatment groups undertook identical tasks, both totalling 180 minutes (see *Training* section) but differing in the distribution of the sessions (all treatment and testing materials are available on IRIS). The ISIs were 7 days and 3.5 days, in line with Suzuki (2017) and reflecting the most common lesson frequency in UK primary schools (one or two lessons per week). The 7-day group completed three sessions of 60 minutes, each occurring 7 days apart, whereas the 3.5-day group completed six sessions of 30 minutes each occurring 3.5 days apart. Figure 1 illustrates the timing of each testing and training session. The timing of the posttest mirrored the respective ISI for each treatment group (ISI:RI ratio = 100%). The delayed posttest took place exactly 6 weeks after the posttest, giving an RI of 42 days for both groups (ISI:RI ratio = 16.7% for 7-day group, 8.3% for 3.5-day group); the ISI:RI ratio was calculated from the first posttest, rather than the final intervention session, as the first posttest provided an additional opportunity for practice. The 3.5-day group's ISI:RI ratio fell just outside Rohrer and Pashler's (2007) observed optimum of 10 to 30%. The timing was chosen to ensure that all classes could adhere to the schedule, whilst fitting within the constraints of the schools' timetables and term dates.

Participants were included in the analysis if they had attended all training sessions and both posttests. One 3.5-day participant did not complete the sentence–picture matching pretest, whilst two 7-day participants and four 3.5-day participants did not complete the acceptability

FIGURE 1
Study Schedule



judgement test (AJT) pretest. Due to limited time available for testing, a number of participants were unable to complete the AJT at post- and delayed posttest; therefore, the participant number for this task is lower.

Target Feature

The target language system being taught and tested was regular French verb inflections in the present and perfect tenses (Table 1) for first- and third-person singular and plural forms: null (*-e*), *-ons*, *-ent* (present tense number inflections) and *ai* and *a* (*avoir* auxiliaries for the perfect tense), in both oral and written forms. This choice was in line with the curriculum, which states that children should be taught the "conjugation of high frequency verbs" (Department for Education, 2013, p. 2). The participants had not previously received explicit instruction in the features. Such features can be problematic for L2 learners due to an overreliance on lexical items that convey the same semantic information (e.g., subject pronouns indicating person and number; temporal phrases indicating tense), as noted in the Lexical Preference Principle (VanPatten, 2015); see also Marsden (2006) for a study using a similar rationale and Processing Instruction to focus on the same inflectional features with slightly older learners in the same educational context. Additionally, associative theories of learning attribute such difficulties to phenomena such as entrenchment, attention blocking, and overshadowing, which account for effects of (L1) prior experience, salience, and frequency in the input (N. Ellis, 2006).

Training

Training for the 7-day and 3.5-day groups was delivered via a bespoke, digital, game-based application containing a series of mini-games, with each game teaching just one particular grammatical contrast that is expressed by one pair of inflections (e.g., first person singular vs. plural present tense inflections; see Table 1). Training was completed on individual laptops with headphones. All sessions were overseen by the first author; class teachers were present during the training sessions but provided technical support only.

Each mini-game utilized form–meaning mapping activities consisting of brief (approximately 2 minute) explicit information followed by referential reading and listening activities. Referential

TABLE 1
Training Activities

Mini-game	Target Features	Cue(s) Removed	Question Sets	Items[a]
A	First person singular (*je jou**e*** 'I play') versus	Pronouns: *je, nous*	R&L (+T), R, L,	48–56
	First person plural (*nous jou**ons*** 'we play')		R&L (recap)	
B	Third person singular (*il/elle jou**e*** 'he/she plays') versus	Pronouns: *il/elle, ils/elles*	R&L (+T), R,	48–56
	Third person plural (*ils/elles jou**ent*** 'they play')		R&L, R&L (recap)	
C	First person present (*je joue* 'I play') versus	Temporal adverbs &	R&L (+T), R, L	36–42
	First person past (*j'**ai** joué* 'I played')	Past participle *é* inflection		
D	Third person present (*il/elle joue* 'he/she plays') versus	Temporal adverbs &	R&L (+T), R, L	36–42
	Third person past (*il/elle **a** joué* 'he/she played')	Past participle *é* inflection		
E	First person past (*j'**ai** joué* 'I played') versus	Pronouns: *je, il/elle*	R&L (+T), R, L,	48–56
	Third person past (*il/elle **a** joué* 'he/she played')		R&L (recap)	

Note. R&L = reading and listening; R = reading only; L = listening only; T = tutorial.
[a]Range from minimum number if all answers correct to maximum number of attempts possible.

activities (a component of Processing Instruction; VanPatten, 2015) are input-based tasks that require learners to notice a feature and connect it with a meaning or function to complete the activity. Numerous studies have demonstrated the effectiveness of such activities for L2 morphosyntax (e.g., DeKeyser & Botana, 2015; Kasprowicz & Marsden, 2018; Marsden & Chen, 2011; Shintani, 2015), including Marsden (2006), who also investigated the teaching of French inflectional verb morphology for person, number, and tense among FL learners aged 13–14 years; and McManus and Marsden (2017, 2018, 2019a, 2019b), who investigated the learning of French *imparfait*. Referential activities make the target grammatical feature task-essential by removing other cues that learners could rely on (e.g., subject pronouns indicating person and number; temporal adverbs indicating tense). For example, in the current study, in mini-game A (first person singular [null] vs. plural [*-ons*] present tense inflections), a robot described the food that it (*je* 'I') or all the robots (*nous* 'we') liked. The learner chose whether to feed only the robot that spoke or all the robots. After the first set of practice items, the subject pronoun was obscured (by *** in the reading version and by a beep in the listening version). The learners, therefore, had to notice the verb inflection, interpret its number meaning, and feed the correct robot(s). The training utilized 12 regular *-er* verbs (see Appendix B), which were chosen because they are commonly taught to beginners (e.g., *aimer* 'to like'), are cognates of English verbs (e.g., *poster* 'to post'), or fit the game context (e.g., *surveiller* 'to watch or survey').

Each mini-game contained three question sets (see Table 1). The first question set included the tutorial (brief explicit information provided alongside the first two question items; see Appendix C). Response options for the tutorial question items were restricted so that the learner had to answer correctly. Learners then completed the main question items. Correct and incorrect responses were indicated aurally by different sounds and visually via the progress bar. Following incorrect answers, learners also received a short explanation (see Appendix D). To successfully complete a question set, the learner had to answer 12 items correctly and received a number of stars upon completion (three stars if all correct, two stars if one incorrect, one star if two incorrect). Learners answered up to two additional questions for each previous item that had been answered incorrectly. If the learner answered three items incorrectly, they lost the question set and had one opportunity to replay the set. After one replay, the learner automatically moved on to the next question set, regardless of their score. The restriction

of one replay was included to ensure that all learners had the opportunity to answer all question sets within the time available. These success and replay features also helped to maintain engagement in the game.

Learners completed all three question sets for one mini-game (one grammatical contrast) before moving onto the next mini-game (and next grammatical contrast). The order of mini-games was counterbalanced across the 7-day and 3.5-day groups, with learners either practising present tense inflections before past tense inflections or vice versa. The 3.5-day group completed one mini-game (three question sets) in each session; the 7-day group completed two mini-games (six question sets) in each session. In the final part of the training (second half of session 3 for the 7-day group; session 3b for the 3.5-day group), the learners completed a final additional question set from mini-games A, B, and E, in order to review each of the grammar features.

Test Materials

Learners completed, on laptops, a sentence–picture matching test and an AJT at pre-, post- and delayed posttest. Three versions of each test were created. Each version contained the same number of items, in the same format, and included stimuli created from the same set of lexical items, but with different noun–verb combinations (see Appendix B for the list of verbs included). The three versions were counterbalanced within each experimental group and class, and each learner completed a different version at each time point. Learners ($n = 22$) of equivalent age and language experience to the main study participants piloted the tests to check the comprehensibility of the instructions, test format, and picture stimuli. Similar tests have been utilized in previous studies with participants of a similar age (e.g., sentence–picture matching task test, Kasprowicz & Marsden, 2018; AJT, Marsden & Chen, 2011).

