COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–26
Copyright © 2015 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12257

The Relationship Between Artificial and Second Language Learning

Marc Ettlinger,[a] Kara Morgan-Short,[b,c] Mandy Faretta-Stutenberg,[d] Patrick C.M. Wong[e,f]

[a]*Research Service, Department of Veterans Affairs, Northern California Health Care System*
[b]*Department of Hispanic and Italian Studies, University of Illinois at Chicago*
[c]*Department of Psychology, University of Illinois at Chicago*
[d]*Department of Foreign Languages and Literatures, Northern Illinois University*
[e]*Department of Linguistics and Modern Languages and the CUHK-Utrecht University Joint Centre for Language, Mind, and Brain, The Chinese University of Hong Kong*
[f]*The Roxelyn and Richard Pepper Department of Communication Sciences and Disorders and the Department of Otolaryngology—Head and Neck Surgery, Feinberg School of Medicine, Northwestern University*

Received 26 December 2012; received in revised form 23 December 2014; accepted 2 March 2015

Abstract

Artificial language learning (ALL) experiments have become an important tool in exploring principles of language and language learning. A persistent question in all of this work, however, is whether ALL engages the linguistic system and whether ALL studies are ecologically valid assessments of natural language ability. In the present study, we considered these questions by examining the relationship between performance in an ALL task and second language learning ability. Participants enrolled in a Spanish language class were evaluated using a number of different measures of Spanish ability and classroom performance, which was compared to IQ and a number of different measures of ALL performance. The results show that success in ALL experiments, particularly more complex artificial languages, correlates positively with indices of L2 learning even after controlling for IQ. These findings provide a key link between studies involving ALL and our understanding of second language learning in the classroom.

Keywords: Second language learning; Artificial grammar; Artificial language learning; Classroom learning; Language pedagogy; Language learning

Correspondence should be sent to Marc Ettlinger, Research Service, 150 Muir Road, Martinez, CA 94553. E-mail: ettlinger@gmail.com

1. Introduction

Research on language has seen a remarkable rise in the use of artificial language learning (ALL) experiments since their introduction nearly 90 years ago (Esper, 1925). The term "artificial language learning" generally refers to an experimental paradigm where participants learn a language, or language-like system, in a lab setting and are then tested on what they learned. The reasons for using artificial languages are diverse and include allowing for a controlled exploration of the principles of universal grammar (Culbertson, Smolensky, & Legendre, 2012; Ettlinger, Bradlow, & Wong, 2014; Finley & Badecker, 2009), of how domain-general cognitive mechanisms might support language and language learning (Saffran, Aslin, & Newport, 1996), principles of language change (Esper, 1925), of the relationship between first and second and adult and child language learning (Finn & Hudson Kam, 2008), and of the processes involved in second language learning (Friederici, Steinhauer, & Pfeifer, 2002; Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter, & Wong, 2014; Morgan-Short, Sanz, Steinhauer, & Ullman, 2010).

Despite the recent ubiquity of this research, there has been little work that has established a clear relationship between performance in lab-based ALL experiments and learning a natural language in an ecologically valid environment, which is the primary question addressed by this study. In considering this question, two related questions also arise: Given interest in a possible dissociation between language learning and general cognitive capabilities (Hauser, Chomsky, & Fitch, 2002; Pinker & Jackendoff, 2009), does a relationship between natural and artificial language learning still hold after controlling for general intelligence? Does this relationship differ for different measures of ALL?

In this study, we seek to bridge this gap between artificial and natural language learning research by examining the performance of adult learners in both second language (L2) and ALL environments. We elicited participation from a cohort of students enrolled in a Spanish language class for our ALL study. We obtained a number of different measures of their classroom performance, their Spanish ability, and general intelligence. The ALL task included a number of different measures in a single task to represent some of the different types of artificial languages that have been used in other studies. Our analysis also examined the relationship between artificial and second language learning measures and a measure of general intelligence, IQ.

1.1. Artificial language learning

Researchers generally use the term "artificial language learning" to refer to an experimental paradigm where participants learn a language, or language-like system, in a laboratory setting and are then tested on what they learned (e.g., Friederici et al., 2002; Gomez & Gerken, 2000). ALL studies, however, take many different forms and go by many different names. The original terminology of "artificial linguistic system" (Esper, 1925, p. 1) has been expanded to include *artificial grammar learning,* which generally focuses on the combinatoric aspects of language, often in the absence of meaning (Reber,

1967; Saffran et al., 1996); *miniature language learning*, which generally refers to learning aspects of an invented or part of a natural language, for example, a determiner system with semantics (Hudson Kam & Newport, 2005) or the case and classifier system of Japanese (Mueller, Hahne, Fujii, & Friederici, 2005); *miniature artificial language*, which refers to made-up grammatical categories and word-order rules (Braine, 1963); and *semi-artificial language learning*, which generally refers to a portion of a natural language, often modified for experimental purposes (Williams, 2005).

The paradigm of the first ALL experiment (Esper, 1925) would be quite familiar to experimenters today. The study examined biases in language learning by exposing participants to pairings of words with pictures of abstract shapes of different colors. After training, participants were presented with just the abstract pictures, which they then had to name. In this, and many of the original ALL experiments, the experimenters looked at the mistakes participants made as indicative of learning biases. For example, in one experiment in Esper (1925), the words presented were bi-morphemic with the first consonant-vowel-consonant-consonant (CVCC) morpheme representing color and the second VC morpheme representing shape. The most common error was that participants would often re-segment the stimuli into two consonant-vowel-consonant (CVC) morphemes, suggesting that a principle of linguistic change involved the alignment of morpheme breaks with syllable breaks. This finding has also been interpreted to reflect a bias against morphemes with complex codas and without onsets.

The 1960s and 1970s saw a significant shift in the nature of ALL experiments. Exemplified by Reber (1967), which is often cited as the first modern ALL experiment, studies started to focus on the combinatoric elements of language, reflecting a shift in the focus of research on language from the study of language change to syntax and generative grammar (Chomsky, 1957, 1965). In Reber's study, participants were trained on sequences of letters generated by a finite-state grammar without any corresponding meaning associated with the sequences. Participants successfully learned which novel letter sequences were valid with respect to the finite-state grammar and crucially reported no explicit awareness of the rules, suggesting an implicit process enabling the extraction of regularities from an input. A similar method of artificial grammar learning has been used with syllables and transitional probabilities (Saffran et al., 1996), words and different types of grammars (Opitz & Friederici, 2003, 2004), musical notes and finite-state grammars (Tillmann, Bharucha, & Bigand, 2000), and various other permutations of these paradigms and stimuli.

The subsequent decades saw a dramatic rise, not only in the frequency but also in the variety of ALL studies. One recent advance is the use of different participant populations, including children and infants (Gomez & Gerken, 2000; Newport & Aslin, 2004; Saffran et al., 1996), primates (Fitch & Hauser, 2004), and songbirds (Gentner, Fenn, Margoliash, & Nusbaum, 2006). ALL paradigms are also used in brain imaging studies (Friederici et al., 2002; McNealy, 2006; Morgan-Short, Steinhauer, Sanz, & Ullman, 2012) to explore the neural bases of language learning. In addition to continued research on the precise syntactic properties that facilitate language learning (Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006; Knowlton & Squire, 1996), ALL studies have explored all levels of linguistic structure from phonetics (Wong et al., 2008) and phonology (Ettlinger et al., 2014;

Finley & Badecker, 2009; Moreton, 2008; Wilson, 2006) to semantics (Mirkovic, Forrest, & Gaskell, 2011; Mori & Moeser, 1983; Petersson, Folia, & Hagoort, 2010) and pragmatics (Galantucci & Garrod, 2011; Nagata, 1987). ALL studies have also been used to explore the degree to which language learning may be explained by domain-general learning mechanisms (Aslin & Newport, 2008; Newport, Hauser, Spaepen, & Aslin, 2004; Saffran et al., 1996). Finally, ALL experiments have recently been incorporated into iterated learning studies where findings come not from seeing whether the languages are learned, but rather, seeing what happens to the artificial languages after several cycles of learning and transmission as a method of exploring principles of language evolution and change (Galantucci, 2005; Kirby, Cornish, & Smith, 2008; Rafferty, Griffiths, & Ettlinger, 2013; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). In addressing these and other issues, artificial languages can be viewed as serving as test-tube models of natural language that allow researchers to examine precise issues about language that are not readily testable with natural language (e.g., Morgan-Short et al., 2012).

