5-2015

Proficiency in Native and Nonnative English Speakers: the Effects of Individual Difference Variables on Performance

Haley E. Barlow

Proficiency in Native and Nonnative English Speakers: the Effects of Individual Difference Variables on Performance

By

Haley E. Barlow

Undergraduate honors thesis under the direction of

Dr. Janet McDonald

Department of Psychology

Submitted to the LSU Honors College in partial fulfillment of the Upper Division Honors Program.

May, 2015

Louisiana State University
& Agricultural and Mechanical College
Baton Rouge, Louisiana

Proficiency in Native and Nonnative English Speakers: the Effects of Individual Difference Variables on Performance

Haley E. Barlow

Louisiana State University

Abstract

Language proficiency in native and early and late learning nonnative speakers may be influenced by a number of variables, including age of exposure to the language, language aptitude, and working memory capacity. This study examined the relationship between performance on a difficult grammaticality judgment task with these variables. We found spread in native and nonnative speakers' scores that corresponded to language aptitude and working memory capacity scores. Also, WMC correlated positively with language aptitude, and high WMC and language aptitude were predictive of high scores within the native group. We also saw a general effect of age of arrival on nonnative speakers' performance, with later arrivers scoring more poorly. In addition, WMC and language aptitude effects were only seen in native speakers. Results supported the idea that individual difference variables are predictive of native performance on a difficult grammar test. The accuracy of the results and generalizability of this study are limited by the small sample of nonnative speakers, n = 10.

Proficiency in Native and Nonnative English Speakers: the Effects of Individual Difference Variables on Performance

For years, the scientific and educational community has attempted to examine the differences between native and non-native speakers of languages in general. There is a general consensus that most non-native speakers are not able to reach the same level of proficiency as native speakers (Johnson & Newport, 1989; Birdsong & Molis, 2001; Abrahamsson & Hyltenstam, 2008), but many researchers agree that some non-native speakers may be able to reach a level of performance indistinguishable from native performance (Birdsong & Molis, 2001; Abrahamsson & Hyltenstam, 2008). Proficiency at this level consists of correct pronunciation, grasp of idiomatic expressions, and, crucially, correct grammar usage; this study will focus on the grammatical aspect of language, and the variables affecting success in this area. These variables may include age of exposure to the second language (Johnson & Newport, 1989; Birdsong & Molis, 2001), language aptitude (Abrahamsson & Hyltenstam, 2008), working memory capacity, ratio of current use of first to second language (Birdsong & Molis, 2001), and other factors.

Before examining these factors in detail, the question remains as to what qualifies as nativelike performance. Most studies concerned with this concept define it quantitatively, as the relationship of a nonnative score on various language tests (grammaticality judgments (Abrahamsson & Hyltenstam, 2008; Johnson & Newport, 1989; Birdsong & Molis, 2001), intuition judgments (Coppetiers, 1987), etc.) to the AVERAGE score of native speakers on the same test. However, if the range, as well as the average of scores is considered, there is wide variation within the native population itself, as seen in Abrahamsson and Hyltenstam's (2008) examination of aptitude effects on second language acquisition-- in fact, even disregarding the

effects of language aptitude in this study, some native speakers scored lower than some non-native speakers on a grammaticality judgment test. Clearly, though this is the case, these low-scoring native speakers should not be classified as nonnative. This begs the question, then, as to what the meaning is of nativelikeness. The answer relies on differences or similarities that exist, regardless of native status, among higher scoring and lower scoring individuals.

Age effects

Age of exposure, or age of onset, refers most often to the date of first immersive exposure to the second language, such as when the individual first moves to a new country or begins attending school in said country; the age of first exposure may affect ultimate achievement in the second language. Critical Period theorists suggest that, after a certain point, usually puberty, it becomes more difficult for individuals to learn a second language (Lenneberg, 1967, as cited in Johnson & Newport, 1989). Johnson and Newport (1989), in an examination of the Critical Period, found that, indeed, Korean- and Chinese-speaking participants who were immersed in English after age 17(their designated age of puberty was 16 years) did score significantly lower than those who had arrived in the US before age 15 as well as native speakers. They also found that the group which had arrived in the US between ages three and seven scored significantly better than any other age group, with scores equal to those of native speakers; no group arriving later than eight years of age saw performance equal to native performance. The results were taken to support the existence of a critical period in language learning.

Birdsong and Molis (2001) attempted to replicate Johnson and Newport's results using participants of a different first language (Spanish), to determine whether such a period exists. Contrary to Johnson and Newport's findings, Birdsong and Molis interpreted the results as precluding a critical period, because some late learners were able to score within the range of

native speakers' scores. According to their interpretation of the Critical Period theory, this should not be possible. However, the overall age effect was supported: the older an individual was at age of first exposure, the lower their proficiency scores later; and this trend continued beyond the critical period (Birdsong & Molis, 2001). Their use of Spanish speakers instead of Korean or Chinese speakers was intended to test the basest form of the Critical Period theory as they saw it: individuals who begin learning a second language during the critical period should be able to reach nativelike levels of performance. If the theory is correct, they postulated, the similarity of the first language to the second language should have no bearing on eventual proficiency (Birdsong & Molis, 2001). However, another interpretation of the theory is possible: that is, if an individual does not begin learning a second language during the critical period, it will be more difficult to attain a higher level of proficiency. The critical period does not necessarily guarantee performance at native levels—it may simply make it easier to reach. If this is the case, it is within the reasoning of the theory that first language may have an effect on the rate at which individuals learn the second language. Though there is certainly some dispute about the legitimacy of a critical period, most researchers agree that there is an expected effect of age on language learning ability—as the individual gets older, not necessarily within the critical period, language performance decreases, eventually leveling off (Johnson & Newport, 1989; Birdsong & Molis, 2001). Thus, we expect to see a general effect of age of exposure on proficiency scores, potentially represented by a negative relationship showing diminishing returns of age.