Sentence–Picture Matching Test. Learners saw a sentence containing a target feature and chose which of two images matched the sentence. The test contained eight items, four for number and four for present or perfect tense inflections (first and third person). The limited time available in class for testing necessitated the low number of items.

For the items targeting number inflections, pronouns were obscured, for instance, *** *joue au foot*, '*** playSING football' and learners chose between a picture of one person and a picture containing three people. For the items targeting the tense inflections, temporal phrases were eliminated to test interpretation of the presence or absence of the perfect tense auxiliary, for instance, *j'ai joué au foot* 'I played football.' The pictures were an arrow pointing down (to indicate happening now) and an arrow pointing to the left (happened in the past). Learners completed training before the test that ensured they consistently understood the meanings of the pictures (e.g., "Which picture means *we*?" [one person or three people]; "Which picture means *happened in the past*?" [down or left arrow]). One point was awarded for selecting the correct image, giving 8 possible points.

Acceptability Judgement Test. Learners were presented with a series of sentences and told, "There may be a mistake in some of the sentences. Decide whether each sentence is right or wrong." Learners answered on a 4-point scale (*definitely right*, *right*, *wrong*, *definitely wrong*). If *wrong* or *definitely wrong* was selected, the learner was asked to "click on any word or words that are wrong" and then write the correct word in the text box provided.

There were six grammatical (G) and six ungrammatical (UG) items. For the number items, the error was due to a mismatch between the pronoun and the inflection (e.g., *je jou***ons*** *au foot* 'I play* football'). For the tense items, the error was due to the absence of the auxiliary (e.g., *Hier, je* joué au foot* 'Yesterday, I play* football').

For G items, learners received 1 point if they correctly selected *right/ definitely right*. For the UG items, learners received 1 point if they correctly selected *wrong/ definitely wrong* and clicked on the correct word(s) in the sentence (e.g., for number items, the pronoun or incorrectly inflected verb; for the tense items, the pronoun, verb or temporal phrase). Learners' corrections of UG items were scored separately, as producing correct versions likely constitutes a slightly different knowledge or skill to recognizing ungrammaticality as it involves production. The presentation of those results is beyond the scope of this article. Note, however, that incorporation of these correction scores in the AJT accuracy scores did not change the patterns of results found and presented here.

Language Analytic Ability Test. The LAA test was a paper-and-pencil test consisting of two parts. Part 1 contained five questions, adapted from the standardized UK Department for Education's spelling, punctuation, and grammar test (Standards and Testing Agency, 2014, 2015, 2016), which targets learners' knowledge of grammatical terminology and concepts in their

L1 English, including grammatical rules relating to pronouns, number, and tense. These questions tested metalinguistic knowledge and grammatical sensitivity, in line with our expanded definition of *LAA* (i.e., including deductive as well as inductive learning abilities; see *LAA* section). Part 2 contained four questions testing learners' ability to spot patterns and apply rules to novel language. The questions, adapted from Tellier (2013) and the UK Linguistics Olympiad (UKLO, 2016), tested learners' ability to separate noun and verb stems from inflections and spot patterns relating to changes in number and tense. This bespoke measure was used as existing LAA measures can be problematic due to their length and difficulty (e.g., LLAMA-F, see Rogers et al., 2017), thus are unsuitable for young children, and do not necessarily measure the full construct of LAA (i.e., including grammatical sensitivity and inductive and deductive learning abilities); for example, MLAT-E (Part 2) focusses solely on grammatical sensitivity.

Each question item was scored 0/1 for incorrect or correct answer, with 30 points available in Part 1 and 14 in Part 2.

Instrument Reliability. Ordinal omega hierarchical was calculated as a measure of test reliability for the sentence–picture matching and AJT tests, as it is considered appropriate for binomial, unit-weighted scales, which do not meet the assumption of unidimensionality (McNeish, 2018): sentence–picture matching, pretest = .28, delayed posttest = .44; AJT (G items), posttest = .81, delayed posttest = .73; AJT (UG items), posttest = .74, delayed posttest = .79. The reliability indices could not be calculated for the matching posttest data or the AJT G and UG pretest data, because R returned an N/A response for these subsets of data, possibly due to problematic factor scores.[1] However, the indices elicited for the same tests at the two other time points give a good indication of the reliability of each measure. The indices yielded for the sentence–picture matching test indicated that the items were not consistent with each other, possibly due to the small number of items or a high incidence of guessing in participants' responses (Bush, 2015). An additional reason may be because the items elicited different verb inflections, some of which may have been more difficult than others. Question was included as a random variable (as described in the next section) in analysis of test performance, to account for variation across question items. Nevertheless, given the low reliability of the sentence–picture matching test and missing indices, the results should be interpreted with caution.

Omega total, which is appropriate for use with unit-weighted, congeneric scales (McNeish, 2018), was calculated as a measure of reliability for the LAA test (.78).

Analysis

Descriptive statistics (means, standard deviations) of raw scores on each test are provided. Effect sizes (Cohen's *d*, calculated using the pooled standard deviation) and their confidence intervals (CIs), for comparisons between groups and between time points, are interpreted based on Plonsky and Oswald's (2014) field-specific medians (between-group: small, $d = 0.40$; medium, $d = 0.70$; large, $d = 1.00$; within-group: small, $d = 0.60$; medium, $d = 1.00$; large, $d = 1.40$; p. 889).[2] (Additionally, between-group effect sizes corrected for differences at pretest are provided in Appendix E. Although these corrected effect sizes give some descriptive indication of change that takes into account baseline differences, there is unfortunately no known way to date for calculating CIs for these corrected effect sizes, making them inappropriate to interpret within the main article.) Data were nonnormally distributed; therefore, Spearman's *rho* (including bootstrapped, bias-corrected, 95% CIs) is provided for analysis of the relationship between performance on the outcome measures and (a) the LAA test, and (b) practice accuracy.[3] The strength of the relationship indicated by Spearman's *rho* is interpreted against the following benchmarks: small = 0.25; medium = 0.4; large = 0.6 (Plonsky & Oswald, 2014, p. 889).

To model the effect of the categorical variables group (7-day, 3.5-day, control) and time (pre-, post-, delayed posttest) on test performance, and to account for random effects of learner, class, and question item, the data were analysed via mixed-effects logistic models fit by maximum likelihood with binomial logit functions using the lme4 package in R 3.4.3 (Bates et al., 2015). The data were binary (correct or incorrect). The models included random intercepts to account for variation in average scores by learner, class, and question item. The base model for analysis of each outcome measure can be described as follows:

```
model < −glmer(Score ∼ Group*Time
+(1|Pupil) + (1|Class) + (1|Question),
data = dataset, family = binomial, control
= glmerControl(optimizer = "bobyqa"))
```

TABLE 2
Descriptive Statistics for Outcome Measures

Test	Group	*n*	Pretest *M*(*SD*)	Posttest *M*(*SD*)	Delayed *M*(*SD*)
Sentence–picture matching (/8)	7-day	38	4.68 (1.35)	4.68 (1.60)	4.53 (1.31)
	3.5-day	41	3.95 (1.36)	4.88 (1.87)	4.27 (1.55)
	Control	34	4.50 (1.29)	4.44 (1.62)	4.65 (1.45)
AJT grammatical (/6)	7-day	20	3.16 (1.46)	2.75 (2.00)	2.70 (1.72)
	3.5-day	26	2.95 (1.28)	3.19 (1.79)	3.23 (1.63)
	Control	16	3.56 (1.15)	3.13 (1.71)	2.25 (1.88)
AJT ungrammatical (/6)	7-day	20	0.51 (0.67)	1.05 (1.00)	0.70 (0.86)
	3.5-day	26	0.83 (0.51)	1.12 (1.21)	0.73 (1.12)
	Control	16	0.56 (0.73)	0.38 (0.62)	0.38 (0.62)

Note. AJT = acceptability judgement test; *M* = mean; *SD* = standard deviation.