1.2. As a predictor of second language acquisition

Despite the use of the ALL paradigm as a way to explore the human language faculty, very little research has explored the relationship between artificial and natural language learning, particularly with respect to second language learning. Indeed, many papers acknowledge the important caveats associated with their findings. For example, Braine (1963) acknowledges that "[a]lthough experiments with artificial languages provide a vehicle for studying learning and generalization processes hypothetically involved in learning the natural language*, they cannot, of course, yield any direct information about how the natural language is actually learned* [emphasis added]" (p. 324). Similarly, Ferman, Olshtain, Schechtman, and Karni (2009) note that "[. . .] one may argue that the simplified language and laboratory conditions afforded in artificial language paradigms may not express the complexity of natural language or of real-life learning conditions. These arguments, however, express a classic dilemma that inevitably arises as the price of experimental control in laboratory research" (p. 387).

A few studies have explored the relationship between artificial and natural language learning indirectly. Some have shown that people with language-related impairments including aphasia (Christiansen, Kelly, Shillcock, & Greenfield, 2010; Dominey, Hoen, Blanc, & Lelekov-Boissard, 2003; Goschke, Friederici, Kotz, & van Kampen, 2001), specific language impairment (Evans, Saffran, & Robe-Torres, 2009), and developmental dyslexia (Pothos & Kirk, 2004) perform worse on ALL tasks than healthy controls. For example, Evans et al. (2009) showed that children with specific language impairment performed worse in an ALL experiment involving the acquisition of transitional probabilities for both syllables and notes. Furthermore, for individuals in both groups, native language receptive vocabulary correlated with their performance in the ALL task. Similarly, Misyak and Christiansen (2012) found a correlation between ALL and some aspects of native language ability, including vocabulary and the comprehension of complex sentences and Misyak, Christiansen, and Tomblin (2010a) found a correlation between learning an

artificial language with non-adjacent dependencies and the processing of long-distance dependencies in natural language.

An indirect relationship between artificial and natural language learning may also be inferred by virtue of the fact that both artificial and second language learning correlate with a third variable, verbal working memory. Examples of the relationship between language learning and working memory are well documented (Michael & Gollan, 2005; Robinson, 2002, 2005a; Williams, 2012) and include both first and second language skill. Examples of research showing a relationship between ALL and working memory include Misyak and Christiansen (2012), where a measure of verbal working memory correlated significantly with the statistical learning of adjacent (r = .46) and non-adjacent (r = .53) transitional probabilities. Thus, artificial and second language learning may be related to each other by virtue of being supported by working memory. Were this to be the case, it raises the question of whether ALL studies tap into language-specific learning abilities or whether ALL studies assess participants' general learning abilities or general intelligence, which in turn play a role in second language learning (Genesee, 1976).

Finally, Robinson (2005b, 2010) directly explored the relationship between artificial and second language learning by comparing performance on two artificial grammar learning tasks with a brief second language learning task. The artificial language component of the study included standard explicit and implicit artificial grammar tasks, which required participants to view and later judge strings of letters generated by an artificial grammar (implicit) or to view a series of letters and choose the letter that best completed the series (explicit). The second language learning task involved exposing the participants to sentences reflecting three different grammatical rules from a natural language (Samoan), and then testing learners on the grammatical rules of the language. In addition, Robinson assessed participants' language learning aptitude, working memory, and intelligence.

No relationship was found between either of the ALL tasks and the natural language learning tasks. In addition, artificial and second language learning tasks correlated with different cognitive abilities: (a) the implicit artificial grammar learning task correlated negatively with IQ, (b) the explicit ALL task correlated positively with aptitude, and (c) the natural language learning task correlated positively with working memory. Robinson suggests that the lack of a relationship between the two learning tasks may be attributed to the fact that the ALL task relied primarily on implicit learning mechanisms and lacked the semantics that are crucially involved in natural second language learning. Similarly, Brooks and Kempe (2013) explored the relationship between learning a small portion of Russian grammar and learning an artificial syntactic grammar using pseudo-words presented auditorily. The results showed that there was no relationship between the auditory sequence learning and L2 learning after controlling for metalinguistic awareness as assessed in a post hoc interview.

The result in these two studies may be limited to the particular artificial grammar learning paradigm, which did not include semantics, and the limited nature of L2 learning assessment. The findings may not necessarily generalize to the relationship between other types of artificial and second language grammar learning or between artificial grammar learning and other aspects of language, such as vocabulary, word segmentation,

phonology and pronunciation, literacy, or any other measures of second language aptitude. Crucially, in these previous studies, the second language learning took place over the course of less than a week, whereas typical second language learning generally occurs over the course of weeks, months, and years.

Thus, previous research on the relationship between artificial and natural language has primarily focused on language disorders or on first language ability, whereas the studies that focused on the relationship between artificial and second language learning showed null results, perhaps due to the limited nature of the experiments used. Indeed, the majority of research in this area has focused on one specific type of ALL, that of artificial *grammar* learning, where no meaning is assigned to the artificial structures being acquired. Limiting research to artificial *grammar* learning studies also fails to provide comparisons of different ALL paradigms.

In this study, we explored (a) the relationship between artificial and second language learning, (b) whether such a relationship still holds after controlling for general intelligence, and finally (c) whether this relationship differs for different measures of ALL. We addressed these issues in the following manner: We used an ALL paradigm that included semantics; we measured participants' IQ to examine the relationship between ALL and L2 controlling for general intelligence; we included a number of different ALL measures including recall versus a simple grammar versus a complex grammar; and we included a more comprehensive assessments of L2 ability. This allowed us to consider which aspects of L2 learning are tapped into by ALL experiments instead of thinking of it as a monolithic cognitive function. Given the diversity of metrics that are used to quantify L2 ability, different measures of ALL may reflect different facets of L2 learning.

Because ALL studies are attempts at simplifying language learning for a laboratory setting, we predicted that the complex ALL measure would most closely correlate with objective measures of L2 ability. By the same token, because classroom performance incorporates a number of different skills, including language learning, homework completion, memorization, test preparation, etc., we predicted that the composite measure of ALL, which includes recall, simple grammar learning and complex learning, would most closely correlate with overall measures of classroom performance. We also predicted that IQ would mediate the relationship between the composite measure of ALL and classroom performance as they both incorporate more general cognitive capabilities beyond just language learning including skills associated with IQ. Conversely, we predicted that IQ would not mediate the relationship between complex ALL and L2 ability, as we hypothesize these measures are indexed more exclusively to language learning.

