EMPIRICAL STUDY

The Roles of Recasts, Task Complexity, and Aptitude in Child Second Language Development

Nektaria-Efstathia Kourtali[a] and Andrea Révész[b]

[a]University of Liverpool and [b]University College London

We would like to thank the anonymous reviewers, Emma Marsden, and Pavel Trofimovich for their insightful comments and suggestions on previous versions of this manuscript.

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

Correspondence concerning this article should be addressed to Nektaria-Efstathia Kourtali, University of Liverpool, School of the Arts, 19 Abercromby Square University of Liverpool, L69 7ZG , United Kingdom. E-mail: nektaria.kourtali@liverpool.ac.uk

This study investigated the effects of task complexity on child learners' second language (L2) gains, the relationship between aptitude and L2 development, and the extent to which task complexity influences this relationship when recasts are provided. Sixty child EFL learners were assigned to two experimental groups. During the treatment, one group completed simple information transmission tasks, whereas the other group performed complex decision-making tasks. In response to errors in the use of the present third person singular verb forms, participants received recasts. L2 development was measured through oral production, written production, and elicited imitation tests. Aptitude was assessed through LLAMA D, LLAMA E, and LLAMA F. Less cognitively demanding tasks were more beneficial. Participants' performance on LLAMA E predicted L2 gains measured through elicited imitation, and their LLAMA D scores predicted development measured through the oral and written production tests under complex task conditions.

Keywords recasts; task complexity; aptitude; child learners; second language; elicited imitation; LLAMA

Introduction

The focus on form approach to second language (L2) teaching has inspired a large body of research in the field of instructed L2 acquisition. Motivated by Long's interaction hypothesis, this approach posits that drawing learners' attention to linguistic elements while engaging in meaningful interaction facilitates subsequent L2 development (Long, 1996, 2015). One way to promote a focus on form is by providing learners with corrective feedback, that is, "responses to learner utterances that contain an error" (Ellis, Loewen, & Erlam, 2006, p. 340). A number of meta-analyses indicate that corrective feedback can assist interlanguage development (e.g., Li, 2010; Lyster & Saito, 2010). However, which types of corrective feedback benefit learners more and under what conditions remains a disputed issue among researchers. A type of corrective feedback that has been the object of a plethora of research is recasts. Recasts are generally defined as reformulations of a learner's utterance in which one or more errors are altered while keeping the original content. Example 1, obtained from data collected for the present study, shows how a learner's utterance is reformulated in order to address an error in the present third person singular verb form by employing a recast.

Example 1

Learner: He **dance** tango.

Researcher:	He **dances**?

Previous research has demonstrated that the effectiveness of recasts in leading to interlanguage development is influenced by several variables (see Loewen & Sato, 2018, for a review). Of these, task complexity (i.e., inherent cognitive demands of tasks) and individual differences in L2 aptitude have been the focus of this investigation. For both variables, previous studies have exhibited mixed findings. Task complexity has been found to influence the efficacy of recasts; nonetheless, it remains unresolved whether recasts are more beneficial for learners during cognitively simple tasks or complex tasks (Baralt, 2013; Kim, Payant, & Pearson, 2015; Révész, 2009; Révész, Sachs, & Hama, 2014). The findings are also contradictory for the relationship between aptitude and the effectiveness of recasts, with some studies observing evidence for this link (Li, 2013; Trofimovich, Ammar, & Gatbonton, 2007; Yilmaz, 2013) and others yielding no relationship between aptitude and learning through recasts (Sheen, 2007; Yilmaz & Grañena, 2016). One reason for the incongruent results regarding task complexity might be that existing studies have not controlled for potential differences in aptitude among learners, because there are indications in the literature that aptitude interacts with different learning conditions (Suzuki & DeKeyser, 2017a). Regarding aptitude and task complexity, it has been proposed that the cognitive complexity of the task during which learners receive feedback might moderate the extent to which aptitude predicts L2 development (e.g., Robinson, 2011).

The novelty of this study was to test this prediction—namely, that recast effectiveness for learners of varying language aptitude is associated with task complexity—by investigating the extent to which task complexity influences the relationship between aptitude and L2 development when learners receive recasts. We also intended to expand on previous research by examining the extent to which aptitude explains L2 learning, regardless of task complexity. An additional contribution of the study is its focus on child EFL learners, a population that remains underresearched both within the context of task-based language teaching and in the larger field of corrective feedback (Li, 2010). In addressing these goals, it was hoped that the study would inform theoretical models of task-based instruction and would help provide guidance to practitioners about how to adapt tasks to maximize learning in language classes where students are likely to have differential aptitude profiles.

Background

Task Complexity and Recasts

Previous research on task complexity and corrective feedback has been informed by two cognitive models: the cognition hypothesis (Robinson, 2001, 2011) and the limited attentional capacity model (Skehan, 2009, 2014). According to Robinson (2001, 2011), task complexity is defined as the cognitive demands imposed on L2 learners by the inherent characteristics of the task in which they engage. Robinson contends that when the cognitive demands of tasks are increased along resource-dispersing factors (e.g., planning time), learners' memory and attentional resources will be dispersed, which will affect the production and uptake of focus on form in negative ways. Conversely, when task complexity is increased along resource-directing dimensions (e.g., reasoning demands), Robinson predicts that learners' attentional and memory resources will be directed to the functional and linguistic demands of the task, leading to positive effects on production and incorporation of information presented through focus on form interventions such as recasts.

Another model that has been used to explain the effects of corrective feedback under simple and complex task conditions is Skehan's (2009, 2014) limited attentional capacity model. Skehan's model is largely inspired by Levelt's (1989) model of speech production, which describes speech production as consisting of four stages: (a) Conceptualization involves planning the content of one's message; (b) formulation refers to the grammatical, lexical, and phonological encoding of the message; (c) articulation is associated with the production of speech sounds; and (d) finally, self-monitoring involves evaluating whether the output produced is accurate and appropriate. Using Levelt’s model, Skehan (2009, 2014) argues that tasks with greater cognitive demands will complexify conceptualizer processes, resulting in fewer attentional resources available for linguistic encoding, and thereby leading to the production of less accurate and/or less complex language. In other words, Skehan (2009, 2014) contends that greater cognitive complexity (e.g., greater reasoning demands) is expected to burden the conceptualizer at the expense of linguistic encoding. Drawing on Skehan’s model, Révész et al. (2014) argued that the provision of recasts may be more effective when learners perform tasks with lower cognitive demands, as the decreased pressure on the conceptualizer may better enable recasts to draw learners’ attention to linguistic encoding.

The combined effects of recasts and task complexity have been examined only in a handful of studies (Baralt, 2013; Kim et al., 2015; Révész, 2009; Révész et al., 2014). Révész (2009) explored the joint impact of recasts and task complexity on L2 development in the use of the past progressive form. One group of students received recasts while engaged in a photo description task without contextual support (i.e., from memory), whereas in the other group, recasts were supplied during the same photo description task accompanied by contextual support (i.e., the photo was available during task performance). The study revealed that participants benefited more from recasts when no contextual support was available. While the results of this study provided clear evidence that lack of contextual support facilitated the efficacy of feedback, they allow for no straightforward conclusions regarding the role of task complexity. The issue of whether the presence or absence of contextual support poses higher cognitive demands remains an object of debate (see Robinson, 2003 vs. Skehan, 2014).

Using a different task manipulation, Révész et al. (2014) examined how increasing the reasoning demands of tasks may affect the efficacy of recasts in developing learners' knowledge of the past counterfactual construction in a computer-mediated context. The participants' task was to identify causes and effects of events based on a picture story they had read. The cause–effect relationships were designed to be more obvious in the simple compared to the complex task condition. The study demonstrated that recasts supplied during simple tasks were significantly more beneficial than those delivered during complex tasks. Drawing on Skehan's limited capacity model and Levelt's speech production model, the researchers argued that when the learners performed the complex task versions, they probably allocated more attentional resources to task completion, leaving less attention available for the processing of recasts and their linguistic target.

Baralt (2013) also explored the combined effects of recasts and task complexity on L2 development. The novelty of this study was to investigate whether any effects of task complexity would differ across face-to-face and computer-mediated environments. The linguistic target was the Spanish past subjunctive, and task complexity was operationalized in terms of the intentional reasoning demands posed by story retelling tasks. Under the complex condition, participants had to think of the characters' intentional reasons while retelling stories, whereas under the simple condition, the characters' intentions were provided in the stories. Interestingly, Baralt found that in the face-to-face mode, recasts were more beneficial when delivered during complex tasks with greater cognitive demands, whereas in the computer-mediated mode, recasts were more successful during simple tasks requiring less reasoning. Baralt attributed this inconsistency in findings to differences in discourse length, number of turns, and contingency of recasts across the two modalities.

Taken together, the results of studies exploring the combined effects of task complexity and corrective feedback on L2 development are inconclusive. Drawing on previous research on aptitude–treatment interaction (see Suzuki & DeKeyser, 2017a), one possible explanation for the mixed findings might be that existing studies did not control for individual differences in cognitive abilities such as working memory capacity (cf. Kim et al., 2015) and L2 aptitude. This study aimed to explore this possibility by investigating how aptitude might moderate the link between task complexity and the effectiveness of recasts.