RESULTS

Influence of Practice Distribution (RQ1)

Sentence–Picture Matching Test. Table 2 details the descriptive statistics for performance on the matching test at pre-, post-, and delayed posttest.[4] Examination of the descriptive statistics indicated minimal changes in group-level mean scores over time and minimal differences between groups, with two notable exceptions: First, the 3.5-day group's performance at pretest was lower than both the 7-day and control groups, with these differences representing small effects (Table 3a). Although the group differences at baseline were generally unreliable as their 95% CIs pass through zero, one effect (3.5-day vs. 7-day) had a reliable, albeit small, effect. The effect of group at pretest was approaching significance, Kruskal–Wallis $\chi^2(2) = 5.396$, $p = .067$. Second, there was an increase in the 3.5-day group's scores between pre- and posttest (Table 3b), with a small within-group effect size whose CIs did not cross zero, indicating a reliable effect.

Analysis, via the Anova() function (Type III) in the car package in R, of the fixed effects within the model of the matching test data revealed no main effects for group, $\chi^2(2) = 4.538$, $p = .103$, or time, $\chi^2(2) = 0.272$, $p = .873$, nor any interaction between group and time, $\chi^2(4) = 5.060$, $p = .281$. Nevertheless, a marginal fixed effect for the interaction between the 3.5-day group at posttest in comparison to the control group was observed (estimate = 0.46, $SE = 0.236$, $z = 1.933$, $p = .053$), reflecting the change in the 3.5-day group's scores from below to above the control group's scores between pre- and posttest. These results mirrored the observations made based on the descriptive statistics, reflecting minimal changes in group-level performance over time and between groups.

Given the difference observed in group scores at pretest, the model was rerun with pretest as a control variable, rather than as part of the independent variable time. However, no significant main effect for pretest was observed, $\chi^2(1) = 1.371$, $p = .242$, suggesting that pretest performance was not an indicator of the groups' performance on the matching test at subsequent time points. Pretest was therefore not included as a control variable in subsequent models.

Acceptability Judgement Test: Grammatical Items. Table 2 details the descriptive statistics for AJT G items at pre-, post- and delayed posttest. These indicate minimal group-level change over time for the 7-day and 3.5-day groups, but a decrease in scores for the control group most notable between pre- and delayed posttest (Table 3b). Further, there was a small difference in the 3.5-day and control groups' performance at pretest (control > 3.5-day) and at delayed posttest (3.5-day > control), although the CIs for both effect sizes cross zero, suggesting that this effect is not reliable (Table 3a).

Analysis of the fixed effects within the model revealed no significant effect of group, $\chi^2(2) = 0.848$, $p = .655$, and no significant interaction between group and time, $\chi^2(4) = 7.611$, $p = .107$; however, a significant fixed effect for time was revealed, $\chi^2(2) = 10.459$, $p = .005$. This effect was qualified by a significant fixed effect for the interaction between the 3.5-day and control groups' scores at delayed posttest (estimate = 1.076, $SE = 0.408$, $z = 2.640$, $p = .008$); reflecting the decrease observed in the control group's scores at delayed posttest, whilst the 3.5-day group maintained their scores.

TABLE 3a
Effect Sizes Comparing Scores Between Groups

Test	Group	Pretest *d* (CIs)	Posttest *d* (CIs)	Delayed Posttest *d* (CIs)
Sentence–picture matching (/8)	7-day / 3.5-day	0.54 (0.08, 0.98)[a]	−0.11 (−0.56, 0.33)	0.18 (−0.26, 0.62)
	7-day / control	0.14 (−0.33, 0.60)	0.15 (−0.32, 0.61)	−0.09 (−0.55, 0.38)
	3.5-day / control	−0.41 (−0.87, 0.05)	0.25 (−0.21, 0.70)	−0.25 (−0.71, 0.21)
AJT grammatical (/6)	7-day / 3.5-day	0.15 (−0.43, 0.74)	−0.23 (−0.81, 0.36)	−0.31 (−0.90, 0.28)
	7-day / control	−0.30 (−0.95, 0.37)	−0.20 (−0.86, 0.46)	0.25 (−0.41, 0.91)
	3.5-day / control	−0.49 (−1.12, 0.15)	0.03 (−0.59, 0.66)	0.57 (−0.08, 1.19)
AJT ungrammatical (/6)	7-day / 3.5-day	−0.55 (−1.13, 0.06)	−0.06 (−0.64, 0.52)	−0.03 (−0.61, 0.55)
	7-day / control	−0.07 (−0.73, 0.59)	0.79 (0.09, 1.45)[a]	0.42 (−0.25, 1.07)
	3.5-day / control	0.45 (−0.19, 1.07)	0.72 (0.06, 1.35)[a]	0.36 (−0.27, 0.98)

Note. AJT = acceptability judgement test; *d* = Cohen's *d* effect size; CIs = confidence intervals.
[a]CIs do not pass through zero.

TABLE 3b
Within-Group Effect Sizes Comparing Scores Between Time Points

Test	Group	Pretest vs. Posttest *d* (CIs)	Posttest vs. Delayed Posttest *d* (CIs)	Pretest vs. Delayed Posttest *d* (CIs)
Sentence–picture matching (/8)	7-day	0.00 (−0.45, 0.45)	−0.10 (−0.55, 0.35)	−0.11 (−0.56, 0.34)
	3.5-day	0.57 (0.12, 1.00)[a]	−0.36 (−0.79, 0.08)	0.22 (−0.22, 0.65)
	Control	−0.04 (−0.52, 0.43)	0.14 (−0.34, 0.61)	0.11 (−0.37, 0.58)
AJT grammatical (/6)	7-day	−0.23 (−0.85, 0.39)	−0.03 (−0.65, 0.59)	−0.29 (−0.91, 0.34)
	3.5-day	0.15 (−0.39, 0.70)	0.02 (−0.52, 0.57)	0.19 (−0.36, 0.73)
	Control	−0.30 (−0.98, 0.41)	−0.49 (−1.18, 0.23)	−0.84 (−1.54, −0.10)[a]
AJT ungrammatical (/6)	7-day	0.63 (−0.01, 1.26)	−0.38 (−0.99, 0.26)	0.25 (−0.38, 0.86)
	3.5-day	0.31 (−0.24, 0.85)	−0.33 (−0.88, 0.22)	−0.11 (−0.66, 0.43)
	Control	−0.27 (−0.95, 0.44)	0.00 (−0.69, 0.69)	−0.27 (−0.95, 0.44)

Note. AJT = acceptability judgement test; *d* = Cohen's *d* effect size; CIs = confidence intervals.
[a]CIs do not pass through zero.

Acceptability Judgement Test: Ungrammatical Items. The descriptive statistics for the AJT UG items revealed low scores on these items across all groups (Table 2). Nevertheless, the effect sizes for between-group comparisons (Table 3a) indicated that both the 7-day and 3.5-day groups scored higher than the control group at posttest and at delayed posttest, although the CIs for the delayed posttest effect sizes crossed zero, indicating less certainty in this effect.

The model of scores on the AJT UG items revealed no significant fixed effect of group, $\chi^2(2) = 1.839$, $p = .399$; time, $\chi^2(2) = 1.020$, $p = .601$; or interaction between group and time, $\chi^2(4) = 3.803$, $p = .433$. Nevertheless, a marginal fixed effect for the 7-day group in comparison to the control group at posttest was observed (estimate = 1.324, $SE = 0.695$, $z = 1.906$, $p = .057$). This reflected the small increase in the 7-day group's scores between pre- and posttest compared to the lower performance of the control group.

Influence of Practice Accuracy (RQ1a)

Analysis was conducted to explore the association between the accuracy of learners' performance during training and subsequent performance at post- and delayed posttest. As the control group did not complete the training activities, it is excluded from this analysis.