2. Methods

2.1. Participants

Participants were 44 adults (23 female) enrolled in a fourth semester Spanish language class at a university in Chicago, Illinois, that focused on learning and using

vocabulary, grammar, and culture for communicative purposes. Participants were recruited over two separate semesters with participants receiving monetary compensation. Participants' average age was 21.7 years (SD = 2.9) and the mean age of initial exposure to Spanish was 13.5 years old (SD = 5). None of the participants had more than 5 years of classroom experience with Spanish, though eight participants indicate ages of acquisition of less than 11 years of age based on general exposure to the language. Thirty-two of the 44 participants were monolingual native English speakers aside from their experience with Spanish and none of them were heritage speakers of the language. The 12 bilingual participants had experience with languages other than Spanish (i.e., 6 Gujarati, 2 Tagalog, 1 ASL, 1 Haitian Creole, 1 Hindi, 1 Tamil). None of the languages participants knew shared the properties critical to the morphophonological system of the artificial language that participants learned in the study.

2.2. Instruments

We evaluated participants using measures of ALL skill, measures of Spanish learning skill, and measures of general intelligence. The ALL test made use of a morphophonological grammar learning paradigm that included a semantic component. Participants were tested on both recall of the artificial language and on generalization for two morphophonological processes—simple and complex—to assess ALL ability. The evaluation of Spanish classroom learning incorporated separate measures of classroom performance, subjective teacher assessments, and objective measures of Spanish interpretation and production. The general intelligence assessment included standardized measures of both verbal and non-verbal IQ.

2.2.1. Artificial language learning

The artificial language in this study has previously been used to explore the relationship between language learning and domain-general cognitive abilities (Ettlinger et al., 2014; Wong, Ettlinger, & Zheng, 2013). In this paradigm, participants were trained on a morphophonological system for combining affixes with words to form new words. Participants were tested on the words they were trained on and then tested on their ability to extend the grammar to another set of withheld words.

2.2.1.1. Artificial language stimulus: The language consisted of 30 noun stems and two affixes: a prefix, [ka-], marking the diminutive (e.g., as in English *doggy*) and a suffix, [-il], marking the plural (e.g., *dogs*). The nouns represented 30 different animals and freely combined with the affixes to produce 120 different words.

The phonological inventory consisted of American English consonants and three American English vowels, [i, e, a] each used within a CVC structure to produce 10 unique nouns for each vowel. No English words or Spanish words were used. Given the diversity of other languages known by bilingual participants, words from other languages were not overtly avoided.

The grammar of the language had two word formation rules as depicted in Fig. 1. The SIMPLE type, applicable to *i*-stems and *a*-stems, consisted of concatenating the stems with the suffix [-il] and/or prefix [ka-] without changing any vowels. The COMPLEX type, applicable to *e*-stems, consisted of concatenation plus changing vowels in the stem and affix. More specifically, the changes reflected two processes absent from English or Spanish. First, vowel harmony changed vowels in the suffix so they had the same (jaw) height as the stem vowel (e.g., the plural of [mez], "cat," became [mez-el] "cats" (compare [vab-**il**] "cows")); second, vowel harmony was also triggered by the prefix [ka-], which changed stem vowels to low (e.g., [ka-maz], "little cat"). When combined, they yielded complex *e*-stem words [ka-maz-el] as contrasted with simple *i*-stem words [ka-bis-il]. Vowel harmony is a relatively common phonological phenomenon and is estimated to occur in hundreds of languages (out of ~6,500) around the world (van der Hulst & van de Weijer, 1995). The particular vowel harmony grammatical system used in this study was based on the language Shimakonde (Ettlinger, 2008).

A native English speaker was recorded saying each of the words spoken at a normal rate with English prosody and phonology so as to sound natural and fluent using Praat (Boersma & Weenink, 2005). Each word had a corresponding picture of an easily recognizable animal/set of animals, with the small animal picture being a shrunken, diminutive version of the standard sized picture. All stimulus and test items are shown in Appendix S1.

2.2.1.2. Artificial language learning procedure: Participants were only told that they would be exposed to a language and then tested on what they learned. They were given no instruction on the rules of the language or told that there were even rules to learn. Auditory stimulus was presented over headphones. Visual stimulus (pictures of the words' meaning) was presented on a computer monitor and participants recorded responses on a button box.

Training consisted of passive exposure to word-picture pairings, with no feedback. During the 20 minutes of exposure, each participant was exposed to four repetitions of 12 nouns in all four forms for a total of 192 tokens in random order. Four nouns were complex (/e/), four were simple with/i/-stems and four were simple with/a/-stems. Each

Fig. 1. Example words from the artificial grammar. Arrows point from the trigger to the target of a pattern. In the plural, the [e] in the stem [mez] changes the suffix to [el]. In the diminutive, the [a] in the prefix [ka-] changes the stem to [maz]. In the diminutive plural, the two combine in a complex fashion: The suffix is still [el] even though the stem vowel has been changed to [a], which normally takes the suffix [il]. Bold indicates vowels that change due to a phonological pattern.

picture was displayed for 3 s with a 1-s ISI. Each audio clip was about 1 s long and started 500 ms after the picture appeared.

At the end of training, participants were tested on their recollection of the 48 words for which they had received training. During the testing of trained items, participants saw a picture and heard two words in a two-alternative forced-choice task. The foil reflected the incorrect form of the suffix (e.g., *kagadel* vs. *kagadil*) or stem (e.g., *kagad* vs. *kaged*)—foils for each item are detailed in the Appendix S1. Each trained item was tested once, in random order. The first alternative was heard 500 ms after the picture appeared, the second word 1,500 ms after the first (with an ISI of around 500 ms, depending on word length). Order of presentation for the answer and foil were randomized and balanced across the study. Participants had 3 s after the beginning on the second word to respond.

After the testing of trained items, participants were tested on their ability to apply the grammar they learned to new words in a version of a *wug*-test (Berko, 1958). *wug*-tests, which are used to assess grammatical knowledge, particularly in children, involve prompting a participant with a new word (e.g., *wug*) and then asking them to produce the word with a modified meaning (e.g., *wugs*), ensuring that participants are displaying knowledge of the grammar, rather than recall of an inflected word. Here, participants saw a picture (from the group of 18 withheld nouns) for 1,500 ms and heard the corresponding new word. After seeing this word-picture pairing, participants saw a blank screen (1 s), then another picture of the same animal but either as a plural, diminutive, or diminutive plural (e.g., first a lion, then many small lions). After seeing the second picture, participants had to choose from two heard alternative words for naming the second picture they saw using a button-press. The trials, which included both simple and complex untrained test items, were presented in random order with no feedback provided. The foils for this two-alternative forced choice *wug*-test are shown in Appendix S1.

2.2.2. General intelligence

After the ALL test, participants were administered the Kauffman Brief Intelligence Test, Second Edition (KBIT; Kaufman & Kaufman, 2004). This test measures IQ, including verbal and non-verbal subcomponents. The test takes approximately 20 min and was administered in English.

2.2.3. Spanish ability

There were several components of the Spanish language assessment used to measure participants' success in learning Spanish in the classroom. One component was the participants' final classroom grades. Instructors reported students' overall final grade in the course, which was comprised of students' grades on (a) chapter exams (55%), which included assessment of vocabulary, grammatical concepts, cultural readings, and videos from particular chapters in the instructional text (VanPatten, Leeser, Keating, & Roman-Mendoza, 2005); (b) pop quizzes (15%), comprised of 10 quizzes administered over the course of the semester that assessed any area that the class instructor wanted to test; (c) online homework (15%), which reflected participants' performance on weekly online homework assignments on vocabulary, grammar, and culture throughout the semester

(students could attempt each homework activity up to three times, with scores reflecting the best attempt only), and (d) participation in class (15%), which was the average of a daily score based on attendance, preparedness, and participation in the Spanish classroom. The instructors also reported the students' grades specific to each of the four graded elements included in the final classroom grade. In addition to reporting student grades, we also asked each instructor to rate participating students on a 1–5 scale based on their reading, writing, speaking and comprehension abilities.