Models and Measures of Aptitude

Language aptitude refers to cognitive and perceptual abilities that facilitate L2 acquisition (Carroll, 1965, 1981; Grañena, 2013). Carroll, one of the pioneers of foreign language aptitude research, conceptualized aptitude as involving four components: (a) phonetic coding ability, (b) inductive language learning ability, (c) grammatical sensitivity, and (d) rote learning ability or associative memory. Phonetic coding entails identification of sounds, making connections between sounds and their symbols, and retaining them. Inductive language learning refers to the ability to induce rules from input. Grammatical sensitivity enables learners to identify the functions of words in sentences. Finally, rote learning is the ability to not only identify connections between sounds and meanings, but also to retain them. The Modern Language Aptitude test (MLAT), which was designed by Carroll and Sapon (1959), remains one of the most influential language aptitude tests. It measures all subconstructs of aptitude put forward by Carroll except inductive language learning ability.

Partly building on Carroll’s work, Skehan (1998) adopted an information processing perspective to model aptitude and described L2 aptitude as a construct involving cognitive differences in memory-as-retrieval, phonetic coding, and language analytic ability, with language analytic ability subsuming Carroll’s grammatical sensitivity and inductive language learning. In an update to the model, Skehan (2002) added attentional control and working memory among the subconstructs of aptitude. Skehan (2002, 2016) also proposed that the different aptitude subconstructs are implicated at various stages of the L2 acquisition process.

In this model, several components of aptitude can be related to the cognitive processes involved in learning grammar through feedback. Attentional control and working memory are likely to be relevant at the stage when feedback is segmented. Working memory and phonetic coding ability may help learners notice the corrective function of feedback and the error highlighted. Recognizing patterns through exposure to feedback may be facilitated by phonetic coding ability, working memory, and language analytic ability. Finally, working memory and retrieval memory are expected to assist learners in avoiding errors.

Although the MLAT is still widely used, several new aptitude tests have been designed over the past two decades, differing in purpose and targeted populations. The CANAL-F test (Grigorenko, Sternberg, & Ehrman, 2000) was created to assess learners' ability to cope with novel L2 learning conditions. The Hi-LAB battery (Linck et al., 2013) aims to identify cognitive abilities that foster the achievement of advanced L2 skills. It includes measures of the central executive component of working memory, phonological short-term memory, associative memory, long-term memory retrieval, processing speed, implicit learning, and auditory discrimination. The present study used the LLAMA test (Meara, 2005) as a measure of aptitude (but see Bokander & Bylund, 2019, for reservations about the validity of LLAMA as a test of aptitude). The LLAMA test, an instrument frequently employed to test aptitude in current L2 acquisition research, is composed of four subtests. These intend to measure rote, associative memory (LLAMA B), the ability to recognize patterns in spoken language (LLAMA D), the ability to associate sounds with symbols (LLAMA E), and inductive language learning ability (LLAMA F). Considering that the learners of the current study were required to process oral feedback targeting a grammatical feature, sound sequence recognition (LLAMA D), phonetic coding (LLAMA E), and grammatical inferencing (LLAMA F) ability were considered relevant to learners' development.

Aptitude, Recasts, Task Complexity and L2 Development

In line with the predictions derived from Skehan's model, previous studies of aptitude and corrective feedback have found that learners with higher aptitude benefit more from the provision of explicit feedback (e.g., Yilmaz, 2013; Yilmaz & Grañena, 2016; Yilmaz & Koylu, 2016). However, existing research has yielded mixed findings for the relationship between aptitude and recasts. While in some studies the effectiveness of recasts was not linked to learners' aptitude (Sheen, 2007; Yilmaz & Grañena, 2016), other studies have shown that the extent to which recasts lead to L2 gains is related to components of aptitude such as language analytic ability (Li, 2013) and attention control (Li, 2013; Trofimovich et al., 2007). One explanation for the contradictory findings appears to relate to the nature of the linguistic feature targeted by recasts. Yilmaz (2013), for example, found that higher language analytic ability facilitated learning from recasts when the focus was a salient feature (Turkish plural morpheme), whereas when recasts targeted a less salient construction (Turkish locative case marker), language analytic ability did not assist L2 learning. Although previous research has explored the effects of corrective feedback on learning linguistic constructions that differ in salience (see Sato & Loewen, 2018), the role of aptitude in facilitating the benefits of corrective feedback for different types of linguistic features has not received sufficient attention.

Besides the nature of the linguistic target, there are indications in the literature that cognitive task complexity is another factor that may moderate the interaction between recasts and aptitude. As part of the cognition hypothesis, Robinson (2011) hypothesized that learner factors will interact with task complexity in determining the extent to which learners benefit from pedagogical interventions. In particular, he predicted that individual differences in cognitive abilities will be increasingly associated with learning as tasks increase in cognitive complexity, resulting in less variation in gains when learners complete tasks with lower cognitive demands than when they perform complex tasks. In line with this prediction, an empirical study by Kim et al. (2015) revealed that learners with high working memory who performed tasks with greater cognitive demands benefitted the most from receiving recasts. To date, however, it remains an empirical question whether task complexity affects the relationship between other aptitude components (e.g., phonetic coding and language analytic ability) and the effectiveness of recasts in developing L2 knowledge.

The Current Study

The current study had three main aims. First, we intended to contribute to the existing research by investigating the effects of task complexity on L2 development when learners receive recasts Our second goal was to expand on previous research by examining the extent to which aptitude accounts for the efficacy of recasts. Finally, the novelty of our study was to explore the extent to which task complexity may influence the relationship between aptitude and L2 development when learners receive recasts during task-based interaction. In addition, we focused on an underresearched population, child EFL learners. To date, only a few studies (Harley & Hart, 1997; Ranta, 2002) have explored the relationship between aptitude and L2 outcomes of child language learners, and none have examined the role of aptitude under different task conditions. Furthermore, our linguistic focus was the third person singular *–s* verb form, a linguistic feature generally considered to be a difficult L2 learning target. The following research questions were formulated:

1. What are the effects of task complexity on developing child L2 learners’ knowledge of the present third person singular?
2. To what extent does aptitude predict development of child L2 learners' knowledge of the present third person singular?
3. To what extent does task complexity influence the relationship between L2 aptitude and development in child L2 learners’ knowledge of the present third person singular?

Method

Design

The study employed a pretest–posttest design with two treatment sessions. Participants were assigned to one of two experimental groups through stratified random sampling, taking into account their pretest, proficiency, aptitude test results, and length of prior English study. One group carried out two simple information transmission tasks with no reasoning demands during the treatment sessions, whereas the other group worked on two complex decision-making tasks, which posed reasoning demands on learners. Both groups received interrogative recasts in response to errors related to the target feature (see Example 1). Each treatment task was followed by a posttask questionnaire, which was included to assess the perceived cognitive demands posed by the treatment tasks, and thereby provide evidence for the validity of the task complexity manipulation (Norris, 2010; Révész, 2014). L2 development was evaluated through an oral production test, a written production test, and an elicited imitation (EI) test. L2 aptitude was measured through the LLAMA test (Meara, 2005).

Participants

The initial pool of participants included 160 EFL learners. All attended English classes in private language schools in Greece. English is also part of the national curriculum in state schools, with students beginning to learn English at the age of seven. In both contexts, the students received form-focused instruction using the communicative approach. The first author recruited participants by sharing the information sheet with the owners of several language schools, child L2 learners, and their parents.

Of the original pool, 60 learners were considered eligible to participate in the study. They were L1 speakers of Greek or bilingual speakers born in Greece, their level was A2 (elementary), they had never lived in an English-speaking country prior to the study, and they did not demonstrate extensive prior knowledge of the target feature on any of the pretests. Participants who did not satisfy these criteria were excluded from the study. Participants' proficiency was assessed through the listening component of the Trinity College ISE Foundation test, which targets level A2 in the Common European Framework of Reference (CEFR). To qualify for the study, students needed to achieve an ISE score in the range of 2 through 4. The pretests assessed participants' prior knowledge of the target feature; those who scored higher than 35% on any of the outcome measures were excluded from the study to avoid ceiling effects. This threshold was determined based on the distribution of scores (the next highest score was 50% on the oral production test and 90% on the written production test). The rationale for excluding participants who achieved high scores on the pretests was that extensive prior knowledge could obscure the role of task complexity and aptitude in L2 learning. Also, we would have encountered a ceiling effect on improvement if we had included participants achieving higher than 90% on one of the pretests.

The final pool of participants included 26 female and 34 male learners. Their ages ranged from 10.5 to 13 years (*M* = 11.46, *SD* = .82). They were all native speakers of Greek, apart from two students who were Greek–Albanian and Greek–Romanian bilinguals, both born in Greece. Participants' length of English study prior to the experiment ranged from 2 to 8.5 years (*M* = 4.65, *SD* = 1.17). The majority of participants also reported learning L2 French and German (*n* = 48). A series of independent samples *t* tests targeting the variables of age, length of previous English study, and performance on the proficiency test confirmed that the two groups were comparable: age, *t* = 0.78, *p* = .44, *d* = 0.14; English study, *t* = 0.33, *p* = .75, *d* = 0.08; proficiency, *t* = 0.07, *p* = .95, *d* = 0.01.

Linguistic Target

The present third person singular verb form was selected as the linguistic target. This feature is regarded as a difficult structure to acquire for several reasons. First, it constitutes a bound morpheme, which is realized through three allomorphs [s], [z], and [əz]. Thus, the grammatical function maps onto multiple forms, making it difficult for learners to discern the form–function relationship through exposure to oral input. Second, the structure lacks physical salience, given that none of the three allomorphs are stressed. As a result, the form may remain unattended by learners. Third, the present third person singular is a communicatively redundant feature (VanPatten, 1996). Its meaning can be conveyed successfully through subject noun phrases, which render the use of verbal inflection in a sentence redundant. The present third person singular is prone to fossilization in adulthood, probably due to these characteristics (Han, 2013). It is important, therefore, to identify which learning conditions (e.g., recasts delivered during simple or complex tasks) may help promote its accurate use among child L2 learners. In addition, it is of theoretical interest whether different L2 aptitude constructs may be related to the learning of a nonsalient, redundant feature.