Practice Accuracy. The learners' global practice accuracy score (i.e., percentage of questions answered correctly out of all those attempted across the training sessions) provided an indication of how successfully the learners completed

the training.[5] The global practice accuracy scores were high for both the 7-day ($n = 38$, $M = 79.6\%$, CI [76.7%, 82.4%], $SD = 8.7\%$) and 3.5-day groups ($n = 41$, $M = 82.5\%$, CI [79.9%, 85.2%], $SD = 8.4\%$), with both groups answering more than 75% of practice items correctly on average. The standard deviation and CI around each mean indicate some variation between individuals' practice scores. The minimum score from any individual within the 7-day group was 62.8% and within the 3.5-day group was 55.4%. An independent samples *t*-test indicated no significant difference in global practice accuracy between the two groups, $t(77) = -1.522$, $p = .132$, $d = -0.34$, CI [−0.78, 0.11].

To examine the relationship between performance during training sessions (i.e., practice accuracy) and performance on the outcome measures, the models of learners' performance on each outcome measure were expanded to include practice accuracy (with scores centred on the grand mean to avoid multicollinearity) as a predictor variable:

```
PracticeAccuracy_model < −glmer(Score ∼
Group*Time + PracticeAccuracy + (1|Pupil)
+(1|Class) + (1|Question), data = dataset,
family = binomial, control
= glmerControl(optimizer = "bobyqa"))
```

Sentence–Picture Matching Test. Including practice accuracy as a predictor variable within the model yielded a significant effect of group, $\chi^2(1) = 5.341$, $p = .021$, qualified by a fixed effect for the 3.5-day group in comparison to the 7-day group (estimate = −0.378, $SE = 0.164$, $z = -2.311$, $p = .021$), reflecting the change in the 3.5-day groups' scores between pre- and posttest. Further, a significant effect of practice accuracy was observed, $\chi^2(1) = 11.039$, $p < .001$, indicating that learners' overall success at completing the training activities predicted performance on the matching test.

A small–medium, significant association with practice accuracy was observed for both groups at posttest and for the 3.5-day group at delayed posttest (Table 4). For the 7-day group at delayed posttest, the association had weakened slightly and the lower CI bound just crossed zero, suggesting a marginally reliable small association.

Acceptability Judgement Test: Grammatical Items. The expanded model of performance on the AJT G items yielded a significant effect of practice accuracy, $\chi^2(1) = 4.026$, $p = .045$, but no significant effect of group, $\chi^2(1) = 0.074$, $p = .786$; or time, $\chi^2(2) = 1.675$, $p = .433$.

Correlations (see Table 4) indicated medium, significant associations with the 3.5-day groups' practice accuracy at post- and delayed posttest. In contrast, for the 7-day group, the association was not reliable or statistically significant at posttest or delayed posttest. The findings suggest that practice accuracy was a significant predictor of posttest performance on this test for the 3.5-day group only.

Acceptability Judgement Test: Ungrammatical Items. For AJT UG items, the expanded model yielded a significant effect of practice accuracy, $\chi^2(1) = 17.441$, $p < .001$, but no effect of group, $\chi^2(1) = 0.635$, $p = .426$; or time, $\chi^2(1) = 4.071$, $p = .131$. A medium–large, significant association between practice accuracy and performance on AJT UG items at posttest was observed for both the 7-day and 3.5-day groups (Table 4). At delayed posttest, this association remained reliable and statistically significant for the 3.5-day group, but weakened considerably for the 7-day group and was no longer reliable or statistically significant.

Influence of Language Analytic Ability (RQ1b)

Although little change was seen in group-level mean scores, the standard deviations (Table 2) suggest substantial within-group variation in performance on each test. We now examine whether this variation can be accounted for by individual differences in LAA.

Language Analytic Ability Test Performance. The descriptive statistics for the three groups' performance on the LAA test are presented in Table 5. The 3.5-day group's performance was marginally higher than both the 7-day ($d = -0.46$, CI [−0.90, −0.01]) and control group ($d = 0.36$, CI [−0.10, 0.81]), although a Kruskal–Wallis test revealed no significant effect of group, $\chi^2(2) = 3.607$, $p = .165$. The CIs around the mean overlapped between all three groups, indicating that performance of all three groups fell within a similar range. Notably, the large standard deviations indicate a large amount of within-group variation on this test (Table 5).

To examine the impact of LAA on performance on the outcome measures, the mixed effects logistic models were expanded to include LAA scores (centred around the grand mean) as a predictor variable[6]:

```
LAA_model < − glmer(Score ∼ Group*Time
+ LAA + (1|Pupil) + (1|Class)
+ (1|Question), data = dataset, family
= binomial, control = glmerControl(optimizer
= "bobyqa"))
```

TABLE 4
Correlation Between Practice Accuracy and Outcome Measures

			Posttest		Delayed Posttest	
Test	Group	*n*	*rho* (CIs)	*p*	*rho* (CIs)	*p*
Sentence–picture matching	7-day	38	.38* (.08, .61)	.018	.32 (−.01, .60)	.052
	3.5-day	41	.36* (.10, .59)[a]	.020	.38* (.11, .59)[a]	.016
	All	79	.38* (.21, .54)[a]	.001	.33* (.12, .52)[a]	.003
AJT grammatical	7-day	20	.19 (−.37, .63)	.429	.16 (−.31, .61)	.508
	3.5-day	26	.41* (−.02, .78)	.040	.51* (.11, .81)[a]	.009
	All	46	.36* (.04, .61)[a]	.014	.37* (.08, .60)[a]	.013
AJT ungrammatical	7-day	20	.53* (.09, .85)[a]	.016	.16 (−.48, .64)	.508
	3.5-day	26	.64* (.34, .83)[a]	.001	.49* (.17, .74)[a]	.010
	All	46	.60* (.34, .77)[a]	.001	.31* (−.02, .59)	.038

Note. AJT = acceptability judgement test; *rho* = Spearman's *rho* correlation coefficient; CIs = confidence intervals.
[a]CIs do not pass through zero.
*Significant at .05 level.

TABLE 5
Performance on Language Analytic Ability Test

Group	*n*	*M* (*SD*)	CIs
7-day	38	26.7 (9.8)	23.4, 29.9
3.5-day	41	30.8 (8.4)	28.2, 33.5
Control	34	27.7 (9.2)	24.5, 30.9

Note. *M* = mean; *SD* = standard deviation; CIs = confidence intervals.

Sentence–Picture Matching Test. The expanded model yielded a significant effect of group, $\chi^2(2) = 6.028$, $p = .049$, which was qualified by a significant fixed effect for the 3.5-day group in comparison to the control group (estimate = −0.329, $SE = 0.167$, $z = -1.967$, $p = .049$). No significant effect of time, $\chi^2(2) = 0.273$, $p = .872$, or interaction between group and time was observed in the expanded model, $\chi^2(4) = 5.020$, $p = .285$. However, a significant effect of LAA was observed, $\chi^2(1) = 5.924$, $p = .015$. Further, comparison, via the Anova() function in R, of the original and expanded models for the matching test revealed that including LAA significantly improved the model fit, $\chi^2(1) = 5.819$, $p = .016$, indicating that learners' performance on the LAA test was a significant predictor of performance on the matching test.

Spearman's *rho* indicated a small but nonstatistically significant association at posttest for both the 7-day and 3.5-day groups and no association for the control group (Table 6). At delayed posttest, a stronger association was observed for the 3.5-day group, which was borderline statistically significant and had 95% CIs for *rho* that only just passed through zero. The association between LAA and matching test performance for the control group at delayed posttest was also stronger with a similar pattern of marginal reliability and significance.

Acceptability Judgement Test: Grammatical Items. The expanded model for the AJT G items revealed a significant effect of time, $\chi^2(2) = 10.460$, $p = .005$, which was qualified by a significant fixed effect of delayed posttest compared to pretest (estimate = −0.997, $SE = 0.314$, $z = -3.178$, $p = .001$) and a significant fixed effect of 3.5-day group compared to control group at delayed posttest (estimate = 1.085, $SE = 0.408$, $z = -2.661$, $p = .008$), reflecting the decrease in the control group's scores. Within the expanded model, there was also a significant effect of LAA, $\chi^2(1) = 4.163$, $p = .041$, and the addition of LAA significantly improved the model fit, $\chi^2(1) = 4.052$, $p = .044$.