Language Aptitude Effects

Language aptitude is a measurement of a person's innate ability to acquire a language, and is typically defined as consisting of four areas: phonetic coding ability, grammatical

sensitivity, rote learning, and inductive learning ability (Abrahamsson & Hyltenstam, 2008; Bylund, Abrahamsson, & Hyltenstam, 2009). Language aptitude tests purportedly determine, overall, how predisposed an individual is to learning a language; this provides an interesting look into how individuals differ, and allows an examination of language performance in regards to these differences. Abrahamsson and Hyltenstam (2008) found that native Swedish speakers with higher language aptitude did not score significantly better on a Swedish grammaticality judgment test than lower aptitude individuals. However, higher aptitude nonnative Swedish speakers, all of whom passed for native in everyday conversation, scored better than lower aptitude nonnative speakers, but only if the speakers had been exposed to the second language after the age of 13 ("late learners"). Early learners were equally likely to score high regardless of aptitude, scored about as well as native speakers, and saw a similar spread in scores; though some early learners did score lower than native speakers, the majority scored higher than the lowest-scoring native speaker. These results suggest an effect of language aptitude on eventual second language proficiency for late learners, though not on eventual language proficiency for native speakers and early learners.

Bylund et al. (2009) performed what was essentially a reversal of Abrahamsson's 2008 study, examining whether language aptitude could predict first language attrition, loss of language ability, after learning a second language at an early age. Participants with higher language aptitude scores did perform significantly better on a grammaticality judgment test of the neglected first language than lower aptitude individuals. They also found that, for lower aptitude individuals only, higher scores were related to more daily use of the first language (Bylund et al., 2009). This finding was counter to Birdsong and Molis' (2001) findings that grammaticality test score and current English usage did not significantly correlate. The results of

this study indicate a clear effect of language aptitude on retention of a first language, and support the idea that aptitude is related to language ability in general. Based on this, we predict an effect of language aptitude on proficiency scores.

Working Memory Capacity Effects

Related to language aptitude, and likely playing some part in it, is working memory capacity. Working memory capacity (WMC) is the ability of an individual to take in new information, hold it in the short-term memory, and manipulate it, while simultaneously retrieving and manipulating information from the long term memory. WMC plays a large part in problem solving ability, and higher WMC is related to better performance in language comprehension, reading comprehension, vocabulary learning, storytelling, and other complex cognitive tasks (Just & Carpenter, 1992; Daneman &Carpenter, 1980; Daneman & Green, 1986; Pratt, Boyes, Robins, & Manchester, 1989; all as cited in Lusk, Evans, Jeffrey, Palmer, Wikstrom, & Doolittle, 2009). WMC is, therefore, especially relevant to this study; due to the nature of language and language learning, it is highly likely that WMC will have an effect on proficiency. In addition, the complex nature of items on the grammaticality test we formulated for this study may have an effect on the scores of low WMC individuals. (Some questions are rather long, and require participants to remember relatively large amounts of information while making alterations to ungrammatical portions of the question.) We expect to see differences in score between high and low WMC participants across the experimental groups, stronger in non-native late learners.

Other Factors

A number of other factors may have an effect on eventual language proficiency. Coppieters (1987) performed an interview and grammar test of several native and nonnative French speakers, all of whom were determined by native speakers to be of high proficiency. The

grammatical items covered areas of the language considered to be more subtle or nuanced. He found that the nonnative speakers, who were from varied linguistic backgrounds, did not conform to the grammaticality judgments of the majority of the native speakers and, interestingly, gave very different reasons for their judgments. Some of the reasoning was based in ideals of their background culture: for instance, a Farsi participant made an incorrect pronoun judgment based on the fact that people in the present deserved the respect of a "person pronoun" instead of an "item pronoun," and people in the past did not. No other person gave a similar reason. This has interesting implications on the effects of background and culture—namely, that differences in linguistic and cultural background may result in differences in underlying linguistic cognitive structures, intuition, and reasoning. Also, in line with the differing results of Birdsong and Molis' study (2001) and Johnson and Newport's study (1989), participants whose first language differed greatly from French did more poorly on the grammar test than those of more similar languages. For instance, Sino-Tibetan and Japonic language group participants did poorly on questions concerning articles, which do not exist in their languages. Interestingly, some of the native speakers also did not conform to the majority native response and reasoning; this implies that there are individual differences in the underlying cognitive structure of language, even within the native group.

There are a number of variables beyond age of exposure, language aptitude, and working memory capacity that could cause differences in language proficiency that we will not be examining in this study. These factors include socioeconomic status, use of language (use of the second language in daily life, as well as the amount an individual reads or writes in any language), academic focus (for instance, linguistics majors as opposed to STEM majors), physical and mental health (see Oertel-Knöchel *et al.*, 2014) the balance between visuospatial

and phonological memory (see Papesh, 2014), and others. Though it is certainly possible that there would be significant effects of these factors on proficiency, we are primarily concerned with language aptitude, age of acquisition, and WMC, as these three seem more likely to render clear data.

Most previous studies used an auditory approach to administering the grammaticality judgment and other test items; this study will utilize both auditory and written format, as Abrahamsson and Hyltenstam (2008) did, in order to present and test certain grammar functions in the most appropriate format for each function (for instance, its vs. it's, their/there/they're, your/you're), as well as to provide means of determining potential effects of reading disability or vision or hearing issues.

Based on the previous literature, we expect to see age effects for nonnative speakers; specifically, the later the age of exposure, the lower the proficiency scores and the higher the variability will be. We also expect to see effects of language aptitude: we predict that all native speakers, all early learning nonnative speakers, and high aptitude late learning nonnative speakers, will score significantly better than low aptitude late learners. In addition, we expect that aptitude effects and score spread will be similar for early learners and native speakers, that early learners will fall largely within native range, and that the overall effect will be greater for late learners than for early learners and native speakers. If we can avoid a ceiling effect by using a difficult grammaticality test, we expect to see effects of working memory capacity (WMC); we hope for a spread in native speakers' scores, and we predict that high WMC participants will score better across groups than low WMC, that high WMC will correlate with high language aptitude, and that WMC will have more of an effect on late learning native speakers than on early learners or native speakers.