Treatment Tasks

Both the simple (information transmission) and complex (decision making) task conditions required participants to talk about habits of fictional characters. In the first decision-making task, participants were asked to act as the administrator of a building, and the researcher (first author) played the role of their assistant. According to the task instructions, they had found several items left by a transportation company at the entrance of their residence, and the owners of these items were the tenants of the apartment block. The participants (i.e., administrators) had to decide which items belonged to whom by using information they had about the tenants' habits (e.g., "The lamp is Mike's because he works at night."). They also had to inform the assistant (i.e., researcher) about these decisions so she could return the items to the tenants. In the corresponding information transmission task, participants played the role of the assistant to the administrator (i.e., researcher), and they were asked to give information about what the tenants usually did on weekends so the administrator could make decisions about who the owners of the items were. For example, participants informed the researcher that Mike works at night, which led the researcher to think that the lamp was Mike's. In the second decision-making task, each participant was asked to imagine that they worked at an airport. They needed to find the owners of lost items and justify their decisions using a table showing the owners' habits. In the information transmission version of the task, each participant was assigned the role of an assistant who worked at the airport. To help another employee at the airport (i.e., the researcher) find the owners of lost items, they were asked to give information about the habits of the owners using a table with pictures. The instructions for all tasks were delivered in the participants' L1 (Greek).

The decision-making tasks were considered more cognitively complex because they required reasoning on the part of participants. On the other hand, no reasoning was needed to complete the information transmission tasks, so this task type was expected to pose lower cognitive demands (Baralt, 2013; Kim et al., 2015; Révész et al., 2014; Robinson, 2001, 2011). The tasks were piloted with a population similar to the participating children and were found to be age appropriate. The tasks also succeeded in eliciting the target structure.

Type of Recasts

During the treatment tasks, participants received a recast when they produced the present third person singular verb form inaccurately, as shown in Example 2.

Example 2

Learner: Mark on Saturdays **play** the guitar at 6 to 7 p.m. He is at home…

Researcher: He **plays**?

In terms of Lyster's (1998) categorizations, the recasts were interrogative and isolated, as they reformulated only the erroneous part of participants' utterances without providing additional information. The recasts were also reduced, consisting of a personal pronoun and a verb in the present third person singular. They focused on one change, either the addition of the allomorph [s], [z], [ɪz], or [əz], or a substitution when the researcher replaced a nontargetlike utterance (e.g., another tense) with the present third person singular. Thus, recasts were delivered in an intensive and focused manner. Given these characteristics and the utterance-final position of the correction (Bardovi-Harlig, 1987), the recasts in the present study could be classified as explicit (Sheen, 2006). Our rationale for utilizing explicit recasts was that previous studies have demonstrated implicit recasts as being less successful in promoting noticing and learning, especially when targeting nonsalient constructions (e.g., Nassaji, 2009; Yilmaz, 2012).

Posttask Questionnaire

After each treatment task, both groups completed a posttask questionnaire to gain information about participants' task perceptions. The questionnaire was adapted from Robinson (2001), and asked participants to provide ratings on a 9-point semantic differential scale about (a) the amount of mental effort required by the task, (b) the overall task difficulty, (c) the linguistic demands posed by the task, and (d) the quality of their own task performance. Given the nature of the task manipulation, it was considered important to differentiate mental effort from task difficulty. While the decision-making tasks were unlikely to be perceived as difficult by the participants, we assumed that this task type would require greater mental effort in comparison to mere information transmission. Participants' responses were converted into a 9-point numerical scale with higher values indicating greater mental effort, task difficulty, linguistic demands, and better perceived performance. The resulting values were used for subsequent statistical analyses. The items were presented in the participants' L1 to facilitate understanding and, during data collection, the researcher provided clarification to participants when needed.

Assessment Tasks

Oral and Written Production Tests

The oral and the written production tests were designed to assess participants' ability to produce the target structure in the oral and written mode, respectively. Each test included 12 pictures that prompted the participants to talk about the habits of fictional characters who were different from those presented in the treatment tasks. For both tests, two versions were designed that were counterbalanced across the pretest and posttest. That is, one version was used as a pretest for half of the participants and as a posttest for the other half. The same procedure was followed for both groups. The two versions included different pictures, but all of the pictures depicted the same actions in order to elicit the same verbs. The number of pictures was the same for the two test versions to generate a similar number of obligatory contexts for the target feature, and each test was designed to elicit verbs with all three allomorphs. The pictures on the written test showed the same habits as the ones on the oral production test, to ensure that the verbs produced were identical in the two modes. The test instructions were delivered in the participants' L1.

The rationale for including an oral as well as a written assessment was to tap into different types of knowledge. Although both tests implicated procedural knowledge (Anderson, 1993; DeKeyser, 2007), the oral production test naturally posed greater time pressure than the written production test. Hence, participants' performance on the oral production test was considered a better indicator of procedural knowledge in the process of automatization. On the other hand, the written production test, in the absence of time pressure, likely better enabled participants to revise their output and thus deploy their explicit declarative and procedural knowledge during production. Using tests in different modes also made it possible to obtain a fuller picture about participants' developing knowledge of the target structure, thereby employing more rigorous inclusion criteria for the study. For example, many participants in the initial pool scored 100% on the written production test but had low scores on the oral tests. These participants were excluded from the study to ensure that participants' prior knowledge did not confound the results obtained for task complexity and aptitude.

Elicited Imitation (EI) Test

The EI test was administered using Microsoft PowerPoint. After listening to a sentence, the participants saw two pictures labeled A and B and were asked to choose the picture that was relevant to the meaning of the utterance they had heard by saying A or B out loud. Next, when the color of the slide changed from white to brown, the students had to produce the sentence in correct English. There was a 4 second time interval between the presentation of the stimulus and the elicited response to ensure that the production of the sentence did not involve rote repetition (Erlam, 2006). The participants had 10 seconds to produce the sentence.

The EI test started with seven practice items that did not include the target structure. The actual EI test consisted of 72 sentences: 24 targeted the present third person singular, and 48 served as distractors. For both target items and distractors, half of the sentences were grammatical and half were ungrammatical. In the target items, the three allomorphs were equally distributed, including three grammatical and three ungrammatical sentences for each of the three allomorphs [s], [z], and [əz]. Following Keating and Jegerski (2014), the items were pseudorandomized so that the same allomorphs were not presented in succession. In the ungrammatical target items, only the base form of the verb was used.

Based on Keating and Jegerski (2014), two versions were designed for each item to combat item effects. The two versions differed as to whether the target verb was presented as grammatical or ungrammatical (e.g., "He always take the bus" vs. "He always takes the bus"). To avoid repetition effects, participants did not encounter both versions of the same item (Keating & Jegerski, 2014). Thus, in both groups, half of the participants were administered Version A and the other half Version B of the test. The test versions and the order of the test items were counterbalanced across the pretest and the posttest.

Drawing on Keating and Jegerski (2014), the critical verbs inflected with the present third person singular were of the same length (i.e., one-syllable verbs) and located in the same position in each target sentence to impose similar processing demands on the participants. According to Vocabprofiler (Cobb, 2016; Heatley, Nation, & Coxhead, 2002), the verbs were among the 1,000 most frequent words in the English language (K1 band), and from the second most frequent 1,000 words (K2 band). The rationale for including words from the K2 band was that they frequently occur in textbooks, and the pilot study revealed that they were familiar to Greek EFL learners with a similar background (e.g., the verb "dance"). In each target sentence, the precritical region (i.e., the structure prior to the present third person singular) consisted of a personal pronoun followed by a two-syllable adverb of frequency (e.g., "He sometimes helps his mum."). The personal pronoun served as the subject of the critical verb, and the adverb was included to create an obligatory context for the present third person singular. The length of the target stimuli and distractors was the same; all 72 sentences consisted of six syllables. The reliability of the two versions of the EI test was found to be high (Version A α = .842; Version B α = .867).

Although both the oral production test and the EI test were delivered in the oral mode, a key difference between the tests was that the oral production test was less controlled, allowing partricipants to produce their own output, whereas the EI test involved the processing of oral input and production of predetermined utterances. The type of knowledge EI tests measure is a matter of debate, namely, whether they capture implicit knowledge (e.g., Erlam, 2006) or automatized explicit knowledge (e.g., Suzuki & DeKeyser, 2015). However, in light of Suzuki's (2017) finding that it takes a number of years in immersion contexts to use implicit knowledge more reliably, it would appear more likely that the child EFL participants in the current study would rely on automatized explicit knowledge given their limited exposure to English.

Aptitude Test

The study employed a computer-administered aptitude test called LLAMA (Meara, 2005). The LLAMA test is language independent and uses visual stimuli and an unfamiliar language. The test instructions were given in the participants' L1, following the LLAMA manual. The following components from the test were administered.

LLAMA D is a sound recognition test providing insights into language users' ability to recognize patterns in spoken language. First, the participants were asked to listen to a string of 10 computer-generated sound sequences only once. Then, they completed a recognition test that required them to distinguish sounds they had already heard from novel ones. This LLAMA test component was considered relevant to the treatment because the recasts intended to facilitate the recognition of a morphological rule through exposure to oral input.