We also used two other more objective assessment instruments. The first was the Elicited Imitation Task (Vinther, 2002). Used for decades as a measure of implicit knowledge of second language (Naiman, 1974), the Spanish version of the Elicited Imitation Task involves listening to sentences in Spanish, then repeating the same sentence in Spanish within a limited time-span. Because people are limited in what they can repeat based on what they can process (Gray & Ryan, 1973), the Elicited Imitation Task serves as a useful tool for rapid language assessment and correlates significantly with more time-intensive measures (Erlam, 2006). The specific Elicited Imitation Task used here was adopted from Ortega (2000) and was modified to reflect Mexican Spanish, which is the dialect reflected in the course textbook. The Elicited Imitation Task is comprised of 30 sentences that vary in their grammatical complexity as well as syllable count (with a range of 7–17 syllables). Participants were instructed to listen to the recorded sentences in Spanish, which were presented one at a time, and to repeat each sentence out loud after hearing a beep that sounded after each sentence. Participants' responses were digitally recorded. The digital recordings were transcribed by two independent raters and then scored following the protocol from Ortega (2000). Each sentence could earn a score of 0, 1, 2, 3, or 4 based on the accuracy of the repetition, yielding a maximum possible score of 120. Any discrepancies between the raters one and two were resolved by a third rater who listened to the recordings independently and provided a final score.

The second objective measure was a brief test of a specific aspect of grammatical knowledge in Spanish, the subjunctive of doubt construction, which is taught explicitly in Spanish language classrooms. It had been originally taught during the previous semester of Spanish and was targeted for review in the present semester, ensuring adequate opportunity to use it. The test was adopted from Farley (2001) and was comprised of two sections: a comprehension portion and a production portion, with the order counterbalanced across participants. For the 24 interpretation items, participants were required to choose between two possible main clauses to begin a sentence whose ending was provided. The critical items were designed to assess knowledge of clauses that require the use of the subjunctive such that participants had to interpret the subjunctive form of the verb in the sentence ending in order to decide which of the two possible main clauses could begin the sentence. Eleven of the test items were distractors and the thirteen critical items were scored as correct or incorrect resulting in a percent correct for each participant.

The production portion of the language assessment was a fill-in-the-blank task where participants were required to complete a provided sentence with the correct form of a verb (provided in infinitive form). This portion of the task included 18 items with 10 crit-

ical items that were designed to elicit a choice between the subjunctive and indicative moods. Examples questions for both tests are provided in Appendix S2.

2.3. Procedure

Participants came in to the lab over a 2-week span in the sixth and seventh weeks of the semester. This served to control for the amount of instruction in this class that they had received prior to the study and to minimize the differences in skill level between participants, though some differences remained. Participants began with the Elicited Imitation Task, then filled out the LEAP-Q questionnaire regarding their language background (Marian, Blumenfeld, & Kaushanskaya, 2007). Participants then took the ALL test followed by the KBIT and Spanish test of the subjunctive construction. Information about the participants that was provided by the instructors (i.e., classroom grades, teacher assessments) was obtained the week after the last day of the semester.

3. Results

The means, standard deviations, and ranges for all measures obtained are provided in Table 1.

For the ALL test participants performed significantly above chance on the recall of the trained items ($t(43) = 7.6$, $p < .001$; Fig. 2). Participants also performed significantly above chance on the simple untrained items ($t(43) = 4.3$, $p < .001$) but were not above chance on complex untrained items ($t(43) = .11$, $p = .91$; Fig. 2). However, 14 of the 44 participants successfully learned the complex grammar and performed significantly above chance for the complex measure (at $p < .05$ for binomial probability, proportion correct $> .66$), reflecting a substantial range in learning success across participants for complex items. As highlighted in Ettlinger et al. (2014), below chance performance on complex items may indicate an interesting aspect of learning and performance. Therefore, we conducted additional statistical analyses taking that into consideration and included these analyses in Appendix S3.

With respect to second language learning, there were significant positive correlations among all measures of Spanish ability (Table 2) suggesting that there is internal consistency among the different measures. The Elicited Imitation Task, in particular, does correlate significantly with almost all of the other measures of Spanish ability. The two measures that do not correlate with the Elicited Imitation Task are homework and class participation grades, which arguably measure effort rather than acumen. All measures of teachers' subjective evaluations of the students in reading, writing, speaking, and comprehension were highly correlated with each other, suggesting minimal distinctiveness among the measures. Also, only final exam score correlates with IQ among the Spanish ability measures, corroborating previous research suggesting that general intelligence or IQ only explains a small portion of the variance observed in language learning (Robinson, 2005a). Correlations also show positive, but not significant, correlations between the three ALL

Table 1
Averages, standard deviations, and ranges for all measures obtained for the 44 participants in our study

	Average (SD)	Range
General info		
Age	21.7 (2.8)	18–27
IQ	96 (7)	81–113
Class grade		
HW	86 (13)	57–100
Participation	95 (7)	65–100
Quiz	87 (12)	50–100
Exam	82 (10)	64–96
Final grade	86 (7)	70–97
Teacher evaluation		
Reading	3.7 (0.99)	2–5
Writing	3.6 (1.00)	1–5
Speaking	3.3 (1.14)	1–5
Comprehension	3.7 (1.09)	1–5
Spanish grammar test		
Comprehension	.49 (0.21)	0.17–1.00
Production	.59 (0.21)	0.17–1.00
Spanish evaluation		
Elicited Imitation Task	25 (16.2)	4–77
Artificial language test		
Recall	0.70 (0.17)	0.43–1.00
Simple	0.63 (0.20)	0.25–1.00
Complex	0.46 (0.23)	0.05–0.93
Average	0.60 (0.13)	0.42–0.89

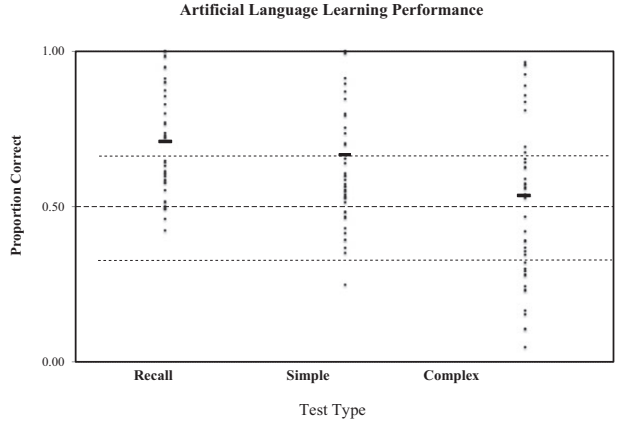

Fig. 2. Performance on artificial language learning tests. Wide marker indicates average. Dotted line indicates significantly above/below chance; dashed line is chance.