Methods

Participants

Our participants consisted of 54 native English speakers, and 10 nonnative speakers. Participants were taken from the student population of Louisiana State University, using the SONA recruiting program and word of mouth. Participants were given candy and awarded extra credit points in their psychology classes for participation. Nonnative speakers who were not currently enrolled in a psychology class were entered into a drawing for a gift card. We recruited more native than nonnative speakers, in order to better examine any individual difference effects within native speakers. Our nonnative speakers came from a variety of backgrounds and first languages, five speaking Spanish, two Arabic, one Urdu, one Telugu, and one Polish. They also had a variety of ages of acquisition, ranging from 0 years to 25 years, with $M = 14.5$ years, and with four participants arriving before age eleven and six arriving after age 15. All participants were adults.

One additional participant was excluded from the nonnative group because she self-reported as dyslexic. No other participants were excluded from the study, despite minor issues. Issues that did not affect any items or potentially affected only one or two items included phones going off in the middle of certain tasks, participants missing part of a list in the WMC task, technical glitches such as audio delays, etc. Issues that may have had a greater effect, but which were not excluded in this study, included people talking in the hallway, or a phone ringing in an office down the hall for some period of time.

Materials

Demographic Questionnaire

The demographic questionnaire is a self-report measure formulated to record a number of potential variables about the participants, including age, gender, age of exposure, major, current use of first language, and others (see Appendix 1). It was administered by personal interview rather than by computer, unlike the other tests, in order to allow the administrator to acquire the most information possible.

Language Aptitude Measure

In order to test participants' language aptitude, we used the LLAMA Language Aptitude Test, based on the Swansea LAT (Meara, 2005). The LLAMA is computer based, and consists of four parts: a vocabulary learning task (LLAMA_B), a sound recognition task (_D), a sound-symbol correspondence task (_E), and a grammatical inferencing task (_F). The LLAMA is formulated to test the four major factors of language aptitude regardless of native language, so each task consists of words, sounds, or sentences from a made-up language and some general images. The vocabulary-learning task presents participants with a group of symbols, with different names; the participants have a limited amount of time to learn the names of the items, which appear in a box in the center of the page when an item is clicked on. In the sound recognition task, participants listen to a number of "words," and then recall whether or not they heard the words in a recognition task. In the sound-symbol correspondence task, participants are exposed to a series of syllables and a series of symbols, and then choose which sounds correspond to which symbols. In the grammatical inferencing task, participants read a number of phrases paired with certain pictures, and then must infer which new phrases are grammatical and

match a set of new pictures. For more information and access to the LLAMA tasks and manual, visit http://www.lognostics.co.uk/tools/llama/.

Working Memory Capacity Measure

Due to time constraints, we tested working memory capacity using a size judgment span task often used in Dr. McDonald's lab, in which participants are presented, orally, with a series of objects, and rearrange them according to size. For instance, when the participant pressed a key to begin, a list like the following would play on the computer speakers:

"Television. Shoe. Bed. Mouse."

The participant would wait until the computer screen said "report," and then say, out loud, "Mouse, shoe, television, bed."

Forced Choice English Grammar Test

The majority of the Forced Choice English Grammar test (FCEGT) consists of questions modeled closely on those in the Verbal section of the Kaplan GMAT Strategies, Practice & Review (2014); we will call our version of the test Difficult Syntactic Structures. Each question used the general construction of items in the GMAT practice book and was based on a grammatical function, or group of functions, found difficult by native English speakers; some structures were selected from those used to test verbal proficiency on the GMAT, and some were selected from other sources (ESL learning materials, or suggestions). The structures chosen were deemed to be difficult not only for ESL learners, but also for native English speakers, to avoid a ceiling effect. The 9 error types included tense agreement, number agreement, conjunctions, word order, parallel structures in lists, comparatives, and relative pronouns, as well as some special structures such as "since/present perfect" and "ranging from/to." Foils for each error type were constructed in the following ways:

1. Tense agreement: Foils were constructed by taking the matching tense format and altering verbs within the answer choices to be future, present, or past, based on the correct time tense, so that the verbs did not match either the context of the sentence or each other.

2. Number agreement: Foils were constructed by changing the tense of verbs referring to plural or singular items so that they no longer matched the number tense.

3. Conjunctions: Foils were created primarily by adding extra conjunctions or by changing the word order of the answer choices.

4. Word order: Foils were created by rearranging the sentence for each answer choice so that clauses were in a non-typical order or sentences were otherwise rendered meaningless.

5. Parallel structures in lists: Foils were created by changing the items of the list so they were no longer parallel, by adding or removing verbs from one or more item.

6. Comparatives: Foils were created by changing the object to which the "than" clause refers, until the sentence no longer makes sense.

7. Relative Pronouns: Foils were created by using the incorrect propositions within the sentence, or by changing the order of the sentence so that it no longer made sense.

8. Since/present perfect: Foils were created by changing the requisite present perfect tense verb to other tenses.

9. Ranging from/to: Foils were created by adding unnecessary conjunctions or removing the "to" before the final clause (the "to" clause).

For any question for which there were no remaining options for foils, in order to reach five foils per question, the remaining foils were created by using incorrect word order, usually formed by rearranging one of the previous foils.