LLAMA E is a sound–symbol correspondence task measuring language users' phonetic coding ability. First, the participants had two minutes to listen to 24 recorded syllables provided in combination with their transliterations, and they were expected to work out the sound–symbol correspondence for each syllable. Then, they had to listen to a two-syllable word while given two possible written representations. The participants were asked to choose the symbols that correctly corresponded to the words they heard. This test would appear pertinent to the treatment given that a prerequisite for learning from recasts is that participants decode the linguistic information entailed in them.

LLAMA F has been designed to measure grammatical inference. The participants were presented with 20 sentences accompanied by pictures. Each sentence described a picture displayed on the screen. The participants had five minutes to read the sentences, look at the corresponding pictures, and figure out the grammatical rules of the new language. This task was followed by a test with items that presented a picture and two sentences, one grammatical and one ungrammatical. The participants had to choose the correct sentence for each picture by making sensible inferences about the grammar and morphology of the novel language used in the test. It seemed reasonable to assume that the ability to make grammatical inference would be related to the ease with which a grammatical structure is acquired.

Procedure

Data were collected over six months. In the first session, the participants filled out a background questionnaire in their L1. They were also administered the proficiency test and the pretests. They first took the oral production pretest, followed by the EI pretest and, finally, the written production pretest. We administered the written production test last to increase the likelihood that participants rely on procedural knowledge during the oral tests. The written production test was more likely to elicit reliance on explicit, declarative knowledge; thus, performing this test prior to the oral assessment may have primed participants into a more explicit mode during the subsequent oral test. The oral production test preceded the EI test for practical reasons: The oral production test was shorter; hence, participants who did not fit the inclusion criteria could be excluded at an earlier point. After the pretests, the participants were administered the LLAMA D, E, and F tests in order to enable stratified random assignment, which controlled for potential differences in aptitude across the two groups.

The participants who met the selection criteria were invited to attend the two treatment sessions. They performed either two information transmission or two decision-making tasks while receiving recasts in response to errors with the target feature. The two treatment sessions lasted for approximately 10 minutes and were separated by a 5 minute interval. Each treatment task was followed by a posttask questionnaire. Once the treatment was completed, there was another 5 minute interval. Then, the participants were administered the posttests (oral and written production tests, EI test) in the same order as during the pretests. For all the sessions, the researcher met the participants individually in her office.

Considering that some of the tests and the treatment involved processing of oral input, it was important to conduct the study in a quiet environment. Figure 1 shows the experimental schedule.

Place Figure 1 near here

Data Analysis

Transcription

All oral production data from the treatment tasks and tests were transcribed by the researcher. To verify the reliability of the transcriptions, another transcriber also transcribed 10 percent of the data, selected through stratified random sampling to ensure equal representation of the experimental groups and various versions of the treatment and testing tasks. The two transcripts were compared with a focus on the target verbs. Two types of discrepancies emerged: (a) Some items were differently transcribed, and (b) some items were present in one transcript but omitted from the other. Intertranscriber agreement was calculated by dividing the total number of items transcribed identically by the total number of items, and it was found to be high (.988). Cohen's kappa was also computed at .962.

Coding and Scoring

As a first step, we considered participants' responses to recasts, in particular, whether recasts led to successful uptake. Following Lyster and Ranta (1997, p. 49), we defined uptake as "a student's utterance that immediately follows the teacher's feedback and that constitutes a reaction in some way to the teacher's intention to draw attention to some aspect of the student's initial utterance." Drawing on previous literature (Ellis, Basturkmen, & Loewen, 2001; Lyster & Ranta, 1997), uptake was coded as successful (see Example 3) when participants corrected their initial error specific to the present third person singular verb form and unsuccessful when they did not (e.g., they repeated the same error or they made a different error, they replied "yes," or they continued with the following picture of the task).

Example 3

Learner: Mark on Saturdays 6 or 7 o'clock is at home and ***play*** the guitar.

Researcher: He ***plays***?

Learner: ***Plays*** the guitar.

For the oral and written production tests, obligatory contexts for the target feature were identified by the researcher. If the target feature was accurately produced, the participants received 1 point. Vocabulary, pronunciation, and orthographical errors were ignored. Given that participants produced different numbers of obligatory contexts for each item, total scores were used in further analyses, making an item-based analysis impossible. The scoring criteria for the EI test were based on Erlam (2006). The participants were given 1 point when choosing the correct picture to ensure they had processed the meaning of the utterance. Only those who scored at least 90 percent were included in the study. For the EI component of the test, the participants received 1 point when they produced the target feature, and 0 when they produced another construction. Participants' EI score was calculated for the grammatical and ungrammatical items separately and combined. Ten percent of the data were coded by another coder for each outcome measure. The Cohen's kappa values were high: .95 for the oral production test, .96 for the written production test, and .97 for the EI test.

Statistical Analyses

We first carried out a power analysis for all statistical tests using GPower 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007). For all tests, the sample size was found to be sufficient to detect effect sizes in the medium range. We used SPSS Version 22 to estimate the reliability of the EI test, compute descriptive statistics, and run a series of independent samples *t* tests to assess whether there were significant differences between the groups in terms of perceived mental effort, number of obligatory contexts elicited, aptitude profiles, and pretest scores. The rest of the statistical analyses entailed constructing linear mixed-effects models employing the *lm* and *glmer* functions from the *lme* package (Version 1.1-21; Bates, Maechler, Bolker, & Walker, 2015) in the R statistical environment (R Core Team, 2018). For the analyses involving the oral and written production test data, multiple regression analyses were used relying on the *lm* function. In each analysis, the pretest scores served as the covariate and the predictors of interest were task complexity (Research Question 1), the LLAMA scores (Research Question 2), or their interactions (Research Question 3) depending on the research question. The dependent variable was the total score achieved on the oral or written production test. For the EI test, mixed-effects logistic regression analyses were employed using the function *glmer* due to the binary nature of the dependent variable. According to the research questions, the fixed effects were task complexity (Research Questions 1 and 3), the LLAMA scores (Research Questions 2 and 3), and/or their interactions (Research Question 3). The random effects included items and participants. The dependent variable was the EI score. Although grammaticality was found to be a significant predictor of the total EI scores in a preliminary analysis using grammaticality as a fixed effect, *Estimate* = –2.78, *SE* = .20, $z = –13.98$, $p < .01$, and participant (*SD* = 1.79) and item (*SD* = .50) as random effects, we report the data for the EI scores combined. Our rationale for this decision was that parallel results were found for the models with and without grammaticality as a random effect as well as for the separate analyses conducted for grammatical and ungrammatical EI items (see Appendix S1 in the Supporting Information online for the results of these analyses).

For the multiple regression analyses, R^2 values were obtained to measure effect sizes. We computed odds ratios (ORs) to assess the magnitude of effects for the mixed-effects logistic regression models. Following Plonsky and Oswald (2014), *d* values of .40, .70, and 1.00 and R^2 values of .06, .16, and .36 were considered small, medium, and large, respectively. Residual plots were used to check the linearity, homoscedasticity, and normality assumptions for the models.

Results

Preliminary Analyses

To validate the task complexity manipulation, we compared participants' perceptions about the mental effort they exerted while completing the treatment tasks. Table 1 shows that as intended, the learners judged the decision-making tasks to be more cognitively demanding than the information transmission tasks. Independent samples *t* tests confirmed that the reported mental effort was significantly higher on the decision making tasks than the information tasks during both treatment sessions: Treatment 1 $t = 4.47$, $p < .01$; Treatment 2 $t = 3.38$, $p < .01$. The effect size was large for Treatment 1 ($d = 1.15$) and medium for Treatment 2 ($d = 0.86$). This means that the task complexity manipulation was successful as far as the participants' perceptions were concerned.

Place Table 1 near here

Table 2 presents descriptive statistics for the number of obligatory contexts that the tasks created for the target feature during the treatment, the number of recasts participants received, and the amount of successful uptake they produced. An independent-samples *t* test yielded no significant difference between the number of obligatory contexts for the two groups, $t = 1.19$, $p = .24$, $d = 0.31$. In other words, the two groups had a comparable number of opportunities to produce the target feature and receive feedback on their use. Furthermore, an independent-samples *t* test showed no significant difference between the number of recasts the two groups received, $t = 1.89$, $p = .39$, $d = 0.48$. Finally, an independent-samples *t* test demonstrated that the two groups produced similar amounts of successful uptake after receiving recasts. In other words, there were no differences between the experimental groups in the extent to which they corrected the target feature in response to recasts, $t = .99$, $p = .33$, *d* = .25. A series of Pearson correlations were also run to explore possible relationships between successful uptake and aptitude in the two groups. In the simple condition, no relationship was found for LLAMA D (*r* = –.34, *p* = .06), for LLAMA E (*r* = .09, *p* = .65), and for LLAMA F (*r* = –.02, *p* = .90). In the complex condition, there was only a weak positive correlation between LLAMA E and successful uptake (*r* = .37, *p* = .05). No significant relationship was identified for LLAMA D (*r* = –.123, *p* = .52) or for LLAMA F (*r* = .26, *p* = .16).

Place Table 2 near here

Research Question 1: Effects of Task Complexity on L2 Development

The first research question investigated the effects of task complexity on the development in the knowledge of the target feature when recasts are provided. Table 3 summarizes the descriptive statistics for the participants' pretest and posttest scores across the two groups. In all the tests, participants showed improvement from the pretest to the posttest, with the information transmission group outperforming the decision-making group considerably on the oral and written productions tests. However, the difference between the two groups' gains on the EI test was small.