Correlations between LAA and AJT G scores revealed significant medium associations for the 3.5-day group at post- and delayed posttest (Table 6). In contrast, for the 7-day and control groups at both posttest and delayed posttest, associations were unreliable and nonstatistically significant.

Acceptability Judgement Test: Ungrammatical Items. For the AJT UG items, the expanded model

TABLE 6
Correlations Between Language Analytic Ability and Outcome Measures

			Posttest		Delayed	
Test	Group	*n*	*rho* (CIs)	*p*	*rho* (CIs)	*p*
Sentence–picture matching	7-day	38	.22 (−.09, .51)	.186	.15 (−.21, .47)	.359
	3.5-day	41	.24 (−.06, .50)	.125	.30 (−.04, .61)	.054
	Control	34	−.02 (−.39, .32)	.906	.29 (−.01, .54)	.087
	All	113	.17 (−.02, .34)	.065	.24* (.07, .41)[a]	.011
AJT grammatical	7-day	20	.18 (−.38, .71)	.447	.10 (−.37, .59)	.691
	3.5-day	26	.48* (.10, .75)[a]	.012	.54* (.19, .81)[a]	.004
	Control	16	.29 (−.24, .73)	.264	−.09 (−.64, .50)	.729
	All	62	.34* (.09, .54)[a]	.007	.22 (−.05, .46)	.080
AJT ungrammatical	7-day	20	.42 (−.07, .74)	.068	.33 (−.14, .73)	.153
	3.5-day	26	.65* (.37, .84)[a]	.001	.28 (−.23, .71)	.167
	Control	16	−.10 (−.54, .42)	.726	.05 (−.65, .54)	.868
	All	62	.40* (.13, .61)[a]	.001	.24 (−.04, .50)	.057

Note. AJT = acceptability judgement test; *rho* = Spearman's *rho* correlation coefficient; CIs = confidence intervals.
[a]CIs do not pass through zero.
*Significant at .05 level.

yielded a significant effect for LAA, $\chi^2(1) = 18.425$, $p < .001$. No significant effect for group, $\chi^2(2) = 1.085$, $p = .581$; time, $\chi^2(2) = 0.978$, $p = .613$; or interaction between group and time, $\chi^2(4) = 3.711$, $p = .447$, was observed within the expanded model, mirroring the findings of the original model; however, the expanded model significantly improved the model fit, $\chi^2(1) = 16.203$, $p < .001$.

For the AJT UG items, at posttest, a large, reliable and statistically significant association was observed with LAA for the 3.5-day group and a medium association for the 7-day group, but this was nonstatistically significant and had borderline reliability as the CIs just passed through zero (Table 6). By delayed posttest, the correlations weakened to small nonsignificant and unreliable associations for both groups. No associations were observed for the control group at post- or delayed posttest. These findings indicate that for both the 7-day and 3.5-day groups, LAA had some relation with learners' performance on the AJT UG items at posttest.

DISCUSSION

This study investigated the impact of practice distribution on grammar learning and explored the extent to which accuracy during training and LAA moderated learning under shorter and longer practice schedules. Our sentence–picture matching test had low reliability, whilst reliability for the AJT G and UG items was acceptable. We were unable to obtain indices for three out of the nine test administrations and both tests had a low number of items, thus we interpret our results with caution whilst also noting that our mixed-model analyses did account for random variation between test items.

Group-Level Performance Compared to Control

Before discussing the findings in relation to the impact of practice distribution, it is necessary to address the (somewhat surprising) finding that, at group-level, minimal differences were observed in the treatment groups' performances on the outcome measures compared to the control group and, further, that minimal changes over time were observed between time points for all groups. The group-level statistics could suggest minimal learning as a result of the intervention, potentially contrary to much of the existing research on form–meaning mapping (a component of the wider Processing Instruction approach; see DeKeyser & Botana, 2015; and Shintani, 2015 for reviews). For example, Shintani (2015) found a large overall effect ($d = 2.60$) for Processing Instruction on receptive knowledge at posttest in a meta-analysis of 42 Processing Instruction studies. Additionally, Marsden (2006) used a similar approach to teach a slightly larger grammatical subsystem (inflectional verb morphology for person, number, tense) and found clear learning gains on a battery of measures. However, a number of considerations should be taken into account when interpreting the current findings.

First, in Marsden's (2006) study, the intervention was considerably longer (4.5 hours over

9 weeks) and the learners had experienced on average 200 more hours of French instruction than the learners in the current study. Their slightly higher proficiency likely provided them with a larger and more stable verb lexicon on which to graft an (already) emerging inflectional system (see Marsden & David, 2008, who showed that the size of the verb lexicon correlated positively with inflectional diversity). Also, perhaps critically, the learners were older (aged 13–14 years), thus potentially more able to draw on their explicit inductive and deductive learning mechanisms. Relatedly, it is also possible that, for (at least some of) the young learners in the present study, the intervention may have been too brief, as explicit learning by young learners is slower than for more cognitively mature, older learners (Lichtman, 2016). Indeed, Kasprowicz and Marsden (2018) observed substantial learning gains following form–meaning mapping practice for 9- to 11-year olds, but after a longer (250 minutes) intervention over 5 weeks, which focussed on one grammatical function (subject or object assignment via German definite articles).

Second, there was a high level of within-group variation in the learners' global practice accuracy scores on the training and in their performance on the outcome measures. This indicated differential benefits of the intervention for individual learners and variation between individuals' success completing the tests. This is discussed further in light of the findings of the LAA analysis.

Third, the analysis of the learners' performance in the training sessions revealed a high level of accuracy during the practice activities in both groups, indicating that the learners were attending to and correctly applying the grammatical rules. It would seem, then, that the 7-day and 3.5-day learners' performance at the posttests (at least at a group level) did not reflect the knowledge being developed during the training sessions. One possible explanation for this discrepancy may be differences between the training and testing activities. Transfer-appropriate processing accounts predict greater success at retrieving previously learnt information when the learning and testing tasks draw on similar processes, skills, and contexts (Lightbown, 2008; Segalowitz, 2003; Spada & Lightbown, 2008). For example, more isolated (decontextualized) instruction may lead to greater gains on explicit, discrete tests, compared to integrated (contextualized) learning activities favouring more communicative tests (Martin–Chang, Levy, & O'Neil, 2007; Spada, Jessop, & Tomita, 2014). Training in the present study involved practice embedded within game-based environments and required repeatedly connecting one inflection, from one particular pair, to a meaning or function in an engaging visual (e.g., robots choosing food in a cafeteria). In contrast, the sentence–picture matching test and AJT required different processes. For example, both tests were in the written modality, in contrast to the training, which had been in both aural and written modalities for each inflection. The matching test required recognition of isolated exemplars, in a decontextualized environment, with two pictures to choose from, and tested all of the target inflections over a small number of items. The AJT G items required learners to recognize correctness and the UG items required them to know that particular features were not grammatical, both knowledge and skills that had not specifically been practiced during the game. Whilst both practice and tests constituted input-based, comprehension activities, the differences in the contexts and actions required may account for why (some of) the learners did not reliably apply knowledge established during the training to the tests.

There are several possible accounts for this, drawing on skill acquisition theory. One is that the learners may not have engaged in transfer-appropriate processing during training and had not acquired representations of the grammatical features that were sufficiently generalizable to the tests. Successfully applying knowledge across different task conditions requires the establishment of relevant and reliable declarative knowledge (as declarative knowledge is transferable to other task conditions and characteristics), yet some learners may not have established either fully accurate or sufficiently robust declarative knowledge for reliable transfer to occur, so recall remained error prone, as is typical of the early stages of skill acquisition. An alternative, though related, explanation might be that during practice, learners established proceduralized and even automatized knowledge of the inflections that was relevant to the game context, but as proceduralized and automatized knowledge is known to be highly specific, this was perhaps not adaptable to the test contexts.

Providing (more) varied practice opportunities may be one way of helping learners consolidate the necessary declarative, proceduralized, and automatized knowledge that is transferable across a wider range of task conditions than found in the current study. This could be built into the current game, as computer delivery provides opportunities to tailor the amount and nature of practice at an individual level (DeKeyser, 2012).