These Difficult Syntactic Structure questions consist of a complex sentence with the target section deleted, and five options to fill in the blank portion of the sentence. Though multiple options may be technically grammatically acceptable, there is only one "best response" to each question. An example of the "word-order" items used here, with the error type listed at the end of each foil (see Table 1 for more examples):

Those who read the biography of J. R. R. Tolkien soon learn that

_____ that he initially wrote for his children.

- before the Hobbit became part of the Lord of the Rings mythology, it was a bedtime story (correct response)
- before it was part of the Lord of the Rings mythology, the Hobbit became a bedtime story (time; the Hobbit could not BECOME a bedtime story, if it was initially written as such.)
- before the Lord of the Rings, it was a bedtime story (Lacks antecedent; what was a bedtime story? The Hobbit.)
- the Hobbit became a bedtime story (time; same as second foil)
- it was a bedtime story that became part of the Lord of the Rings mythology (time and lacks antecedent)

The remainder of the written portion was formulated based on areas of difficulty for English-as-a-second-language learners, as presented in Advanced English Grammar for ESL Learners (Lester, 2011). These stimuli, here called Homophone Items, were formulated as simple sentences examining proper usage of its/it's, your/you're, and their/there/they're. In each of these stimulus items, there was only one correct answer; they were presented in a multiple-choice format. The incorrect responses were simply formed by adding the wrong option of each word group. An example of the sentences used here:

She knew by _____ luster that the gold was fake.

a) its

b) it's

Auditory Grammaticality Judgment Test

The auditory section of the examination was taken from the grammatical constructions presented in Advanced English Grammar for ESL Learners (Lester, 2011), and is a simple auditory grammaticality judgment test (AGJT). Eight constructions were tested: irregular plurals, comparative and superlative forms of adjectives, number agreement, relative pronouns, pronoun case, word order, subject/verb agreement, and verb tense. We created grammatical and ungrammatical versions of sentences for each construct, with a total of 68 sentence bases. The auditory items were recorded by a native English speaker, and each construction had from two to twelve base sentences. One example of a sentence set:

Todd was angrier than he had ever been.

Todd was more angry than he had ever been.

See Table 2 for more examples and methods of formulating stimuli.

Procedure

First, participants took all four portions of the LLAMA language aptitude test (Meara, 2005).

Next, the Forced Choice English Grammar Test was administered in forced-choice multiple-choice format through a computer program, wherein the participant hit the key denoting the response they thought was correct before the program moved on to the next item. Each item was presented only once to each participant, in a random order. Answer choice and reaction time was recorded for each item. The Difficult Syntactical structures were administered first, followed immediately by the Homophones structures.

Then, the Auditory Grammaticality Judgment Test stimulus sentences were presented via computer program to each participant in separate random order. A grammaticality judgment was required for each item before the program continued to the next; the grammaticality judgments were made on a keyboard attached to the computer, with the 'd' key signaling grammatical correctness and the 'k' key signaling incorrectness. Participants were told to respond as quickly as possible to each item by pushing the appropriate key. Each item was presented only once. The program recorded the responses, as well as response time for each item.

Next, participants were asked to take a Working Memory Capacity test consisting of a size judgment task. Participants heard a list of words, and repeated the words back in order of smallest to largest item. There were three lists on each level, and each subsequent level consisted of longer lists than the previous level. The size judgment task was scored based on the number of words correctly recalled in each list regardless of the exact position, as long as the participant actually made a clear effort to rearrange and recall the items.

Finally, participants filled out a demographics questionnaire, answering questions about country of origin, age of arrival in the US, number of years studying English (in classes or by immersion), native language, current daily use of native language, gender, major, social activeness, frequency of reading in first and second language, etc.

Results

Analysis - Native Speakers

After collecting the data for the native speakers, we took the weighted average scores of the eight difficult syntactical structures of the Forced-Choice English Grammar Test (conjunctions, from/to, number agreement, word order, parallel structures, relative pronouns, verb tense, and sentences involving "than"), the three homophone structures (its/it's, their/there/they're, and your/you're), and the overall scores on the Auditory Grammaticality Judgment Test. We also separately investigated the grammatical and ungrammatical items on the latter test.

Next, we performed a correlation for each average with the native speakers' raw scores on the LLAMA_B, _D, _E, and _F, and the size judgment task. We then split the participants into groups High or Low LLAMA_B, _D, _E, and _F and WMC, in order to run repeated-measures ANOVAs for the significant correlations; this resulted in 33 High and 21 Low _B, 12 High and 42 Low _D, 35 High and 19 Low _E, 31 High and 23 Low _F, and 23 High and 31 Low WMC. For LLAMA_B, _D, and _F, those participants who scored at the standard Good level (provided in the LLAMA manual) or above were designated as High, and those who scored below were designated as Low. For LLAMA_E, in which the average score was considerably higher than the Good level, participants who scored at Outstanding or above were designated as High and those who scored below were designated as Low. For WMC, we took the average score

and designated those above that score as High and those below as Low. Finally, we ran repeated-measures (RM) ANOVAs between the pairs of measures for which significant correlations were seen.

Individal difference measures. We performed a correlation between the various sections of the LLAMA and the WMC task . The LLAMA_B, _E, and _F tasks showed significant positive relationships with one another; and the LLAMA_D was correlated only with the Size Judgment task. We believe this is because the LLAMA_D is a totally auditory task, and the other LLAMA tasks are all at least partially visual. In addition, the WMC task was significantly positively correlated with each of the LLAMA tasks. This indicates a desirable relationship between the Working Memory Capacity and Language Aptitude Tasks. See Table 3.

Forced-choice English grammar test. A bivariate correlation matrix for performance on the Difficult Syntactical Structures from the forced-choice English grammar test showed significant positive relationships with LLAMA_B, _E, and F (all partially or completely visual tasks, with LLAMA F testing syntactic inferencing), as well as Size Judgment.

After examining the Difficult Syntactical Structures, we ran correlations also with the less-difficult homophone structures. A bivariate correlation matrix for performance on the Homophone structures and the LLAMA and WMC tasks showed a significant positive relationship only with the LLAMA_E, the sound-symbol recognition task, We think this may be due to the common elements of scrutinizing the symbols associated with the proper spelling of the homophones and the nonsense words given in the LLAMA_E.

The correlation results indicate that there is a connection, as we hoped, between Language Aptitude and Working Memory Capacity and overall scores on the Forced Choice English Grammar Test. LLAMA_D, the only LLAMA task that was strictly auditory, was not

found to have a significant correlation with either structure group within the Forced Choice English Grammar Test, $p > .05$; this is not surprising, given that the structures of this test were all present visually. See again Table 3.

Difficult syntactic structures. Having run bivariate correlations for the Forced Choice English Grammar test, we then ran RM ANOVAs, based on the significant correlations, between participants' weighted average scores and Hi vs Low scores on the individual difference tasks.