Place Table 3 near here

The independent-samples *t* tests, which were carried out to investigate whether the two groups had differential prior knowledge of the target structure at the time of the pretest, found no significant difference for any of the three tests: oral production *t* = 0.12, *p* = .95, *d* = 0.03; written production *t* = 0.35, *p* = .73, *d* = 0.09; EI overall *t* = 0.15, *p* = .88, *d* = 0.06. Thus, any significant difference detected at the time of the posttest could not be attributed to differential knowledge of the third person singular –*s* at the outset of the experiment.

To assess whether task complexity affected the participants' development in the use of the third person singular –*s* on the oral and written production tests, linear regression analyses were conducted for the two tests separately using the relevant pretest score and task complexity as predictor variables. In both analyses, the pretest score served as the covariate, and task complexity was the predictor of interest. As shown in Table 4, the regression analyses yielded a significant effect for task complexity for both the oral and written productions tests, with effect sizes being in the small range.

Place Table 4 near here

To examine the effects of task complexity on participants’ production for the EI test, a mixed-effects regression analysis was carried out. The dependent variable was the EI score, the fixed effects were participants’ pretest scores and task complexity, and the random effects were participant and item. Again, the pretest scores served as the covariate, and the predictor of interest was task complexity. No significant effects emerged for task complexity, as shown in Table 5.

Place Table 5 near here

Taken together, these findings indicate that participants demonstrated significantly greater gains on the oral and written production tests when they completed information transmission compared to decision-making tasks during the treatment sessions, but they had parallel gains on the EI test regardless of group assignment.

Research Question 2: Aptitude as a Predictor of L2 Development

The second research question examined the extent to which aptitude predicts development in the knowledge of the target structure when learners receive recasts As a preliminary step, we ran a series of Pearson correlations to examine the relationships among the various LLAMA components. As shown in Table 6, no significant correlations emerged, suggesting that the different LLAMA tests tapped different constructs.

Place Table 6 near here

Table 7 presents the descriptive statistics for the LLAMA aptitude scores by group. A series of independent-samples *t* tests found no significant difference between the two groups on any of the LLAMA test components: LLAMA D *t* = 0.41, *p* = .68, *d* = 0.11; LLAMA E *t* = 0.04, *p* = .97, *d* = 0.01; and LLAMA F *t* < 0.01, *p* = 1.00, *d* = 0.00). Thus, the two groups, overall, had similar aptitude profiles, excluding the possibility that differences in aptitude between the groups might account for the differential gains observed in the two groups. It is also worth noting that the standard deviations were large for both groups, indicating considerable within-group variation in aptitude scores. This justified our decision to include aptitude as a fixed effect in further analyses.

Place Table 7 near here

Multiple regression analyses were conducted to examine whether aptitude accounted for the extent to which participants showed development in the use of the target structure on the oral and written production tests as a result of completing communicative tasks and receiving feedback (a correlation matrix summarizing the relationships between the LLAMA and gain scores is also provided in Appendix S2 in the Supporting Information online). The predictors in the analyses for both tests were the pretest scores and the LLAMA D, LLAMA E, and LLAMA F aptitude scores. The pretest scores served as the covariate, and the aptitude scores were the predictors of interest. None of the aptitude measures emerged as a significant predictor (see Table 8).

Place Table 8 near here

To test the extent to which aptitude explained participants’ development on the EI test, we carried out another mixed-effects regression analysis. The dependent variable was the EI score, and the fixed effects were participants’ pretest scores and the LLAMA D, LLAMA E, and LLAMA F scores. Participant and item were added as random effects to each model. The pretest scores served as the covariate, and our predictors of interest were the aptitude scores. Only LLAMA E emerged as a significant predictor in the analyses (see Table 9). Participants with higher LLAMA E scores performed better on the EI test, OR = 1.02, $CI_{.95}$ = [1.00, 1.03]. Those who achieved one point higher on the LLAMA E test were 2% more likely to score a point higher on the EI test. The LLAMA D scores, OR = 1.01, $CI_{.95}$ = [0.99, 1.04], and the LLAMA F scores, OR = 1.00, $CI_{.95}$ = [0.99, 1.02], were not found to be significant predictors of participants' development in the use of the linguistic target. In sum, only the LLAMA E scores accounted for the extent to which participants benefited from the treatment, and this finding only surfaced on the EI test. Participants who scored higher on the LLAMA E test showed greater development on the EI test.

Place Table 9 near here

Research Question 3: Relationship Between Aptitude and L2 Development With Task Complexity Used As a Moderator

The third research question asked the extent to which task complexity influenced the relationship between L2 aptitude and development in the knowledge of the target feature when learners received recasts during task completion. To address this research question, we ran separate multiple regression analyses for the oral and written production tests. The predictors in the models were the pretest scores, task complexity, the three LLAMA scores, and the interactions between task complexity and the LLAMA scores. The pretest scores served as the covariate, and the predictors of interest were the interactions. As shown in Table 10, the analyses yielded a significant interaction between task complexity and LLAMA D.

Place Table 10 near here

To explore the interaction effects, follow-up regression analyses were conducted for the two experimental groups separately using the pretest scores and LLAMA D as predictors. For both the written and oral production data, LLAMA D emerged as a significant predictor of development for the decision-making group in the oral production test, *Estimate* = 0.72, *SE* = .25, *t* = 2.90, *p* < .01, R^2 = .19, and in the written production test, *Estimate* = 1.10, *SE* = .40, *t* = 2.73, *p* = .01, R^2 = .20, but not for the information transmission group in the oral production test, *Estimate* = –.45, *SE* = .48, *t* = –0.93, *p* = .36, R^2 = .03, or the written production test, *Estimate* = –0.72, *SE* = .74, *t* = –0.98, *p* = .34, R^2 = .03. That is, participants with higher LLAMA D scores showed greater gains on both the oral and written production tests in the decision-making group, but LLAMA D made no difference in gains in the information transmission group. The effect sizes were in the small range.

Next, to test whether task complexity moderated the extent to which aptitude predicted participants' development on the EI test, a mixed-effects logistic regression analysis was conducted. The dependent variable was the EI score, the fixed effects were participants' pretest scores, task complexity, the LLAMA scores (LLAMA D, LLAMA E, or LLAMA F), and the interactions between task complexity and the LLAMA scores. Participant and item served as random effects in the model. The pretest scores served as the covariate, and our predictors of interest were the interactions. As shown in Table 11, the models did not yield a significant interaction effect.

Place Table 11 near here

Discussion

Task Complexity and L2 Development

The first research question investigated the effects of task complexity on child EFL learners' development in the knowledge of the target structure when they received recasts. For the oral and written production tests, the results revealed that the learners benefited more when they carried out tasks and received recasts in the simple condition involving mere information transmission rather than in more complex, decision-making tasks imposing greater reasoning demands. The effect size for this difference was small, consistent with the small overall effect size found for accuracy in Jackson and Suethanapornkul's (2013) meta-analysis investigating the effects of task complexity on L2 production.

These results reflect the predictions of Skehan's (2009, 2014) limited attentional capacity model. As discussed earlier, this model predicts that greater cognitive demands will put increased pressure on the conceptualizer, leaving fewer attentional resources for linguistic encoding processes. Following Skehan, learners under the complex condition likely had less capacity left to pay attention and process feedback focusing on linguistic errors, given the increased effort the task required in terms of conceptualization. In other words, it seems that because of the need to devote greater attention to the communicative demands of the task, learners had few attentional resources available to allocate to the target structure, a communicatively redundant element not needed for successfully completing the task. On the other hand, under the simple condition, given that learners had to pay less attention to conceptualization, learners likely had more cognitive capacity to notice and process feedback focusing on a nonsalient grammatical structure.

The findings for the first research question are also in line with those of Révész et al. (2014). Interestingly, both the child L2 learners of the current study and the adult learners in Révész et al. benefitted more from recasts under the simple than the complex task conditions. Similar to the current study, Révész et al. employed simple and complex monologic tasks during which the participants received a recast when they produced a nontarget structure. The results obtained here and in Révész et al., however, run counter to those of Baralt (2013), who found that oral face-to-face recasts were more effective under complex interactive dialogic tasks than simple ones. Further research is needed to illuminate possible effects of task design (e.g., monologic vs. dialogic tasks) on the efficacy of feedback supplied during cognitively simple and complex tasks.

It should also be noted that although the simple condition was found to be more beneficial than the complex condition, both groups exhibited considerable L2 gains despite the short intervention. This was probably due to the manner in which recasts were provided. As explained in the methodology section, the recasts in the present study could be classified as explicit (Sheen, 2006), reduced (Lyster, 1998), focused on one change (Lyster, 1998), entailing the correction in utterance-final position (Bardovi-Harlig, 1987), and supplied in an intensive and focused manner (Han, 2002).

Finally, it is worth discussing why participants (regardless of group assignment) improved the most on the written production test but showed somewhat less improvement on the oral production test and demonstrated little progress on the EI test. The written production test, unlike the oral production and the EI tests, allowed for production of the target structure under less time pressure, thereby providing more opportunities for the deployment of declarative knowledge. As declarative knowledge is amenable to transfer across modalities (Anderson, 1993), learners probably experienced no problem accessing the declarative knowledge they had obtained through exposure to oral input when engaged in a task in the written mode. Due to the short intervention, however, it is likely that the treatment did not enable participants to achieve the stage of automatic production, which would have allowed them to perform better on the oral tests that imposed more time pressure.