Table 2
Correlations across all measures of artificial and second language learning

	Class Grade						Teacher Eval				Span Test		
	IQ	HW	Partic	Quiz	Exam	Overall	Read	Write	Speak	Comp	Comp	Prod	EIT
All													
Recall	−0.21	0.29	0.05	0.29	0.12	0.18	0.31*	0.14	0.36*	0.47**	0.09	0.12	0.04
Complex	0.00	−0.01	0.02	0.39*	0.4*	0.31*	0.29	0.32*	0.20	0.33*	0.06	0.29	0.54**
Simple	0.02	0.11	0.13	0.26	0.31*	0.30	0.56**	0.58**	0.26	0.28	0.06	0.29	0.13
Composite	−0.08	0.17	0.10	0.49**	0.44**	0.41*	0.58**	0.53**	0.4*	0.55**	0.11	0.36*	0.41*
General													
IQ		0.08	−0.05	0.15	0.3*	0.24	0.11	0.04	0.12	0.14	0.22	0.21	0.14
Class grade													
HW			0.36*	0.10	0.35*	0.58**	0.21	0.21	0.22	0.13	0.01	0.08	−0.09
Partic				0.13	0.23	0.47**	0.05	0.23	0.13	0.19	0.02	0.10	−0.15
Quiz					0.29	0.37*	0.33*	0.55**	0.33*	0.47**	−0.02	0.37*	0.32*
Exam						0.9**	0.58**	0.52**	0.53**	0.68**	0.32*	0.58**	0.51**
Overall							0.52**	0.56**	0.56**	0.61**	0.25	0.52**	0.37*
Teacher evaluation													
Read								0.68**	0.61**	0.69**	0.22	0.42**	0.51**
Write									0.56**	0.57**	0.12	0.43**	0.45**
Speak										0.65**	0.04	0.31*	0.45**
Comp											0.19	0.37*	0.41*
Spanish grammar test													
Comprehension												0.61**	0.44**
Production													0.58**

Notes. *Indicates a *p*-value below .05.
**Indicates a *p*-value below .001 (significant at .05 with Bonferroni correction).
ALL, artificial language learning; Comp, comprehension ability; EIT, elicited imitation task; HW, homework grade; Partic, participation grade; Span Test, objective test of subjunctive construction.

measures (recall vs. simple: $r(43) = .14$, $p = .34$; recall vs. complex: $r(43) = .24$, $p = .11$; simple vs. complex: $r(43) = .03$; $p = .87$).

Our primary interest is in the relationship between the ALL and natural language learning measures. The three main questions that are being addressed are as follows: Is there a relationship between ALL and natural language learning? Does this relationship still hold after correcting for IQ? Does this relationship differ for different measures of ALL?

As a preliminary exploratory analysis to address the first question, we consider the overall correlations shown in Table 2. There were a number of positive correlations between the three ALL measures and the different measures of Spanish learning ability. These extend up to $r = .49$ for class grades and $r = .58$ for teacher evaluated performance, which compares favorably to previous studies showing a relationship between natural language learning ability and measures of working memory and artificial grammar learning, which show correlations around $r = .40$ (Misyak & Christiansen, 2012; Robinson, 2005b).

To address the second question on the relationship between ALL and second language learning, independent of the effects of general intelligence, we performed a first-level analysis using correlations among key measures controlling for IQ (Table 3). Importantly, there was still an overall significant correlation between overall final class grade and composite ALL performance ($r(41) = .44$, uncorrected $p = .001$).

A more conservative analysis utilizes Bonferroni correction, given the large number of comparison involved in this study. For the 50 comparisons (10 classrooms measures $\times$ 4 artificial language learning measures + IQ), a very conservative threshold of $p < .001$ may be used. After this correction, and when controlling for IQ, Composite ALL was still significantly correlated with Exam Score, Quiz Score, and Reading, Writing and Comprehension assessment score, and Complex ALL was still significantly correlated with Elicited Imitation Task.

The relatively low number of participants for this individual differences study motivates additional significance testing. First, a Monte Carlo simulation can be used to estimate the likelihood of obtaining the correlations reported in Table 3 simply by chance. For 10,000 simulation iterations of the correlations, scores were generated for 44 participants for 17 performance measures using R (R Development Core Team, 2010). The scores were generated by randomly sampling, with replacement, a value from the actual scores for each performance measure. Thus, the Recall scores for each of the 44 simulation participants were generated by randomly selecting from one of the 44 actual Recall ALL scores; then the Complex ALL scores were randomly selected from the actual Complex ALL scores, and so on for all 17 measures. This ensures the distributional properties of the scores are retained, even if they are not normal, and simulates what the results of our study would be had there been no relationship between ALL, natural language learning, and IQ for each participant. The correlations between these randomly generated scores were calculated in the same manner as the results in Table 3. A histogram of the correlation coefficients is shown in Fig. 3. This histogram shows the correlation values one would get for conducting these analysis on random results.

Table 3
Correlations across artificial language learning measures and classroom performance, correcting for IQ

	Class Grade					Teacher Eval				Span Test		
	HW	Partic	Quiz	Exam	Overall	Read	Write	Speak	Comp	Comp	Prod	EIT
All												
Recall	0.32*	0.04	0.33*	0.19	0.25	0.34*	0.15	0.4*	0.52**	0.15	0.17	0.07
Complex	0.01	0.02	0.4*	0.42**	0.32*	0.29	0.32*	0.20	0.34*	0.07	0.3*	0.49**
Simple	0.11	0.14	0.26	0.32*	0.30	0.56**	0.58**	0.26	0.28	0.06	0.29	0.12
Cormposite	0.18	0.09	0.51**	0.49**	0.44**	0.59**	0.54**	0.42*	0.56**	0.13	0.39*	0.39*

Notes. *Indicates a *p*-value below .05.
**Indicates a *p*-value below .001 (significant at .05 with Bonferroni correction).
Abbreviations as in Table 2.

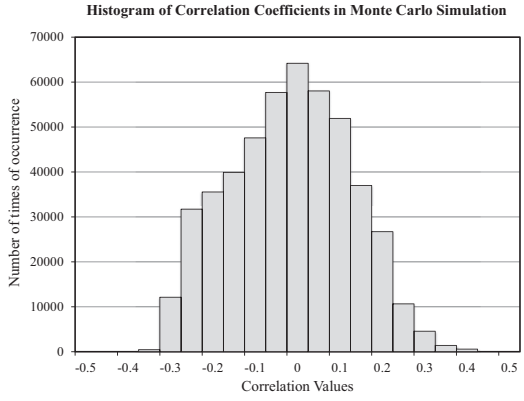

Fig. 3. Histogram of correlation coefficients obtained from a Monte Carlo simulation replicating this study.

As expected, most of the correlations are close to zero. Correlations above .27 were present in 5% of the time, correlations above .39 were present in only 0.1% of the simulations, and there were no correlations as large as .45 in any of the 10,000 simulations of 48 correlations. This suggests that the correlations observed in Table 3, particularly those above .39, are extremely unlikely to be due to chance.

Second, the possibility of any individual participant significantly biasing the findings can be mitigated using a leave-one-out analysis. The correlations between composite ALL score and overall classroom grade and Elicited Imitation Task performance were calculated, controlling for IQ, 44 times, each time leaving out one participant. The correlation between composite ALL score and overall classroom performance ranged from .40 to .51, with a mean of .44 and standard deviation of .020 and the correlation between composite ALL score and Elicited Imitation Task performance ranged from .35 to .52, with a mean of .39 and standard deviation of .025. While leaving out certain participants changed the significance value for these correlations, the results were still significant and there is no evidence that a small number of participants drove the correlations found.

Considering the third question, on the differing relationship between the different measures of ALL and classroom learning, there were a number of positive correlations to consider among the different ALL measures.

After controlling for IQ, Complex ALL was found to be correlated with exam grade ($r(41) = .42$, $p = .005$), overall class score ($r(41) = .32$, $p = .03$), teacher-rating of comprehension ability ($r(41) = .34$, $p = .025$), production on the Spanish test of subjunctive ($r(41) = .30$, $p = .048$), and the Elicited Imitation Task ($r(41) = .49$, $p < .001$).

Simple ALL performance was correlated with exam grade ($r(41) = .32$, $p = .041$) and teaching ratings for reading and writing ($r(41) = .56$, $p < .001$, and $r(41) = .58$, $p < .001$, respectively).

Recall of trained items was related to homework grade ($r(41) = .32$, $p = .04$), quiz grade ($r(41) = .33$, $p = .035$), and teaching ratings of reading, speaking, and comprehension ability ($r(41) = .34$, $p = .030$; $r(41) = .40$, $p = .001$; and $r(41) = .52$, $p < .001$, respectively).