Impact of Practice Distribution

Our analyses of lag effects did not yield convincing evidence that our different practice distributions affected learners' group performance differentially, at least on the outcome measures utilized here. On the sentence–picture matching test, a small advantage was observed for the 3.5-day group, with group-level improvement between Pre- and Posttest 1 compared to minimal group-level change in the 7-day group's (and control group's) scores over time. Note that the 7-day group's pretest score was higher than that of the 3.5-day group, and therefore the 3.5-day group's gains brought them to a similar level to the 7-day group at posttest. The advantage of the 3.5-day spacing was not maintained at delayed posttest. On the AJT test, no practice distribution effect was observed; neither group showed significant group-level change over time (although there was a small increase in both groups' scores at posttest on the UG items and for the 7-day group, this was a small effect).

Accuracy during training predicted performance for both groups, on the matching and AJT UG items at posttest. Spearman's *rho* indicated that the effects were smaller for both groups at delayed posttest, perhaps due to decay of declarative knowledge developed during the training (Suzuki & DeKeyser, 2017a), though a small effect remained for the 3.5-day group. For the AJT G items, accuracy during training was related to posttest performance for the 3.5-day group only. In sum, we observed no clear advantage in knowledge retention for spacing of 3.5 or 7 days on our tests, despite some tentative benefits for the 3.5-day group on a number of findings.

Our tentative finding of some advantage for the 3.5-day group (most clearly, pre- to posttest gains on the sentence–picture matching test) could align with findings by Suzuki and DeKeyser (2017a) and Suzuki (2017), who both observed benefits for spacing that was shorter than 7 days (1 day and 3.3 days respectively), also with beginner learners and focusing on morphology (see also Toppino & Bloom, 2002, for an account of why longer spacing may lead to more forgetting). However, this finding is contrary to Bird (2010) and Rogers (2015), who found advantages for spacing of 7 days or more with intermediate learners. However, when interpreting the relationship between our findings and previous studies we must bear in mind the methodological differences between these studies and our study (see Appendix A). In particular, our study compared less frequent, longer sessions (7-day ISI, 3×60 minutes) to more frequent, shorter sessions (3.5-day ISI, 6×30 minutes), both distributed over the same period of time (3 weeks); a comparison that is perhaps more reflective of the decisions that teachers have to make regarding how to distribute the curriculum within allocated teaching time. These differences point to the general need for increased replication in our field (as illustrated by Marsden et al., 2018).

Another issue to consider in interpreting the lack of clearer lag effects and the lack of overall gains over time is the mixed findings regarding our instrument reliability (with some data missing and the sentence–picture matching test index falling below the recommended level of acceptability). This could indicate that the tests may not have been able to show robust change over time. However, this concern is mitigated by the fact that sufficient variance and change over time in the scores was observed for both accuracy during training and LAA to be significant predictors of learning, the latter finding to which we now turn.

Impact of Language Analytic Ability

LAA improved the fit of our mixed-effects models and was a significant predictor for both outcome measures, suggesting that LAA significantly influenced learners' test scores. This finding is consistent with that of Tellier and Roehr–Brackin (2013), who observed that LAA significantly predicted learning for young learners in a classroom context similar to that of the present study. Additionally, the correlations observed between the outcome measures and LAA for the treatment groups (particularly the 3.5-day group and the AJT G and UG items) were similar to the overall association between aptitude and L2 grammar learning ($r = .31$) observed in Li's (2015) meta-analysis. We found no significant correlations between LAA and pretest scores (see Appendix F).

Considering the explicit nature of the training, which potentially drew on all three constructs elicited by our LAA test (grammatical sensitivity, and deductive and inductive analytic abilities), learners with higher LAA probably better understood the explicit information provided or were more efficient at identifying and applying rules during practice, which in turn led to improved posttest scores, particularly at Posttest 1. This observation aligns with previous studies observing strong relationships between LAA and learning under explicit instruction (e.g., Erlam, 2005; Li, 2015; Robinson, 1997). Further, components of aptitude, such as LAA, have been

argued to become increasingly important as tasks increase in complexity and place a higher cognitive burden on the learner (Suzuki & DeKeyser, 2017b). In line with this argument, the significant associations with posttest performance may in part reflect the complexity inherent in transferring knowledge between training and tests, as discussed previously.

Two additional observations merit discussion. First, the correlations with LAA were strongest for the AJT, at least for the 3.5-day group. This is likely due to similarities between the two tests, as both elicited learners' explicit ability to spot (violations in) patterns. This observation aligns with Granena's (2013) finding that demonstrated a relationship between aptitude tests that draw on more explicit processes and language tasks that encourage analysis of language form.

The second important observation is that LAA was associated more strongly with outcomes for the 3.5-day group than for the 7-day group. Recall that Suzuki and DeKeyser's (2017b) adult learners showed an association between LAA under their longer ISI (7 days), not their shorter ISI (1 day). They argued that learners with higher LAA developed a more accurate and reliable initial understanding of the structures, resulting in more successful retrieval after longer spacing, whereas high LAA was not as important (useful) when practice was repeated just a day later. In contrast, in the present study, stronger associations were, overall, observed between LAA and learning under the shorter distribution. But our shorter distribution was 3.5 days, rather than Suzuki & DeKeyser's 1 day, and our learners were much younger. It is possible that for our learners, the 3.5-day interval showed differential sensitivity to LAA as learning was susceptible, at some level, to the capacity to establish accurate and robust knowledge in the first place, whereas the 7-day interval may have washed out any such sensitivity due to its overall heavier demands on recall. These suggestions are speculative, and further research is needed into how individual differences such as age and, related to age, working memory capacity may interact with distribution of practice effects.

Limitations

Although our study had high ecological validity, this inevitably came with some costs due to practical constraints of carrying out classroom studies, such as participant attrition (with fewer participants completing the AJT test) and the use of intact classes rather than randomization at the individual level. It is also important to acknowledge the brevity of the intervention, which may in part account for the minimal, group-level learning gains. We also note, however, that 180 minutes of instruction focused solely on a subset of inflectional verb morphology over 3 weeks is greater than is likely to occur within the time-limited FL primary school context. Two other limitations of the study are that we have only examined comprehension-based tests (not production) and our sentence–picture matching test had low internal reliability.

CONCLUSION

This study investigated distribution of practice effects for learning L2 French inflectional morphology, extending previous research by investigating younger learners in ecologically valid FL classrooms. Results showed minimal differences between performance under shorter (3.5-day) and longer (7-day) practice schedules on the outcome measures utilized in this study but provided tentative evidence that shorter spacing may have been slightly more helpful for these young learners. Furthermore, the results indicated that learning under both practice schedules was moderated by individuals' training success (i.e., practice accuracy) across both groups and by LAA particularly for the 3.5-day condition. This underlines the importance of considering individual learner differences in the development of instructional materials and the potential benefits of utilizing adaptive digital tools.

ACKNOWLEDGMENTS

This research was supported through the Engineering and Physical Sciences Research Council-funded Digital Creativity Labs at the University of York (Grant number: EP/M023265/1). We are grateful to the teachers and pupils who participated in this study and to Dr. Abigail Parrish for her research assistance. Our thanks also go to Professor Peter Cowling, Lynn Yun, Kacper Sagnowski, and Dr. Sebastian Deterding for their support in developing the digital grammar learning game utilized in the study's intervention. We would also like to thank the guest editors and two anonymous reviewers for their valuable comments.

NOTES

[1] As ordinal omega hierarchical reliability indices could not be provided for the sentence–picture

matching posttest, AJT G or AJT UG pretest data, here we provide the corresponding indices that were returned: sentence–picture matching posttest, ordinal Cronbach's alpha = .46; AJT G pretest, Cronbach's alpha = .29; AJT UG pretest, ordinal Cronbach's alpha = .39. However, we strongly emphasize that these should be treated with caution, as the most suitable reliability index for our data is ordinal omega hierarchical.