Aptitude as a Predictor of L2 Development

The second research question explored the extent to which aptitude predicted child EFL learners' knowledge of the target feature when they received recasts. The study demonstrated that better phonetic coding ability, as assessed by the LLAMA E test, was found to be associated with higher scores on the EI test. In particular, participants who scored one point higher on the LLAMA E test were 2% more likely to achieve a point higher on the EI test. This relatively small effect size is consistent with the results of Li's (2016) meta-analysis, which yielded only a moderate overall association between aptitude and L2 grammar learning (*r* = .31).

The results obtained here for the LLAMA E test are not unexpected. The LLAMA E test assessed learners' phonetic coding ability, which enables learners to divide words into phonetic units and analyze how these relate to symbolic units. Participants engaged in similar processes when processing recasts during the treatment and when they received oral input on the EI test. Learners needed to divide words into phonetic units and analyze the meanings associated with these units. Thus, participants with higher LLAMA E scores were probably better able to decode recasts, recognize the sound patterns associated with the present third person singular morpheme and associate this form with its function. This ability also likely helped them process the oral input on the EI test more successfully and, hence, to cope with the demands of the task (i.e., comprehending oral input and producing output). In sum, given the overlaps in the constructs measured by LLAMA E, the EI test, and the nature of the treatment, it appears reasonable that a positive relationship between this aptitude component and L2 development, albeit relatively small in size, would be found.

Grañena's (2013) observations regarding the potential relationships between the LLAMA test components and implicit versus explicit learning also appear pertinent to the findings obtained here. Grañena argued that the LLAMA E test might be more sensitive to tapping aptitude components involved in explicit learning (learning via deduction of rules), whereas the learning conditions created by the LLAMA D test resemble implicit learning environments (learning through exposure) to a greater extent. Following this line of thought, we would expect that those with higher LLAMA D rather than LLAMA E scores would achieve greater gains on the EI test, given that the EI test is traditionally regarded as a test of implicit knowledge (e.g., Ellis, 2005; see, however, Suzuki & DeKeyser, 2015). A possible way to explain this seemingly contradictory finding is that the EI test, as argued by Suzuki and

DeKeyser (2015), may elicit, at least in part, the use of automatized explicit knowledge. Considering that the participants in the present study were from a foreign language environment that had trained them to analyze language explicitly, it is not unlikely that those with high LLAMA E scores were better able to develop their explicit, declarative knowledge through exposure to explicit recasts. In turn, the increased explicit knowledge that high LLAMA E students had gained might have helped them to analyze the oral input of the EI test and achieve better scores.

Considering that the participants of the present study were low-proficiency learners, the facilitative role of LLAMA E is also in line with Skehan’s (2002) suggestion that phonetic coding is expected to assist L2 outcomes in the first stages of L2 development. A role for phonetic coding ability, as measured by LLAMA E, has been attested in previous empirical studies as well. Saito (2017) found a positive relationship between the ability to associate sounds with symbols and morphological accuracy. Likewise, Yilmaz and Koylu (2016) showed that phonetic coding ability was related to the extent to which learners benefited from feedback. It is worth noting that Yilmaz and Koylu’s experiment and the present research both employed explicit forms of feedback. Thus, the findings of the two studies are consistent with Grañena’s (2013) proposal that LLAMA E might be implicated in explicit learning.

Last but not least, it is worth discussing why the LLAMA F scores failed to emerge as predictors of learner gains, either for the two groups combined or one of the groups (as LLAMA D did). This finding might be related to the nature of the target structure and/or the type of feedback supplied in the study. Although the morpheme *–s* is associated with several meanings which might pose difficulty in the acquisition of the third person *–s*, the rule associated with the present third person singular is relatively simple (i.e., add a morpheme to the verb). Thus, high grammatical inferencing ability might not have played a crucial role in figuring out the linguistic pattern in the present study. As Saito (2017, p. 670) argued, analytic ability might be more instrumental in facilitating “more diverse, sophisticated, and complex lexicogrammar usage.” In line with this reasoning, Suzuki and DeKeyser (2017b) found that LLAMA F was a significant predictor of developing automatized explicit knowledge of complex Japanese constructions. Another possible reason for the lack of relationship between LLAMA F and learners’ gains might be associated with the nature of the recasts provided. The explicit, partial recasts may have helped learners notice the nonsalient form, thereby neutralizing the role of analytic ability. It is for further research to explore whether analytic ability is of greater importance if participants are exposed to more implicit recasts.

Relationship Between Aptitude and L2 Development: Task Complexity as a Moderator

The third research question delved into the impact of task complexity on the relationship between L2 aptitude and development in child L2 learners’ knowledge of the target structrure when they received recasts. No link was found between aptitude and L2 development under the simple task condition, whereas when learners carried out complex tasks, higher LLAMA D was associated with greater L2 outcomes on the oral and written production tests. The effect size, again, was found to be small, consistent with the results of Li’s (2016) meta-analysis.

From a theoretical perspective, our findings confirm Robinson’s (2011) proposal that individual differences in cognitive abilities will be related to learners’ performance and development when tasks pose greater cognitive demands. The results are also well aligned with Kim et al. (2015), who found that learners with better working memory benefitted the most from recasts when engaged in cognitively complex tasks. From a pedagogical perspective, our study suggests that carrying out tasks with low cognitive demands might minimize the extent to which L2 aptitude predicts learners’ benefits from feedback. Previous research has also shown that learning conditions may influence whether aptitude accounts for

L2 gains. For example, explicit instruction was found to compensate for low aptitude in Erlam (2005).

Albeit only observed for the complex task version, the positive relationship found between LLAMA D and development in the use of the present third person singular form is not surprising. When taking the LLAMA D test and when receiving recasts, the participants needed to process sounds and retain them in long-term memory. This finding also accords with Meara's (2005) prediction that the ability to recognize patterns in spoken language, which LLAMA D intends to measure, should assist learners in recognizing morphological variation in languages. It is worth noting, however, that Saito (2017) found no correlation between morphological accuracy during oral production and learners' LLAMA D scores. This suggests that the extent to which sound recognition ability plays a role in L2 development may depend on the nature of the grammatical structure. Unlike in the present study, Saito operationalized accurate use of morphology as a ratio of morphological errors in a wide range of constructions, including less salient (plural, subject–verb agreement) and more salient (modality, aspect) features. In light of this, one possible explanation for the discrepancy in the findings of the two studies may be that the ability to recognize sound sequences may facilitate the acquisition of morphemes realized through redundant, nonsalient forms (e.g., the present third person singular) but may be less helpful in developing knowledge of more salient grammatical constructions. In Saito's work, combining more and less salient features in one accuracy ratio might have concealed a relationship between the LLAMA D scores and grammatical accuracy ratings.

It is also worthwhile to consider why the positive role of LLAMA D in the complex task group's performance was only observed on the oral and written production tests. Invoking the notion of transfer appropriate processing may be one way of explaining this finding. According to this principle, learners will better transfer and remember what they have learned “if the cognitive processes that are active during learning are similar to those that are active during retrieval” (Lightbown, 2007, p. 27). In light of the transfer appropriate processing principle, then, the additional learning gains that participants were able to accrue due to their superior ability to recognize sounds during the treatment might have been easier to deploy on the oral and written production tests. This was probably because the completion of these tests, as compared to the EI test, involved cognitive processes more similar to the ones in which participants engaged during the treatment.

Limitations and Future Research

There are several limitations to the present study that should be acknowledged, and these should be considered in future research. One shortcoming is that only tests that required learners to produce the L2 were administered. Future studies could additionally utilize comprehension-based outcome measures (e.g., a grammaticality judgment task) in order to obtain a more complete picture of learners’ development. The study would also have benefited from examining the longer-term effects of the treatment. However, using a delayed posttest would not have generated valid results in the present study because the researchers could not control for exposure to the target structure between the immediate and the delayed posttest. Another weakness of the study is that individual differences in cognitive abilities were only captured by an aptitude test. A follow-up study could, besides indices of aptitude, involve measures of working memory and attentional control to provide a more comprehensive account of the role of cognitive individual differences in child L2 learners’ ability to benefit from recasts and task-based interaction. A further limitation of the study is that we utilized only a single type of grammatical structure and one type of recasts (interrogative, partial recasts). Future studies could explore other target constructions (e.g., more salient grammatical features) and the effectiveness of different focus on form interventions. Both direct and conceptual replications of the current study are also warranted (see Marsden, Morgan-Short, Thompson, & Abugaber, 2018), for example, using other types of task manipulations, adult learners, and participants from different L1 backgrounds or proficiency levels.

Conclusion

This study aimed to investigate the extent to which (a) task complexity influences L2 development resulting from task-based interaction, (b) aptitude predicts L2 gains, and (c) task complexity influences the relationship between aptitude and L2 outcomes when recasts are provided. We focused on child EFL learners, an underresearched population, and the third person singular *–s* structure, a linguistic feature that involves several challenges in L2 acquisition. In line with Skehan's (2009, 2014) prediction, we found that those learners who completed less cognitively demanding tasks improved their knowledge of the present third person singular to a greater extent than learners who engaged in tasks with greater cognitive demands. Interestingly, however, high-aptitude learners, especially those with superior sound recognition ability, were slightly better able to compensate for the increased demands posed by the complex tasks than their low-aptitude counterparts, reflecting Robinson's (2011) predictions regarding the relationship between aptitude and task complexity. To put it differently, our results suggest that low-complexity tasks have the capacity to minimize the degree to which learner differences in L2 aptitude predict development in task-based contexts when feedback is available. Thus, a tentative pedagogical implication, if this finding is replicated in other studies, is that the use of less complex tasks might be more beneficial for developing the grammatical knowledge for language learners in an entire language class, unless individual learners are assigned to instruction types based on their L2 aptitude profiles.