A repeated-measures ANOVA between Syntactical Structure sentence type and High or Low LLAMA_B score showed a significant main effect of sentence type, $F(7, 364) = 29.695$, $p < .05$, and a significant main effect of High or Low _B, $F(1, 52) = 4.928$, $p < .05$, but no significant interaction effect, $F(7, 364) = .991$, $p > .05$. Another RM ANOVA between Syntactical Structure sentence type and High or Low LLAMA_E score showed a significant main effect of sentence type, $F(7, 364) = 27.225$, $p < .05$, but no significant main effect of High or Low _E, $F(1, 52) = 1.179$, $p > .05$, and no significant interaction effect, $F(7, 364) = .208$, $p > .05$. An RM ANOVA between Syntactical Structure sentence type and High and Low LLAMA_F scores showed a significant main effect of sentence type, $F(7, 364) = 35.076$, $p < .05$, and a significant main effect of High or Low _F score, $F(1, 52) = 6.738$, $p < .05$, but no significant interaction effect, $F(7, 364) = 2.071$, $p > .05$. A final repeated-measures ANOVA for Syntactical Structure sentence type and High or Low WMC showed a significant main effect of sentence type, $F(7, 364) = 28.578$, $p < .05$, and a significant main effect of High or Low WMC, $F(1, 52) = 5.271$, $p < .05$, but there was no significant interaction found between sentence type and WMC, $F(7,364) = 1.573$, $p > .05$.

As the reader will notice, the main effects of sentence type, High or Low Working Memory Capacity, and High or Low LLAMA _B, _E, and _F were significant each time. In

addition, there were no interaction effects, which we would have expected—generally, we would expect that Low language aptitude or WMC individuals would perform doubly worse on the harder structures, but we saw no such effects. See table 4.

We then ran a one-sample t-test to determine whether difficult structures were significantly higher than chance. Scores on two of the Difficult Syntactical Structures were found to be equal to chance--Than and Number Agreement. The Than category consisted of only two questions, and the question that saw performance at chance had only one unequivocally correct answer--the other responses, while technically possible, made the main sentence nonsensical. The Number Agreement category consisted largely of questions with more than one target structure to be matched; we suspect the difficulty of these items is due to the multiple structures, and that the items may have overloaded participants' working memory. As there was no significant difference between high and low WMC individuals' performance on number agreement items ($t(52) = -1.432$, $p > .05$), we concluded that the number agreement structures may have been too taxing for all native participants. See again Table 4.

A one-sample t-test with test value =1 indicated that the easiest syntactical structure for native speakers was Conjunctions, which was not significantly different from a perfect score of 1, $M = .975$, $p > .05$. This indicates that Conjunctions were not difficult enough to effectively distinguish between individual Native speakers. See again Table 4.

In addition to showing levels of difficulty, the Bonferroni-corrected RM ANOVA indicated that scores on “Than,” “Number Agreement,” “Tense,” and “Parallel Structure” items were statistically the same; scores on “Tense,” “Parallel Structure,” and “From/to” items were the same; and scores on “Parallel,” “From-to,” “Relative Pronoun,” “Word Order,” and “Conjunction” items were the same. See again Table 4.

Homophone items. As LLAMA_E was the only test that correlated significantly with the Homophone structures, we ran a repeated-measures ANOVA between Homophone sentence type and High or Low LLAMA_E score. The results showed a significant main effect of sentence type, $F(2, 104) = 7.987$, $p < .05$, but no significant effect of High or Low _E, $F(1, 52) = 2.809$, $p > .05$, and no significant interaction effect, $F(2, 104) = 2.150$, $p > .05$.

A one-sample t-test with test value = .5 indicated that all structures were different from chance, and another t-test with test value = 1 indicated that scores on all homophone structures but "your/you're" were significantly different from perfect ($M = .9886$, $t(56) = -1.753$, $p > .05$). See Table 5.

The RM ANOVA with Bonferroni corrections showed that 'its/it's' items, $M = .895$, were significantly harder than 'your/you're' items, $M = .988$, $p < .05$. However, neither 'its/it's' items nor 'your/you're' items were significantly different from 'their/there/they're' items, $M = .969$, $p > .05$. See again Table 5.

Auditory grammaticality judgment test. A bivariate correlation between the overall weighted average of scores on the auditory GJT did not show significant relationships with any of the LLAMA or WMC task scores; the same result was found for the weighted average of the scores on ungrammatical items. However, the weighted average of scores for grammatical items showed a significant positive relationship with the Size Judgment scores. This indicates a connection between Working Memory Capacity and the Auditory Grammaticality Judgment Test as well as on the Forced Choice English Grammar Test.

Because the correlation with LLAMA_D (the auditory task) was marginally significant, we explored the unweighted averages of the Grammatical items in relation to LLAMA_D. In line with a test of the un-weighted averages that also indicated a marginal relationship between the

auditory portion of the GJT and the LLAMA_D (the auditory task), r = .267, p = .051, there was only a nearly-significant relationship between Language Aptitude and the auditory portion of the GJT, p = .059. See Table 3.

Based on the significant correlations with WMC, we then ran RM ANOVAs between the Grammatical weighted-average scores and High and LOW WMC and between Grammatical scores and LLAMA_D. A repeated-measures ANOVA between grammatical auditory sentence type and High or Low WMC showed a significant main effect of sentence type, $F(11, 572) = 19.912$, $p < .05$, and a significant main effect of High or Low WMC, $F(1, 52) = 5.809$, $p < .05$, but no significant interaction effect, $F(11, 572) = .538$, $p > .05$. A repeated-measures ANOVA between grammatical auditory sentence type and High or Low LLAMA_D score showed a significant main effect of sentence type, $F(11, 572) = 13.356$, $p < .05$, and a significant main effect of High or Low _D, $F(1, 52) = 4.368$, $p < .05$, but no significant interaction effect, $F(11, 572) = .855$, $p = .588$.