[2] Confidence intervals for Cohen's *d* were calculated using an effect size calculator (https://www.cem.org/effect-size-calculator; accessed July 2018)

[3] Correlations were calculated using learners' raw scores at post- and delayed posttest. We also calculated correlations using gains scores on each of the outcome measures to account for baseline differences and found a similar pattern of results.

[4] As the sentence–picture matching test was a two-way multiple-choice test, a single-sample *t*-test was run to compare the learners' scores at each time point to a 50% chance-level score. A significant difference compared to chance was observed at all time points, pretest: $t(112) = 2.820$, $p = .006$; posttest: $t(112) = 4.255$, $p = .001$; delayed posttest: $t(112) = 3.463$, $p = .001$.

[5] Detailed presentation of the learners' performance within each training session is beyond the scope of this article.

[6] The mixed-effects logistic models including LAA did not include practice accuracy as an additional control variable so as to enable inclusion of control group data.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*, 481–509.

Australia Curriculum, Assessment and Reporting Authority. (2011). *The shape of the Australian curriculum: Languages*. Accessed 17 May 2019 at http://docs.acara.edu.au/resources/Languages_-_Shape_of_the_Australian_Curriculum_new.pdf

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*, 228–247.

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

Bjork, E. L., & Bjork, R. A. (2014). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher & J. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 59–68). New York: Worth.

Bush, M. E. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education, 40*, 218–231.

Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. S. Parry & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–29). Upper Saddle River, NJ: Prentice Hall.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning a temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102.

Collins, L., Halter, R. H., Lightbown, P., & Spada, N. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly, 33*, 655–680.

Collins, L., & White, J. (2011). An intensive look at intensity and language learning. *TESOL Quarterly, 45*, 106–133.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499–533.

DeKeyser, R. (2007). Conclusion: The future of practice. In R. DeKeyser (Ed.), *Practice in a second language* (pp. 287–304). Cambridge: Cambridge University Press.

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

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94–112). London: Routledge.

DeKeyser, R., & Botana, G. P. (2015). The effectiveness of processing instruction in L2 grammar acquisition: A narrative review. *Applied Linguistics, 36*, 290–305.

Department for Education. (2013). *Languages programmes of study: Key stage 2. National curriculum in England*. London: Crown Copyright.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*, 795–805.

Doughty, C. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. Doughty & M. Long (Ed.), *The handbook of second language acquisition* (pp. 256–310). Oxford, UK: Blackwell.

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Ellis, N. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*, 164–194.

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, *40*, 83–107.

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9*, 147–171.

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Journal of Educational Psychology*, *59*, 290–296.

Granena, G. (2013). Language aptitude and long-term achievement in early childhood L2 learners. *Applied Linguistics*, *35*, 483–503.

Hanan, R. E. (2015). *The effectiveness of explicit grammar instruction for the young foreign language learner: A classroom-based experimental study* (Unpublished doctoral dissertation). University of York, Heslington, York, UK.

Harley, B., & Hart, D. (1998). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition*, *19*, 379–400.

Kasprowicz, R. E., & Marsden, E. (2018). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form–meaning practice. *Applied Linguistics*, *39*, 886–911.

Kiss, C., & Nikolov, M. (2005). Developing, piloting, and validating an instrument to measure young learners' aptitude. *Language Learning*, *55*, 99–150.

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.

Lichtman, K. (2016). Age and learning environment: Are children implicit second language learners? *Journal of Child Language*, *43*, 707–730.

Lightbown, P. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27–44). Clevedon, UK: Multilingual Matters.

Lotfolahi, A. R., & Salehi, H. (2016). Learners' perceptions of the effectiveness of spaced learning schedule in L2 vocabulary learning. *SAGE Open*, *6*, 1–9.

Marsden, E. (2006). Exploring input processing in the classroom: An experimental comparison of processing instruction and enriched input. *Language Learning*, *56*, 507–566.

Marsden, E., & Chen, H.–Y. (2011). The roles of structured input activities in processing instruction and the kinds of knowledge they promote. *Language Learning*, *61*, 1058–1098.

Marsden, E., & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal*, *36*, 181–198.

Marsden, E., Morgan–Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning*, *68*, 321–391.

Martin–Chang, S. L., Levy, B. A., & O'Neil, S. (2007). Word acquisition, retention, and transfer: Findings from contextual and isolated word training. *Journal of Experimental Child Psychology*, *96*, 37–56.

McManus, K., & Marsden, E. (2017). L1 explicit instruction can improve L2 online and offline performance. *Studies in Second Language Acquisition*, *39*, 459–492.

McManus, K., & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*, *40*, 459–475.

McManus, K., & Marsden, E. (2019a). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*, 205–234.

McManus, K., & Marsden, E. (2019b). Using L1 explicit instruction to reduce crosslinguistic effects in L2 grammar learning. *Modern Language Journal*, *103*, 459–480.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*, 578–596.

Muñoz, C. (2014). The association between aptitude components and language skills in young learners. In M. Pawlak & L. Aronin (Eds.), *Essential topics in applied linguistics and multilingualism: Studies in honour of David Singleton* (pp. 51–68). Cham, Switzerland: Springer.

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

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

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. 160–181). Philadelphia/Amsterdam: John Benjamins.

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

Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics*, *29*, 173–199.

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*, 857–866.

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*, 906–911.

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude

tests. *Journal of the European Second Language Association*, *1*, 49–60.

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*, 183–186.

Segalowitz, N. (2003). Automaticity and second languages. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Oxford, UK: Blackwell.

Serrano, R., & Huang, H.–Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, *52*, 971–994.

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

Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, *36*, 306–325.

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. 69–96). Philadelphia/Amsterdam: John Benjamins.

Spada, N., Jessop, L., & Tomita, Y. (2014). Isolated and integrated form-focused instruction: Effects of different types of L2 knowledge. *Language Teaching Research*, *18*, 453–473.

Spada, N., & Lightbown, P. (2008). Form-focused instruction: Isolated or integrated? *TESOL Quarterly*, *42*, 181–207.

Standards and Testing Agency. (2014, 2015, 2016). *Key stage 2: Grammar, punctuation and spelling (Levels 3–5)*. London: Crown Copyright.

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning*, *67*, 512–545.

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

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

Swanson, P., & Mason, S. (2018). The world language teaching shortage: Taking a new direction. *Foreign Language Annals*, *51*, 251–262.

Tellier, A. (2013). Developing a measure of metalinguistic awareness for children aged 8–11. In K. Roehr & G. A. Gánem–Gutiérrez (Eds.), *The metalinguistic dimension in instructed second language learning* (pp. 15–43). London: Bloomsbury.

Tellier, A., & Roehr–Brackin, K. (2013). The development of language learning aptitude and metalinguistic awareness in primary-school children: A classroom study. *Essex Research Reports in Linguistics*, *62*, 1–28.

Tellier, A., & Roehr–Brackin, K. (2017). Raising children's metalinguistic awareness to enhance classroom second language learning. In M. d. P. G. Mayo (Ed.), *Learning foreign languages in primary school: Research insights* (pp. 22–48). Bristol, UK: Multilingual Matters.

Tinsley, T., & Board, K. (2017). *Language trends 2016/17: Language teaching in primary and secondary schools in England; Survey report*. London: British Council.

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.

UKLO. (2016). *The United Kingdom Linguistics Olympiad: Test papers for 2016*. Accessed 11 January 2017 at http://www.uklo.org/problems-2016

VanPatten, B. (2015). Foundations of processing instruction. *International Review of Applied Linguistics in Language Teaching*, *53*, 91–109.