Final revised version accepted 13 June 2019

References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*, 689–725. https://doi.org/10.1017/S0272263113000429

Bardovi-Harlig, K. (1987). Markedness and salience in second-language acquisition. *Language Learning*, *37*, 385–407. https://doi.org/10.1111/j.1467-1770.1987.tb00577.x

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://www.jstatsoft.org/article/view/v067i01

Bokander, L., & Bylund, E. (2019). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*. Published online https://doi.org/10.1111/lang.12368

Carroll, J. B. (1965). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training, research, and education* (pp. 87–136). New York, NY: Wiley.

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

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test.* San Antonio, TX: Psychological Corporation.Cobb, T. (2016). VocabProfiler [Computer software]. http://www.lextutor.ca/vp/eng

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

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/S0272263105050096

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, *51*, 281–318. https://doi.org/10.1111/1467-9922.00156

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28,* 339–368. https://doi.org/10.1017/S0272263106060141

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9,* 147–171. https://doi.org/10.1191/1362168805lr161oa

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27,* 464–491. https://doi.org/10.1093/applin/aml001

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39,* 175–191. https://doi.org/10.1037/0033-2909.112.1.155

Grañena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test. In G. Grañena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–130). Amsterdam, Netherlands: John Benjamins.

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The canal-F theory and test. *The Modern Language Journal*, *84,* 390–405. https://doi.org/10.1111/0026-7902.00076

Han, Z. (2002). A study of the impact of recasts on tense consistency in L2 output. *TESOL Quarterly*, *36,* 543–572. https://doi.org/10.2307/3588240

Han, Z. (2013). Forty years later: Updating the fossilization hypothesis. *Language Teaching, 46*, 133–171. https://doi.org/10.1017/S0261444812000511

Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition, 19*, 379–400. https://doi.org/10.1017/S0272263197003045

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range and frequency programs [Computer software]. http://www.victoria.ac.nz/lals/resources/range.aspx

Jackson, D. O., Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, *63*, 330–367. https://doi.org/10.1111/lang.12008

Keating, G. D., & Jegerski, J. (2014). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, *37*, 1–32. https://doi.org/https://doi.org/10.1017/S0272263114000187

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition*, *37*, 549–581. https://doi.org/10.1017/S0272263114000618

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

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal*, *97*, 634–654. https://doi.org/10.1111/j.1540-4781.2013.12030.x

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 801–842. https://doi.org/10.1017/S027226311500042X

Lightbown, P. M. (2007). Transfer appropriate processing as a model for classroom second language acquisition. In Z.-H. Han (Ed.), *Understanding second language process* (pp. 27–44). Clevedon, UK: Multilingual Matters.

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

Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching, 51*, 285–329. https://doi.org/10.1017/S0261444818000125

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In W. C. Ritchie & B. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York, NY: Academic Press.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Hoboken, NJ: Wiley Blackwell.

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition, 20*, 51–81. https://doi.org/10.1017/S027226319800103X

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19*, 37–66. https://doi.org/10.1017/S0272263197001034

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32*, 265–302. https://doi.org/10.1017/S0272263109990520

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. https://doi.org/10.1111/lang.12286

Meara, P. (2005). *Llama language aptitude tests: The manual.* Swansea, UK: Lognostics.

Nassaji, H. (2009). Effects of recasts and elicitations in dyadic interaction and the role of feedback explicitness. *Language Learning, 59*, 411–452. https://doi.org/10.1111/j.1467-9922.2009.00511.x

Norris, J. M. (2010, September). *Understanding instructed SLA: Constructs, contexts, and consequences*. Plenary address delivered at the annual conference of the European Second Language Association (EUROSLA), Reggio Emilia, Italy.

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

R Core Team. (2018). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org

Ranta, L. (2002). The role of learners’ language analytic ability in the communicative classroom. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 159–180). Amsterdam, Netherlands: John Benjamins.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*, 437–470. https://doi.org/10.1017/S0272263109090366

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, *35*, 87–92. https://doi.org/10.1093/applin/amt039

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning*, *64*, 615–650. https://doi.org/10.1111/lang.12061

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57. https://doi.org/10.1093/applin/22.1.27

Robinson, P. (2003). The cognitive hypothesis, task design, and adult task-based language learning. *Studies in Second Language Acquisition*, *21*, 45–105. https://doi.org/http://hdl.handle.net/10125/40656

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning and performance. In P. Robinson (Ed.), *Second language task complexity:*

Researching the cognition hypothesis of language learning and performance (pp. 3–38). Amsterdam, Netherlands: John Benjamins.

Saito, K. (2017). Effects of sound, vocabulary, and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning*, *67*, 665–693. https://doi.org/10.1111/lang.12244

Sato, M., & Loewen, S. (2018). Metacognitive instruction enhances the effectiveness of corrective feedback: Variable effects of feedback types and linguistic targets. *Language Learning*, *68*, 507–545. https://doi.org/10.1111/lang.12283

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research*, *10*, 361–392. https://doi.org/10.1191/1362168806lr203oa

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 301–322). Oxford, UK: Oxford University Press.

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

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–93). Amsterdam, Netherlands: John Benjamins.

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

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp.

211–260). Amsterdam, Netherlands: John Benjamins.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Grañena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 17–40). Amsterdam, Netherlands: John Benjamins.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics, 38,* 1229–1261. https://doi.org/10.1017/S014271641700011X

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65,* 860–895. https://doi.org/10.1111/lang.12138

Suzuki, Y., & DeKeyser, R. (2017a). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics, 38,* 27–56. https://doi.org/10.1017/S0142716416000084

Suzuki, Y., & DeKeyser, R. (2017b).The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67,* 747–790. https://doi.org/10.1111/lang.12241

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

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Westport, CT: Ablex.

Yilmaz, Y. (2012). The relative effects of explicit correction and recasts on two target structures via two communication modes. *Language Learning, 62,* 1134–1169. https://doi.org/10.1111/j.1467-9922.2012.00726.x

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics*, *34*, 344–368. https://doi.org/10.1093/applin/ams044

Yilmaz, Y., & Grañena, G. (2016). The role of cognitive aptitudes for explicit language learning in the relative effects of explicit and implicit feedback. *Bilingualism: Language and Cognition*, *19*, 147–161. https://doi.org/10.1017/S136672891400090X

Yilmaz, Y., & Koylu, Y. (2016). The interaction between feedback exposure condition and phonetic coding ability. In G. Grañena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 303–326). Amsterdam, Netherlands: John Benjamins.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher’s website:

Appendix S1. Additional Results From Linear Mixed-Effects Models.

Appendix S2. Correlations Between L2 Development and Aptitude.

Table 1 Descriptive statistics for the perceived mental effort scale

		Treatment task 1			Treatment task 2		
Group	*n*	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Information transmission	30	3.83	1.74	[3.18, 4.48]	3.43	1.71	[2.79, 4.07]
Decision making	30	5.93	1.91	[5.22, 6.45]	5.23	2.35	[4.35, 6.11]

Note. CI = confidence interval.

Table 2 Descriptive statistics for obligatory contexts, number of recasts, and successful uptake (percent) per group

Group	*n*	*M*	*SD*	95% CI
Obligatory contexts				
Information transmission	30	30.66	5.01	[28.79, 32.53]
Decision making	30	32.66	7.71	[29.78, 35.54]
Number of recasts				
Information transmission	30	15.40	8.41	[12.26, 18.54]
Decision making	30	19.73	9.28	[16.27, 23.20]
Successful uptake				
Information transmission	30	50.47	36.65	[36.78, 64.16]
Decision making	30	41.37	34.88	[28.35, 54.40]

Note. CI = confidence interval.

Table 3 Descriptive statistics for pretest and posttest scores on the assessment tasks per group

		Pretest			Posttest		
Group	*n*	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Oral production							
Information transmission	30	3.99	8.19	[0.93, 7.06]	33.61	30.74	[22.14, 45.09]
Decision making	30	3.75	6.91	[1.18, 6.34]	17.31	26.54	[7.40, 27.22]
Written production							
Information transmission	30	4.10	9.22	[0.66, 7.55]	57.05	44.32	[40.50, 73.60]
Decision making	30	4.93	9.30	[1.46, 8.41]	26.77	39.50	[12.02, 41.52]
Elicited imitation							
Information transmission	30	4.26	4.25	[2.67, 5.85]	6.80	5.30	[4.81, 8.78]
Decision making	30	4.00	4.10	[2.46, 5.53]	6.00	5.20	[4.05, 7.94]

Note. The total score was 24 for the EI test. CI = confidence interval.

Table 4 Results for the linear regression models examining the effects of task complexity on the oral and written production tests

Factor	*Est*	*SE*	*t*	*p*	R^2
Oral production					
Intercept	44.01	11.17	3.94	< .01	
Pretest	1.39	0.47	2.99	< .01	.13
Task complexity	–15.97	6.95	–2.30	.03	.08
Overall R^2					.20
Oral_Post ~ 1 + Oral_Pre + Task_complexity					
Written production					
Intercept	85.63	17.28	4.95	< .01	
Pretest	0.52	0.60	0.87	.39	< .01
Task complexity	–30.71	10.87	–2.83	< .01	.12
Overall R^2					.13
Written_Post ~ 1 + Written_Pre + Task_complexity					

Table 5 Results for the linear mixed-effects models examining the effects of task complexity on the elicited imitation test

	Fixed effects				Random effects	
Factor	*Est*	*SE*	*z*	*p*	Factor	*SD*
Intercept	–1.46	0.51	–2.83	< .01	Participant	1.10
Pretest	2.53	0.20	12.94	< .01	Item	0.21
Task complexity	–0.25	0.33	–0.78	.44		
EI_Post ~ 1 + EI_Pre + Task_comp + (1|Participant) + (1|Item)						

Table 6 Correlations among LLAMA scores

	LLAMA D	LLAMA E
LLAMA E	.082 (.534)	
LLAMA F	–.222 (.089)	.124 (.346)

Note. *p* values appear in parentheses.