This different set of relationships for complex and simple ALL (e.g., both show a relationship with Exam grade but only complex ALL shows a relationship with Final Grade, see Fig. 4) suggests that second language learning is not a unitary process; it involves a number of different skills and abilities, including understanding, speaking, written communication, and explicit knowledge of the language (i.e., what is tested on exams).

Bivariate correlations do not take collinearity into consideration, and our data had a large number of collinearities, particularly as some measures are embedded in others by design (e.g., aggregate and component ALL scores). Furthermore, there was variability in the linguistic background of the participants: Some were bilingual and they all had different amounts of prior exposure to Spanish. Therefore, we used a step-wise regression to look for unique variance explained. We also included number of years of Spanish and whether the participant was bilingual as covariates. This also allowed us to address the question of how well ALL tests predict second language learning and the reverse question of what aspects of second language learning are tapped into when conducting an ALL test.

We conducted three step-wise multiple regressions with different dependent variables that incorporated both forward selection and backward elimination. The first regression has composite ALL score as the dependent variable, and the initial model included all of the measures of Spanish ability plus IQ, number of years of exposure to Spanish and whether the participant was bilingual as independent measures. After the regression, the final model included Quiz score, Exam score and IQ (Table 4) and accounted for a significant amount of variance in ALL performance (R^2 = .41, p = .0001). Crucially, none of the language experience measures—years of Spanish and bilingualism—factored into performance, possibly due to the narrow standard deviation of years of exposure to Spanish and low number of bilinguals.

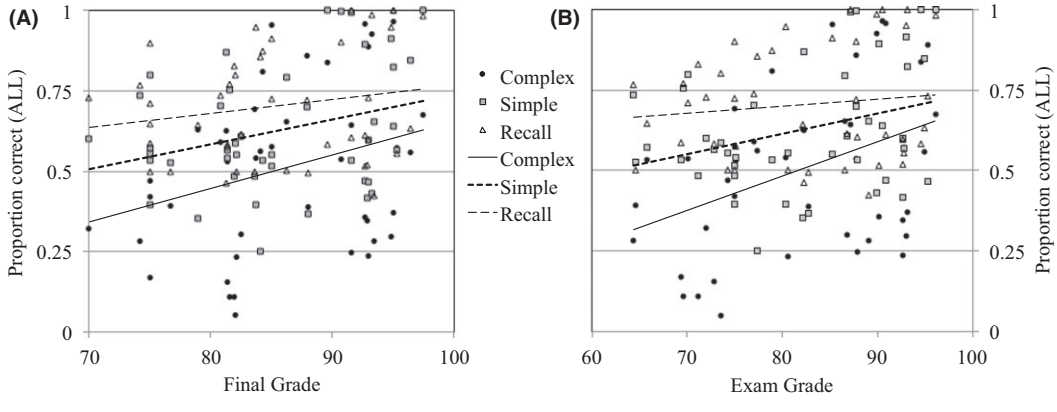

Fig. 4. Correlation between the three artificial language learning measures (recall of training items, generalization for simple items and generalization of complex items) and (A) Final Grade and (B) Total Exam Grade.

Table 4
Step-wise multiple regressions showing the unique variance in composite artificial language learning accounted for by IQ, classroom quiz, and exam scores

Model	Predictor	R	R^2	ΔR^2	B	β	p
1		.497	.244	.244***			
	Quiz				.58	.49	< .001
2		.584	.342	.098*			
	Quiz				.46	.39	.006
	Exam				.49	.33	.021
3		.639	.408	.066*			
	Quiz				.48	.411	.003
	Exam				.59	.398	.005
	IQ				.51	.270	.046

Notes. ΔR^2 = unique variance accounted for by additional step employed by model.
***$p < .001$, *$p < .05$.

Table 5
Step-wise multiple regressions showing the unique variance in composite Final Grade accounted for by IQ, Complex artificial language learning, and Simple artificial language learning

Model	Predictor	R	R^2	ΔR^2	B	β	p
1		.305	.093	.093*			
	Complex				.089	.31	< .001
2		.420	.176	.083*			
	Complex				.087	.30	.002
	Simple				.11	.29	.004
3		.483	.234	.058*			
	Complex				.088	.30	.004
	Simple				.11	.28	.04
	IQ				.25	.24	.05

Notes. ΔR^2 = unique variance accounted for by additional step employed by model.
*$p < .05$.

The second regression included Final Grade as the dependent variable, with the different ALL scores, plus IQ, bilingualism, and previous Spanish exposure as the independent measures in the initial model. After the step-wise regression, the final model for this second regression (Table 5) accounted for a significant amount of variance in classroom performance ($R^2 = .23$, $p = .01$). As shown, IQ does play a role in Final Grade, as expected, but ALL explained variance beyond IQ. In our study, the Complex and Simple AGL scores were included in the model as explaining this variance, whereas the recall score did not explain any additional variance in performance. As above, years of experience with Spanish and bilingualism did not account for any additional variance.

Finally, the third step-wise multiple regression included Elicited Imitation Task score as the dependent variable, also with the different ALL scores, plus IQ, bilingualism, and previous Spanish exposure as the independent measures in the initial model. The final model for this third step-wise multiple regression (Table 6) accounted for significant variance in Spanish ability ($R^2 = .24$, $p < .001$). The final model included only Complex

Table 6
Step-wise multiple regressions showing the unique variance in Elicited Imitation Task accounted for by Complex AGL

Model	Predictor	R	R^2	ΔR^2	B	β	p
1		.485	.235	.235**			
	Complex				1.86	.484	< .001

Notes. ΔR^2 = unique variance accounted for by additional step employed by model.
**$p < .01$.

AGL as predicting performance on the Elicited Imitation Task, suggesting that this Complex AGL measure is a useful innovation over previous ALL experiments. The fact that IQ and the other ALL measures did not explain additional variance in Elicited Imitation Task performance suggests that this language learning ability (as contrasted with classroom ability) is distinct from other measures of intelligence such as IQ and recall ability.

Although bilingualism and previous Spanish exposure accounted for no additional variance in any of the multiple regression analyses, we compared performance between bilinguals and monolinguals and correlated performance to years of Spanish exposure to further ensure that these are not factors. There was no evidence of a difference between bilinguals and monolinguals for overall ALL (unpaired t-test $t(42) = .29$, $p = .78$), classroom performance ($t(42) = .62$, $p = .53$) or Elicited Imitation Task ($t(42) = 1.3$, $p = .20$) and there was also no evidence for a relationship between previous Spanish exposure and overall ALL ($r(41) = .11$, $p = .47$), overall classroom performance ($r(41) = -.08$, $p = .60$), or Elicited Imitation Task ($r(41) = -.15$, $p = .31$).

Thus, these results show a strong relationship between performance on an ALL task and L2 learning. ALL performance correlated with a broad range of L2 measures, including classroom performance, teacher evaluation of language ability, and more objective measures of language ability. The Complex ALL task showed the strongest relationship with more objective measures like the Elicited Imitation Task. This is in accordance with the idea that ALL paradigms are simplified versions of language learning, and thus the most complex ALL will most closely resemble L2 learning, though correlations were found for Simple and Recall aspects of ALL as well. Conversely, measures of overall classroom performance, which are based on a number of non-language-learning-related skills, correlated most closely with Composite measures of ALL. Crucially, these relationships still hold when controlling for general intelligence, or IQ. Finally, we considered which different aspects of L2 learning are captured by ALL overall. The results of a multiple regression with ALL as the dependent variable suggested that ALL taps primarily into IQ and classroom performance on quizzes and exams.