A one-sample t-test with test value = .5 showed that all structures were significantly different from chance (all $t > 15.937$, all $p < .05$), and another with test value = 1 showed that all grammatical auditory structures were significantly different from perfect (all $t < -2.059$, all $p < .05$). See Table 6.

Analysis – Nonnative Speakers

After collecting the data for the nonnative speakers, we took the weighted average scores of the eight difficult syntactical structures of the Forced-Choice English Grammar Test (conjunctions, from/to, number agreement, word order, parallel structures, relative pronouns, verb tense, and sentences involving “than”), the three homophone structures (its/it’s, their/there/they’re, and your/you’re), and the overall scores on the Auditory Grammaticality

Judgment Test. We also separately investigated the grammatical and ungrammatical items on the latter test.

We performed a correlation for each average with the nonnative speakers' raw scores on the LLAMA_B, _D, _E, and _F, and the size judgment task. There were a number of correlations which were of high magnitude, but which were, because of low sample size, not significant (e.g. Difficult Syntactical structures with LLAMA_E, Age of Exposure with Grammatical auditory items, etc.). See table 7. Because we found no significant relationships between the WMC and LLAMA tasks and the Grammar Test items, we did not split the nonnatives into High and Low WMC or LLAMA designations. We also did not split participants by early or late arrival in an English-speaking country because we had only ten participants, and five per group would not have provided enough power for strong analyses. We did run RM ANOVAs on the pairs for which we found significant correlations.

Forced choice English grammar test. A bivariate correlation matrix showed no significant correlations for scores on the Forced-Choice English Grammar Test (Syntactic or Homophone structures) with any of the other measures; as such, we did not run any further tests on the data from this Test other than to determine order of difficulty of items. See Table 7.

***Difficult syntactic structures*.** In order to determine the order of difficulty of items, we ran one-sample t-tests and a Bonferroni-corrected RM ANOVA. A one-sample t-test with test value = .5 showed that only Conjunctions, Relative Pronouns, and Tense were found to be different from chance, $p < .05$. Another one-sample t-test with test value = 1 showed that scores on From/to, Number agreement, Parallel structure, Tense, and Than items were different from perfect, all $t > 2.449$, $p < .05$. See Table 8.

The Bonferroni corrected RM ANOVA showed that the easiest syntactical items were Conjunction items, that Parallel, Number Agreement, and Tense items were the most difficult, but that the other items were equal to both the easiest and most difficult tasks, $p > .05$. See again Table 8.

Homophone items. A one-sample t-test showed that scores on all homophone items were different from chance, all $t >$ or $= 19$, $p < .05$, and another one-sample t-test showed that none of the homophone structures were significantly different from perfect, all $p > .05$.

A Bonferroni corrected RM ANOVA showed that all of the homophone structures were statistically the same, $p > .05$.

Auditory grammaticality judgment test. A bivariate correlation between the overall weighted average of scores on the auditory grammaticality judgment test did not show significant relationships with any of the LLAMA or WMC task scores; the same result was found for the weighted averages of the scores on ungrammatical items and on the grammatical items. However, a significant negative relationship was present between Age of Arrival (AoA) and overall weighted average score, weighted average of scores on grammatical items, and weighted average scores on ungrammatical items. This indicates that there is a relationship between AoA, as previous researchers suggested, and performance on a grammaticality judgment task. See again Table 7, and figures 1, 2 and 3.

An RM ANOVA for nonnatives' scores on ungrammatical AGJT sentence types showed no main effect of sentence type, $F\ (11, 99) = 7.242$, $p > .05$. The hardest structures were Tense, Comparatives, Superlatives, Irregular Plurals, and Pronoun case, but they were not significantly different from the rest of the structures. The structure with the highest mean score was Regular Plurals, $M = .963$, $p > .05$. See Table 9.

An RM ANOVA for nonnatives' scores on grammatical AGJT sentence types showed no main effect of sentence type, $F(11, 99) = 1.278$, $p > .05$. None of the auditory structures were significantly different from one another, $p > .05$; however, the structure with the lowest mean score was Few/Little ($M = .75$), and the structure with the highest mean score was Regular Plurals, $M = .925$. See Table 10.

A one-sample t-test with test value .5 showed that ungrammatical Less/fewer, Superlative, Tense, and Word order items were not significantly different from chance, all $t < 2.148$, all $p > .05$. Another one-sample t-test with test value 1 showed that ungrammatical items Few/little, many/much, and subject/verb agreement, and grammatical Irregular plural, Less/fewer, Regular plural, and Word order items were not significantly different from perfect, all $t < 2.236$, all $p > .05$. See again tables 9 and 10.

Analysis – Native vs. Nonnative Speakers

After running the analyses on the separate speaker groups, we ran independent-samples t-tests between native and nonnative speakers on the various test items, to see if and where native and nonnative participants' performance differed.

Individual difference variables. An independent-samples t-test between native and nonnative speakers for WMC and LLAMA scores showed that native and nonnative speakers did not score significantly differently on WMC or LLAMA_B, _D, _E, or _F measures, all $p > .05$, indicating that language aptitude and WMC are not affected by native or nonnative speaker status.

Forced choice English grammar test. Another independent-samples t-test showed that native speakers did not score significantly higher on the difficult syntactical structures ($M = .6632$) than nonnative speakers ($M = .6208$), $t(62) = .821$, $p > .05$. In addition, nonnative

speakers did not score significantly better on homophone structures (M = .9929) than native speakers (M = .9582), t (62) = -1.568, p > .05.

Auditory grammaticality judgment test. An additional independent-samples t-test between native and nonnative participants for auditory task scores showed that native speakers performed significantly better on the overall auditory test (M = .8535) than nonnative speakers (M = .7840), t (62) = 4.160, as well as on grammatical items (M_{nat} = .8983, M_{non} = .8253, t (62) = 3.899) and on ungrammatical items (M_{nat} = .8149, M_{non} = .7483, t(62) = 2.797), all p < .05.