Table 7 Descriptive statistics for L2 aptitude scores per group

Component	Group	*n*	*M*	*SD*	95% CI
LLAMA D	Information transmission	30	23.66	11.36	[19.42, 27.91]
	Decision making	30	22.16	16.33	[16.07, 28.26]
LLAMA E	Information transmission	30	42.66	26.25	[32.86, 52.47]
	Decision making	30	43.00	31.85	[31.10, 54.90]
LLAMA F	Information transmission	30	21.66	20.35	[14.07, 29.27]
	Decision making	30	21.66	19.84	[14.26, 29.08]

Note. The total score was 75 for LLAMA D and 100 for LLAMA E and F. CI = confidence interval.

Table 8 Results for the multiple regression analyses examining the relationship between aptitude and gains on the oral and written production tests

Factor	*Est*	*SE*	*t*	*p*	R^2
Oral production					
Intercept	6.02	9.99	0.60	.55	
Pretest	1.42	0.48	2.94	< .01	.13
LLAMA D	0.35	0.27	1.30	.20	.03
LLAMA E	0.14	0.13	1.09	.28	.02
LLAMA F	< 0.01	0.19	< 0.01	1.00	< .01
Overall R^2					.18
Oral_Post ~ 1 + Oral_Pre + LLAMA_D + LLAMA_E + LLAMA_F					
Written production					
Intercept	17.62	15.65	1.13	.27	
Pretest	0.37	0.64	0.57	.57	.01
LLAMA D	0.60	0.43	1.39	.17	.04
LLAMA E	0.15	0.20	0.71	.48	.01
LLAMA F	0.12	0.31	0.40	.69	< .01
Overall R^2					.06
Writ_Post ~ 1 + Writ_Pre + LLAMA_D + LLAMA_E + LLAMA_F					

Table 9 Results for the linear mixed-effects models examining the relationship between aptitude and gains on the elicited imitation test

	Fixed effects				Random effects	
Factor	*Est*	*SE*	*z*	*p*	Factor	*SD*
Intercept	–2.85	0.43	–6.61	< .01	Participant	0.99
Pretest	2.52	0.19	12.98	< .01	Item	0.21
LLAMA D	0.01	0.01	1.19	.23		
LLAMA E	0.01	< 0.01	2.78	< .01		
LLAMA F	< 0.01	0.01	0.36	.72		

EI_Post ~ 1 + EI_Pre + LLAMA_D + LLAMA_E + LLAMA_F + (1|Participant) + (1|Item)

Table 10 Results for the multiple regression models examining the effect of task complexity on the relationship between aptitude and development in the oral and written production tests

Factor	*Est*	*SE*	*t*	*p*	R^2
Oral production					
Intercept	99.61	36.18	2.75	< .01	
Pretest	1.40	0.46	3.06	< .01	.13
Task complexity	–55.29	20.36	–2.72	< .01	.08
LLAMA D	–1.82	0.98	–1.86	.07	.03
LLAMA E	–0.19	0.41	–0.46	.65	.02
LLAMA F	–0.25	0.58	–0.44	.67	< .01
Task complexity × LLAMA D	1.25	0.56	2.25	.03	.05
Task complexity × LLAMA E	0.21	0.25	0.84	.41	.01
Task complexity × LLAMA F	0.09	0.37	0.23	.82	< .01
Overall R^2					.33
*Oral_Prod_Post ~ 1 + Oral_Prod_Pre + LLAMA_D * Task_comp + LLAMA_E * Task_comp + LLAMA_F * Task_comp*					
Written production					
Intercept	202.40	55.97	3.62	< .01	
Pretest	0.61	0.59	1.03	.31	< .01
Task complexity	–109.98	31.72	–3.47	< .01	.12
LLAMA D	–2.96	1.51	–1.95	.06	.04
LLAMA E	–0.73	0.63	–1.15	.26	.01
LLAMA F	–0.81	0.91	–0.89	.38	< .01
Task complexity × LLAMA D	2.00	0.86	2.34	.02	.07
Task complexity × LLAMA E	0.54	0.39	1.41	.17	.03

Task complexity × LLAMA F	0.46	0.57	0.80	.43	< .01
Overall R^2					.28
*Written_Prod_Post ~ 1 + Written_Prod_Pre + LLAMA_D * Task_comp + LLAMA_E * Task_comp + LLAMA_F * Task_comp*					

Table 11 Results for the linear mixed-effects models examining the effects of task complexity on the relationship between aptitude and development on the EI test

	Fixed effects				Random effects	
Factor	*Est*	*SE*	*z*	*p*	Factor	*SD*
Intercept	–0.99	0.47	–2.12	.03	Participant	0.97
Pretest	0.95	0.07	12.99	< .01	Item	0.21
Task complexity	0.18	0.59	0.30	.76		
LLAMA D	–0.27	0.30	–0.92	.36		
LLAMA E	–0.01	0.51	–0.01	.99		
LLAMA F	0.10	0.50	0.21	.83		
Task complexity × LLAMA D	< –0.01	0.34	–0.01	.99		
Task complexity × LLAMA E	0.28	0.32	0.90	.37		
Task complexity × LLAMA F	–0.06	0.32	–0.18	.86		

*EI_Post ~ 1 + EI_Pre + Task_comp * LLAMA_D + Task_comp * LLAMA_E + Task_comp * LLAMA_F + (1|Participant) + (1|Item)*

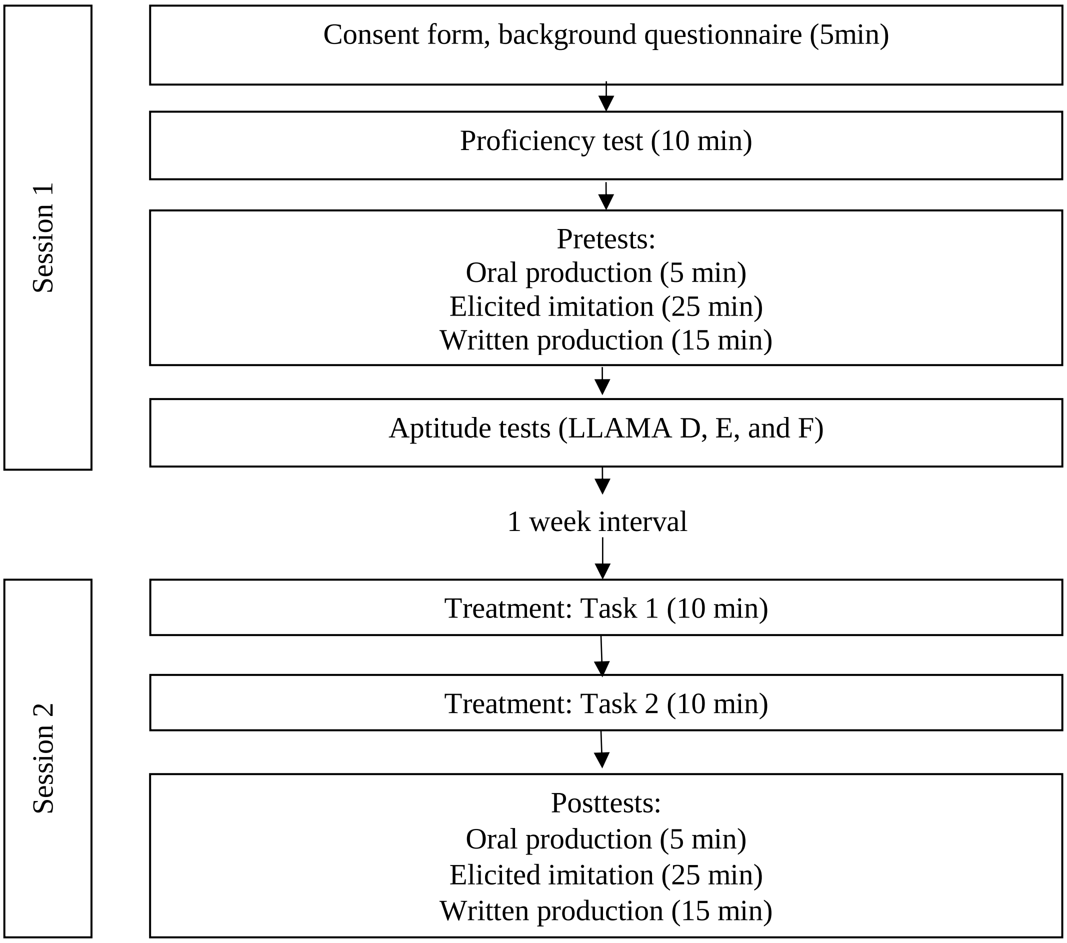
Session 1
Consent form, background questionnaire (5min)
Proficiency test (10 min)
Pretests:
Oral production (5 min)
Elicited imitation (25 min)
Written production (15 min)
Aptitude tests (LLAMA D, E, and F)
1 week interval
Session 2
Treatment: Task 1 (10 min)
Treatment: Task 2 (10 min)
Posttests:
Oral production (5 min)
Elicited imitation (25 min)
Written production (15 min)

Figure 1 Experimental schedule.