4. Discussion

In this study, we examined the relationship between ALL and natural language learning, how it may differ for different measures of ALL, and how the relationship may be

mediated by IQ. Our primary finding—a positive correlation between performance on an ALL task and second language learning in an ecologically valid environment—provides a key link between studies that use ALL experiments and our understanding of second language learning in the classroom.

By virtue of using an ALL task with several measures and meaning, we were also able to show that the more complex grammar tracked most closely with classroom performance and Spanish ability. This suggests that ALL studies that incorporate a semantic component and involve more complicated grammatical systems may closely resemble second language learning. On the other hand, the composite measure of ALL, which included recall and simple and complex grammatical generalization, is most closely related to overall classroom performance, which includes study skills, motivation, etc. Because different aspects of ALL were related to different aspects of second language learning, we may conclude that not all ALL paradigms would be expected to approximate language learning.

Further research can explore the generalizability of these results to other ALL paradigms. A more comprehensive study would be longitudinal and follow students over the course of a number of semesters, to observe changes in proficiency, which is more reflective of learning, and would include a larger sample size, as is important in individual differences studies (e.g., Desmond & Glover, 2002).

Juxtaposing the differences between the present ALL paradigm and other studies, which (a) found no relationship between artificial and second language learning (Robinson, 2005b), (b) found a relationship mediated by other factors (Brooks & Kempe, 2013), or (c) found an indirect relationship (Evans et al., 2009; Misyak, Christiansen, & Tomblin, 2010b; Misyak et al., 2010a) suggests that the specific methods used in an ALL paradigm matter in terms of engaging natural language learning processes. The current paradigm differs from previous studies by having a semantic component, by being multimodal with auditory and visual picture referents, and by focusing on morphophonology. Future research manipulating modality, semantics, and the parts of artificial language acquired can provide further clarity on what is necessary to best approximate natural language learning.

Finally, the results address our question on whether a relationship between ALL and classroom learning still holds when factoring out general intelligence. The correlations are still significant after controlling IQ, suggesting that the ability being tapped into by ALL and L2 learning is distinct from general intelligence.

Further characterizing the nature of this language learning ability remains an interesting challenge. The results of this study could mean that there is a distinct language learning skill underpinning second language learning ability and that ALL studies are a useful method of exploring and evaluating that ability. This is generally the assumption made by authors using ALL studies to explore language function, including those showing an overlap between neural mechanisms associated with language processing and neural mechanisms associated with ALL (e.g., Friederici et al., 2002).

Alternatively, the results could be interpreted to mean that there is some third skill or capability that is crucial for both ALL and second language learning distinct from IQ.

This skill may be related to perceptual learning in the auditory system (Maye, Werker, & Gerken, 2002), general pattern matching, or different memory subsystems.

Indeed, auditory working memory has been argued to be involved in both first and second language learning (Baddeley, 1992; Baddeley, Gathercole, & Papagno, 1998; Ellis & Sinclair, 1996) as well as in ALL success (Amato & MacDonald, 2010; Misyak & Christiansen, 2012).

The procedural and declarative memory systems have also been suggested to play a role in both artificial and second language learning (Conway, Bauernschmidt, Huang, & Pisoni, 2010; Ettlinger et al., 2014; Morgan-Short et al., 2014; Ullman, 2004, 2005). In particular, previous research has suggested that L2 learning is supported by procedural memory (Ettlinger, 2008; Morgan-Short et al., 2014); that procedural memory is an important component of ALL (Reber, 1967); and that procedural memory is distinct from general intelligence (Cohen & Squire, 1980). Thus, procedural memory may be the common substrate for ALL and L2 learning that is distinct from IQ. However, the fact that the inclusion of semantics may be an important part of a predictive ALL paradigm suggests that there may be more than procedural learning involved.

Implicit statistical learning may also play an important role (Conway et al., 2010; Misyak et al., 2010b) as it has also been shown to be distinct from IQ (Kaufman et al., 2010). This is further supported by evidence showing a role for dopamine in second language learning (Wong, Morgan-Short, Ettlinger, & Zheng, 2012) and in more general implicit learning processes (Jocham et al., 2009).

Ultimately, there may be some unique learning mechanism (Hauser et al., 2002) or unique combination of general mechanisms (Pinker & Jackendoff, 2009) that is specific to acquiring linguistic systems. This study provides no evidence to distinguish these possibilities, but understanding the interaction between general cognitive capabilities underlying ALL and second language learning will provide insight into understanding human language learning as a unique ability (Hauser et al., 2002) or as an ability primarily shaped by domain-general cognitive function (Elman, Bates, Johnson, & Karmiloff-Smith, 1996).

5. Conclusion

The results of this study provide evidence for a relationship between ALL and second language learning. This suggests that previous research using ALL experiments may provide insight into naturalistic language learning, particularly when they include a semantic component and complexity. However, the full theoretical implications of this finding still remain unclear given that the nature of this relationship is still unknown. Future research and larger, longitudinal studies can provide more insight into the specific cognitive components involved in artificial and natural language learning. These future studies can then address the question of whether ALL experiments provide insight into language-specific learning abilities or whether it is more a function of motivation, different

memory subsystems, perceptual abilities, some other cognitive ability, or, as is likely, some combination of these abilities.

Acknowledgments

This work was supported by the Liu CheWoo Institute of Innovative Medicine at The Chinese University of Hong Kong, the US National Institutes of Health grants R01DC008333 and R01DC013315, the Research Grants Council of Hong Kong grants 477513 and 14117514, and the Global Parent Child Resource Centre Limited to PCMW and US National Institute of Health grant T32 NS047987 supporting ME.

References

Amato, M. S., & MacDonald, M. C. (2010). Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition*, *116*(1), 143–148. doi:10.1016/j.cognition.2010.04.001.

Aslin, R. N., & Newport, E. L. (2008). What statistical learning can and can't tell us about language acquisition. In P. M. J. Colombo & L. Freund (Eds.), *Infant pathways to language: Methods, models, and research directions* (pp. 15–29). Mahwah, NJ: Lawrence Erlbaum Associates.

Baddeley, A. D. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173.

Berko, J. (1958). The child's learning of English morphology. *Word-Journal of the International Linguistic Association*, *14*(2–3), 150–177.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer (Version 4.3.19). Available at http://www.praat.org. Accessed 3/12/2008.

Braine, M. D. (1963). On learning the grammatical order of words. *Psychological Review*, *70*, 323–348.

Brooks, P. J., & Kempe, V. (2013). Individual Differences in adult foriegn language learning: The mediating effect of metalinguistic awareness. *Memory and Cognition*, *41*, 281–296.

Chomsky, N. (1957). *Syntactic structures*. The Hague, the Netherlands: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Christiansen, M. H., Kelly, M. L., Shillcock, R. C., & Greenfield, K. (2010). Impaired artificial grammar learning in agrammatism. *Cognition*, *116*(3), 382–393. doi:10.1016/j.cognition.2010.05.015.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, *210*(4466), 207–210.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371. doi:10.1016/j.cognition.2009.10.009.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. doi:10.1016/j.cognition.2011.10.017.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, *118*(2), 115–128.

Dominey, P. F., Hoen, M., Blanc, J. M., & Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies. *Brain and Language*, *86*(2), 207–225.

Ellis, N. C., & Sinclair, S. G. (1996). Working Memory in the Acquisition of Vocabulary and Syntax: Putting Language in Good Order. *Quarterly Journal of Experimental Psychology*, *49A*(1), 234–250.

Elman, J. L., Bates, E. A., Johnson, M. H., & Karmiloff-Smith, A. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: The MIT Press.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464–491.