APPENDIX A

Key Features of Previous Studies Investigating Distribution of Practice Effects

Study	Proficiency (as Authors Describe)	Age of Learners	L1-L2 of Learners	Language Feature	Amount and Type of Instruction	Outcome Measures	ISI	RI 1 (ISI:RI)	RI 2 (ISI:RI)	Findings
Bird (2010)	Intermediate	Adult (19–23 years)	L1 Malay; L2 English	Simple past / present perfect / past perfect verb morphology	300 minutes over five sessions; written error identification and correction tasks	Written error identification and correction test	3-day; 14-day	7-day (42%; 200%)	60-day (5%; 23%)	Longer > shorter spacing
Rogers (2015)	Intermediate	Adult (17–28 years)	L1 Arabic; L2 English	Complex syntactic structures (e.g., *Where Sue is in the car not on the boat*)	75 minutes over five sessions; written sentence-level comprehension questions	Written grammaticality judgement test	2.25-day; 7-day	0-day	42-day (5%; 17%)	Longer > shorter spacing

(Continued)

APPENDIX A
(Continued)

Study	Proficiency (as Authors Describe)	Age of Learners	L1-L2 of Learners	Language Feature	Amount and Type of Instruction	Outcome Measures	ISI	RI 1 (ISI:RI)	RI 2 (ISI:RI)	Findings
Suzuki & DeKeyser (2015)	Beginner	Adult (M = 21 years)	L1 English (one L1 Nepali, one L1 Romanian); L2 Japanese	*te-* prefix indicating a realized state or activity	90–100 minutes over two sessions; oral vocabulary training, written explicit information, aural comprehension practice	Oral rule application test; oral picture-sentence completion test	1-day; 7-day	7-day (14%; 100%)	21-day (3%; 25%)	Accuracy: longer = shorter spacing; response time (picture–sentence completion test): shorter > longer spacing
Suzuki (2017)	Beginner	Adult (M = 19.63 years)	L1 Japanese; L2 Supurango (artificial)	Present progressive inflections	204 minutes over four sessions; oral vocabulary practice, written explicit information, oral grammar practice	Oral rule application test; oral picture description test	3.3-day; 7-day	7-day (47.1%; 100%)	28-day (11.8%; 25%)	Accuracy: shorter > longer spacing; response time (both tests): shorter = longer spacing

(Continued)

APPENDIX A
(Continued)

Study	Proficiency (as Authors Describe)	Age of Learners	L1-L2 of Learners	Language Feature	Amount and Type of Instruction	Outcome Measures	ISI	RI 1 (ISI:RI)	RI 2 (ISI:RI)	Findings
Current study	Beginner	Child (8–11 years)	L1 English; L2 French	Verb inflections for number and tense (first/third person)	180 minutes over three or six sessions; explicit information and aural and written input-based form-meaning mapping practice	Written sentence–picture matching test; written acceptability judgement test	3.5-day; 7-day	3.5-day; 7-day (100%)	42-day (8.3%; 16.7%)	Longer = shorter spacing

Note. ISI = intersession interval; RI = retention interval.

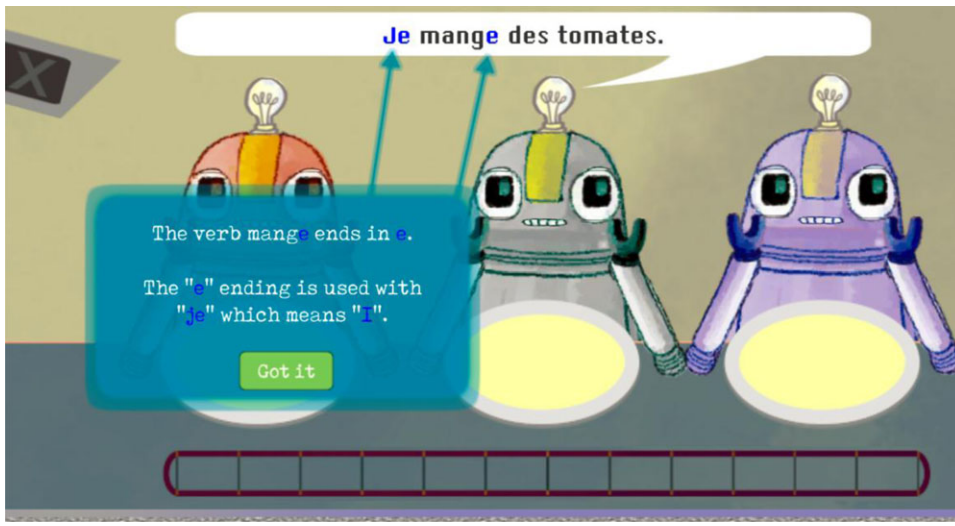
APPENDIX B

Verbs Included in Training and Testing Materials

Training & Testing Materials	*adorer*	to love	*parler*	to talk
	aimer	to like	*porter*	to wear
	chercher	to look for	*poster*	to post
	jouer	to play	*surveiller*	to watch
	manger	to eat	*trouver*	to find
	marcher	to walk	*visiter*	to visit
Testing Materials Only	*chanter*	to sing	*nager*	to swim
	danser	to dance	*pêcher*	to fish
	dessiner	to draw	*promener*	to walk (the dog)
	écouter	to listen	*regarder*	to watch (TV)
	embrasser	to kiss	*téléphoner*	to telephone
	laver	to wash		

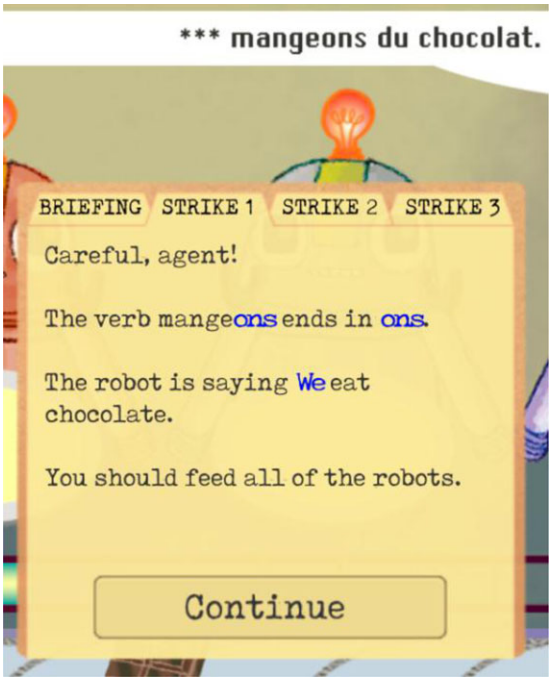
APPENDIX C

Example of Explicit Information Provided During Training Tutorial



APPENDIX D

Example of Feedback Provided for Incorrectly Answered Items During Training



APPENDIX E

Effect Sizes Corrected for Baseline Differences for Comparison Between Groups at Post- and Delayed Posttest on Outcome Measures

Test	Group	Posttest *d*	Delayed Posttest *d*
Sentence–Picture Matching (/8)	7-day vs. 3.5-day	−0.65	−0.36
	7-day vs. control	0.01	−0.23
	3.5-day vs. control	0.66	0.16
AJT Grammatical (/6)	7-day vs. 3.5-day	−0.38	−0.46
	7-day vs. control	0.10	0.55
	3.5-day vs. control	0.52	1.06
AJT Ungrammatical (/6)	7-day vs. 3.5-day	0.49	0.52
	7-day vs. control	0.86	0.49
	3.5-day vs. control	0.27	−0.09

Note. Correction calculation = effect size at (delayed) posttest − effect size at pretest (see Table 3a).

APPENDIX F

Correlation Between Language Analytic Ability and Pretest Scores on Outcome Measures

Test	Group	*n*	*rho* (95% CIs)	*p*
Sentence–Picture matching	7-day	38	−.13 (−.48, .24)	.423
	3.5-day	41	−.19 (−.47, .13)	.241
	Control	34	−.13 (−.48, .22)	.451
	All	113	−.18 (−.35, −.02)	.055
AJT Grammatical	7-day	20	.02 (−.39, .48)	.950
	3.5-day	26	.09 (−.35, .53)	.666
	Control	16	.31 (−.20, .76)	.251
	All	62	.05 (−.22, .32)	.692
AJT Ungrammatical	7-day	20	.19 (−.23, .55)	.405
	3.5-day	26	−.02 (−.46, .48)	.924
	Control	16	.18 (−.50, .79)	.492
	All	62	.10 (−.17, .38)	.426

Note. *rho* = Spearman's *rho* correlation coefficient; CIs = bootstrapped 95% confidence intervals.