Specific item types. A final independent-samples t-test showed that, while natives scored better than or equal to nonnatives on all AGJT items and on most Forced-Choice English Grammar Test items, there were a few FCEGT items wherein nonnatives scored better than natives. Specifically, nonnative speakers scored significantly better on "its/it's" items (M = .9750) than natives (M = .9074), t = -1.102, and better on "their/there/they're" items (M = 1.0) than natives (M = .972), t = -1.037, p < .05.

Discussion

Hypotheses Addressed

Based on the previous literature, we expected to see age effects for nonnative speakers; specifically, the later the age of exposure, the lower the proficiency scores and the higher the variability would be. We did see an effect of age of arrival (immersed exposure) on nonnative performance, with a significant negative relationship between Age of Arrival and the various AGJT scores. This indicates that, as previous literature supports, later AoA corresponds to lower scores. However, because the sample was so small, we were not able to examine early and late learners separately. We did separate them in a t-test for curiosity's sake, and saw significant differences between early and late learners in the AGJT, and the expected wide variability in the

Forced-Choice English Grammar test and in those who arrived after age 11. In addition, there was no clear effect of AoA on Forced-Choice English Grammar Test scores. We suspect this is due primarily to the small sample size, though it could also be due to the large number of Foreign-language graduate students in the nonnative sample, who would have had extensive linguistics training as part of their curricula. Additionally, though the native speakers did score higher than nonnatives on the AGJT, there was no significant advantage of native speakers over nonnatives in the Forced Choice English Grammar Test items.

Johnson and Newport (1989) had data that suggested the presence of a critical period for language learning, after which age effects became variable and unpredictable. As mentioned before, we did see suggestions of a possible critical period after age 11, but because our sample size was so small, it is impossible to say definitively.

We expected to see effects of language aptitude: we predicted that all native speakers, all early learning nonnative speakers, and high aptitude late learning nonnative speakers, would score significantly better than low aptitude late learners. Again, we were unable to examine the early vs late learner effects on nonnatives, though native speakers did score significantly better on the AGJT than nonnatives. Due, we think, to the small sample size of our nonnative group, we did not see any effects of WMC or language aptitude on nonnative scores on the AJGT or FCEGT; if a larger sample size had been obtained, we might have seen the expected relationships. However, there was a clear effect of WMC and LLAMA on the native speakers—which was not present in the previous literature (Abrahamsson and Hyltenstam, 2008—they saw no relationship between Language Aptitude and performance for natives). This may mean that with a harder language test, individual difference variables will come more into play and individual differences overall will be more apparent.

In addition, we expected that aptitude effects and score spread would be similar for early learners and native speakers, and that early learners would fall largely within native range. Indeed, we did not see any significant differences in WMC and LLAMA scores between natives and nonnatives, and we did see a wide range of scores for both natives and nonnatives on the proficiency measures, wherein all nonnatives fell well within the spread of natives for the Forced Choice English Grammar Test, and some for the AGJT. We also expected that the overall effect would be greater for late learners than for early learners and native speakers, when in fact we only saw a significant effect of language aptitude for native speakers.

Finally, we expected to see effects of working memory capacity (WMC); we hoped for a spread in native speakers' scores, and we predicted that high WMC participants would score better across groups than low WMC, and that high WMC would correlate with high language aptitude. We found that there were no participants who made a perfect score on either the Forced Choice English Grammar Test or the Auditory Grammaticality Judgment Test, and that there was a good spread of scores among both the native and nonnative speakers. There was also a positive relationship between weighted average score on the FCEGT and WMC and LLAMA scores for native speakers, and WMC correlated significantly with high LLAMA scores. This is as was expected—and we are glad to have avoided the ceiling effect seen in previous literature. This means that native participants do actually differ in linguistic ability, WMC, and language aptitude, casting ever more doubt on the concept of "native-like" performance in nonnative speakers. We also predicted that WMC would have more of an effect on late learning native speakers than on early learners or native speakers, but because we were unable to examine early vs late learner performance, and because our small nonnative sample resulted in no significant

correlations between proficiency measure scores and WMC or LLAMA scores, we were unable to explore this prediction.

Order of Difficulty Effects

Though both native and nonnative speakers had the easiest time with conjunctions, nonnative speakers found the five hardest structures (Parallel Structure, Number Agreement, Than, From/to, and Tense) to be equal to all the other structures; native speakers found Parallel Structure, Number Agreement, Than, and Tense to be the hardest, differing significantly from the other item types. The order of difficulty for syntactical structures was largely the same for both groups, so it is possible that with more nonnative participants, we could have seen more difference in difficulty of items. Auditory items were, for nonnative speakers, largely the same in difficulty (with the exception of a few items, they were all statistically the same as the other structures within their grammaticality groups), whereas native speakers found Superlatives, Comparatives, Relative Pronouns, Many/Much, Subject/Verb Agreement, Pronoun Case, and Less/Fewer items to be more difficult than the others. Both, however, found Regular Plural items to be among the easiest. These results differed somewhat from those of Johnson and Newport, 1989.

The hardest structures for native speakers were Than, Tense, and Number Agreement items; we expected Tense and Number Agreement due to the longer sentences and multiple structures in each item, but we suspect Than may have been equal to chance because there were only two items, one of which could also qualify as Parallel Structures (rather than because Than items are actually difficult). The Than item in question is as follows:

When Black Friday sales rolled around, Aunt June bought over $2000 worth of items, 21% more than she did last year, 30% more than Grandma Lois did, and almost twice as much as Uncle Jim's purchases.

- than Grandma Lois did, and almost twice as much as Uncle Jim's purchases.
- than Grandma Lois's purchases, and almost twice as much as Uncle Jim.
- than Grandma Lois bought and almost twice as much as Uncle Jim's purchases.
- than Grandma Lois and almost twice as much as Uncle Jim's purchases were.
- than Grandma Lois bought, and almost twice as much as Uncle Jim bought.