Esper, E. A. (1925). A technique for the experimental investigation of associative interference in artificial language material. *Language Monographs*, *1*, 1–47.

Ettlinger, M. (2008). Input-driven opacity. Ph.D. thesis, University of California, Berkeley.

Ettlinger, M., Bradlow, A. R., & Wong, P. C. M. (2014). Variability in the learning of complex morphophonology. *Applied psycholinguistics*, *35*(4), 807–831.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical Learning in Children With Specific Language Impairment. *Journal of Speech Language and Hearing Research*, *52*(2), 321–335. doi:10.1044/1092-4388(2009/07-0189).

Farley, A. P. (2001). Authentic processing instruction and the Spanish subjunctive. *Hispania*, *84*, 289–299.

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*(4), 384–412. doi: Doi 10.1016/J.Jneuroling.2008.12.002

Finley, S., & Badecker, W. (2009). Artificial grammar learning, and feature-based generalization. *Journal of Memory and Language*, *61*, 423–437.

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*(5656), 377–380. doi:10.1126/science.1089401.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(7), 2458–2463.

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(1), 529–534.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767. doi:10.1207/s15516709cog0000_34.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience*, *5*, 11. doi:10.3389/fnhum.2011.00011.

Genesee, F. (1976). The role of intelligence in second language learning. *Language Learning*, *26*, 267–280.

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, *440*(7088), 1204–1207. doi:10.1038/nature04675.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Science*, *4*(5), 178–186.

Goschke, T., Friederici, A. D., Kotz, S. A., & van Kampen, A. (2001). Procedural learning in Broca's aphasia: Dissociation between the implicit acquisition of spatio-motor and phoneme sequences. *Journal of Cognitive Neuroscience*, *13*(3), 370–388.

Gray, B., & Ryan, B. (1973). *A language program for the nonlanguage child*. Champaign, IL: Research Press.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Hudson Kam, C. L., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learner in language formation and change. *Language Learning and Development*, *1*, 151–195.

van der Hulst, H., & van de Weijer, J. (1995). Vowel harmony. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 495–534). Cambridge, MA: Blackwell.

Jocham, G., Klein, T. A., Neumann, J., von Cramon, D. Y., Reuter, M., & Ullsperger, M. (2009). Dopamine DRD2 polymorphism alters reversal learning and associated neural activity. *Journal of Neuroscience*, *29* (12), 3695–3704.

Kaufman, S. B., Deyoung, C. G., Gray, J. R., Jimenez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. doi:10.1016/j.cognition.2010.05.011.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman brief intelligence test* (2nd ed.). Bloomington, MN: Pearson.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, *105*(31), 10681–10686.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 169–181.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech Language and Hearing Research*, *50*(4), 940–967.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

McNealy, K. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience*, *26*(29), 7629–7639. doi:10.1523/jneurosci.5501-05.2006.

Michael, E. B., & Gollan, T. H. (2005). Being and becoming bilingual: Individual differences and consequences for language production. In J. F. Kroll & A. M. B. d. Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 389–407). New York: Oxford University Press.

Mirkovic, J., Forrest, S., & Gaskell, M. G. (2011). Semantic regularities in grammatical categories: Learning grammatical gender in an artificial language. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 324–329). Austin, TX: Cognitive Science Society.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302–331.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010a). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, *1*(31).

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010b). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*, 138–153.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*, 83–127.

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*(1), 56–72.

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. *Language Learning*, *60*(1), 154–193.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience*, *24*(4), 933–947.

Mori, K., & Moeser, S. D. (1983). The role of syntax markers and semantic referents in learning an artificial language. *Journal of Verbal Learning and Verbal Behavior*, *22*(6), 701–718.

Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, *17*(8), 1229–1244. doi:10.1162/0898929055002463.

Nagata, H. (1987). Extraction of linguistically relevant pragmatic contrast through language learning. *Journal of Psycholinguistic Research*, *16*(1), 43–61.

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, *2*, 1–37.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, *49*(2), 85–117. doi:10.1016/j.cogpsych.2003.12.002.

Opitz, B., & Friederici, A. D. (2003). Interactions of the hippocampal system and the prefrontal cortex in learning language-like rules. *NeuroImage*, *19*(4), 1730–1737.

Opitz, B., & Friederici, A. D. (2004). Brain correlates of language learning: The neuronal dissociation of rule-based versus similarity-based learning. *Journal of Neuroscience*, *24*(39), 8436–8440. doi:10.1523/JNEUROSCI.2220-04.2004 24/39/8436 [pii].

Ortega, L. (2000). Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners. Ph.D. dissertation, University of Hawaii.

Petersson, K. M., Folia, V., & Hagoort, P. (2010). What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language*, *120*(2), 83–95.

Pinker, S., & Jackendoff, R. (2009). The reality of a universal language faculty. *Behavioral and Brain Sciences*, *32*(5), 465–466. doi:10.1017/s0140525x09990720.

Pothos, E. M., & Kirk, J. (2004). Investigating learning deficits associated with dyslexia. *Dyslexia*, *10*(1), 61–76. doi:10.1002/dys.266.

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria. Available at http://www.R-project.org/. Accessed 7/1/2005.

Rafferty, A. N., Griffiths, T. L., & Ettlinger, M. (2013). Greater learnability is not sufficient to produce cultural universals. *Cognition*, *129*(1), 70–87. doi:10.1016/j.cognition.2013.05.003.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. doi:10.1016/j.cognition.2009.02.012.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Robinson, P. (Ed.) (2002). *Individual differences and instructed language learning*. Amsterdam: Benjamins.

Robinson, P. (2005a). Aptitude and second language acquisition. *Annual Review of Applied Linguistics*, *25*, 46–73.

Robinson, P. (2005b). Cognitive abilities, chunk-strength and frequency effects during implicit Artificial Grammar, and incidental second language learning: Replications of Reber, Walkenfeld and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance to SLA. *Studies in Second Language Acquisition*, *27*, 235–268. doi:10.1017/S0272263105050126.

Robinson, P. (2010). Implicit Artificial Grammar and incidental natural second language learning: How comparable are they? *Language Learning*, *60*(Supplement 2), 245–263.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449. doi:10.1016/j.cognition.2010.06.004.

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*(4), 885–913.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1–2), 231–270.

Ullman, M. T. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In C. Sanz (Ed.), *Mind and context in adult second language acquisition: Methods, theory, and practice* (pp. 141–178). Washington, DC: Georgetown University Press.

VanPatten, B., Leeser, M. J., Keating, G. D., & Roman-Mendoza, E. (2005). *Sol y viento: Beginning Spanish*. New York: McGraw-Hill.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics (INJAL)*, *12*, 54–73.cogs

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*(2), 269–304. doi:10.1017/S0272263105050138.

Williams, J. N. (2012). Working memory and SLA. In S. M. Gass & A. Mackey (Eds.), *The handbook of second language acquisition* (pp. 427–442). New York: Routledge.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*, 945–982.

Wong, P. C., Ettlinger, M., & Zheng, J. (2013). Linguistic grammar learning and TAQ-IA polymorphism. *PLoS ONE*, *8*(5), e64983. doi:10.1371/journal.pone.0064983.

Wong, P. C., Morgan-Short, K., Ettlinger, M., & Zheng, J. (2012). Linking neurogenetics and individual differences in language learning: The dopamine hypothesis. *Cortex*, doi:10.1016/j.cortex.2012.03.017.

Wong, P. C., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., & Zatorre, R. J. (2008). Volume of left Heschl's Gyrus and linguistic pitch learning. *Cerebral Cortex*, *18*(4), 828–836.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Stimulus items for artificial language

Appendix S2. Sample questions for test of Spanish subjunctive

Appendix S3. Results after removing below-chance participants