Johnson and Newport (1989) chose to create a GJT made up of basic English structures, which they found was extremely simple for native speakers. In contrast, we tried, and evidently we succeeded, to make up a test that would be extremely difficult for both native and nonnative speakers; as such, some of the structures tested in this study differed from those of Johnson and Newport, as did the most difficult items. Though they found Determiners and Plurals to be the most difficult items for late learners, and all other items to vary widely in performance for Native Speakers, we had clear-cut data indicating that Tense, Than, and Number Agreement are consistently more difficult for Native Speakers. As an example of how our items differed, compare the following examples of Johnson's Verb Tense items to examples of ours:

Johnson's item:

The little boy is speaking to a policeman.

The little boy is speak to a policeman.

Yesterday the hunter shot a deer.

Yesterday the hunter shoots a deer.

Our difficult syntactical item:

Honor society members who attended this year's officer elections meeting find, to their surprise, that they were not familiar with any of the candidates.

- meeting find, to their surprise, that they were not familiar with any of the candidates.
- meeting find, to their surprise, that they are not familiar with any of the candidates.
- meeting were surprised to find that they are not familiar with any of the candidates.
- meeting found, to their surprise, that they were not familiar with any of the candidates.
- meeting found, to their surprise, that they are not familiar with any of the candidates.

Our auditory item:

She has never heard the joke I am about to tell.

She had never heard the joke I am about to tell.

She have never heard the joke I am about to tell.

Though our auditory GJT items do not differ much in structure or difficulty from those of Johnson and Newport, our Forced-Choice English Grammar test items (of which Tense was found to be the most difficult) are much more complex, and may often result in increased memory load and difficulty. In the auditory task, we did find that Tense was not the most difficult item group for natives or for nonnatives—however, Plurals were not the most difficult in this case, either. An example of our Plurals items follows:

Four geese chased us around the lake.

Four gooses chased us around the lake.

Four geeses chased us around the lake.

In comparison to our Difficult Syntactic Structures, it is not difficult to see that Plurals would be significantly less taxing for participants. Of course, we must also take into account the native languages from which our participants came—a large number of our nonnative participants had romance languages as their mother tongues, which tend to have structures in common with English and which may have made plurals and other such items comparatively easier for our participants. In contrast, Johnson and Newport (1989) chose all Chinese and Korean speakers as participants, specifically in order to judge the performance of participants from a highly dissimilar native language.

In all, that the order of difficulty seen in our study is so dissimilar from what we may have expected based on Johnson and Newport is not highly surprising—with the increased complexity of our test items, varying linguistic background of participants, and the increased memory load of some of our double-structure test items, it seems almost certain our results would differ.

Some Limitations

One important limitation to consider is the inherent confound in examining Age of Arrival; the age when a person arrived in a country is usually also related to the amount of time the person has, at present day, been in the country. Typically, the earlier the person came to the country, the longer they have been living there, and practicing the language in an immersion setting. The later a person has arrived, typically, the less time they have spent immersed. We ran an additional bivariate correlation showing that, indeed, the number of years in the US (or other English-speaking country) correlated positively and significantly with scores on the AGJT. The length of time in the English-speaking area, therefore, may actually be the predictor of performance, rather than the *age* at which a person was immersed.

The English structures used were based on "textbook" English, meaning that participants from different dialectical backgrounds may have had a harder time; also, since common usage is not always necessarily grammatically correct, native speakers may not have scored well on certain structures. For instance, nonnatives scored better on "its/it's" and "their/there/they're" items in the Homophones structures, which might suggest that the widely-accepted incorrect usage of these words by native speakers does not cross over to nonnative speakers. We think its/it's structures may have been harder (significantly for natives, nominally for nonnatives) due to the seemingly arbitrary assignment of the apostrophe to "it is" rather than possessive 'its'. In English, possessives are generally formed by adding 'apostrophe s' to the word or name possessing the item—in this case, it may be unclear to participants whether 'it's' refers to the possessive or contractive form. Interestingly, while their scores on Difficult Syntactic items varied widely, all but one nonnative (who missed one 'its/it's' item) scored perfectly on the Homophone items. Nonnatives may have had a stricter rote rule-based English education than natives, which could account for their better performance on the Homophones.

An additional limitation is the small number of nonnative participants; we experienced unexpected difficulty in recruiting nonnative participants, resulting in a very small sample size, n = 10. This will decrease our generalizability. In addition, some of our nonnatives (due to the small pool we managed to draw from) may not have been functionally fluent in English. One nonnative participant stated that she believed she had done poorly on the AGJT, because she is used to speaking to nonnative English speakers and overlooking the types of errors that were tested in the AGJT (which lines up with the overall poorer performance of nonnatives on the AGJT).

Another potential limitation comes in the form of one of our individual difference measures, the LLAMA. A number of researchers in the previous literature, including Bylund et al. and Abrahamsson and Hyltenstam, used the Swansea test, which is why we chose to use the LLAMA (derived from the Swansea) as the Language Aptitude measure for this study. One drawback of the LLAMA, however, is that its manual provides no guide for a standardized overall score, giving different score levels for each of the tasks. Previous researchers used an older version of the examination, from 2005, whose overall score was expressed as 100%, and assigned High and Low labels to participants based on the average score (Bylund et al., 2009; Abrahamsson & Hyltenstam, 2008). This may not have been the best approach—given that the tasks are split into four distinct aspects of aptitude, and that the four tasks are not consistently visual or auditory in nature, an overall score may not be representative of individual differences in each of the functions. However, this means that it is difficult to get a measure of overall language aptitude or of the LLAMA's overall effectiveness, and causes the data to be clunky and difficult to work with. Another issue with the LLAMA is that it is not truly a test of pure language aptitude; the manual states that scores above a certain level are indicative of linguistic training. If training can alter Language Aptitude, perhaps it is the level of training, and not an innate individual factor, that is predictive of performance in a language; and given that some of our nonnative participants were foreign language TAs, their scores may have been affected by their linguistics training and heightened awareness of grammatical issues.

Based on the correlational data with the difficult syntactical items (which seem to discriminate between individuals better than any of the other structure types), it would seem that the most important tasks for performance prediction are, in order, LLAMA_F (grammatical

inferencing), LLAMA_B (vocabulary learning), LLAMA_E (symbol-sound association), and then LLAMA_D (sound recognition); LLAMA_D is ranked the lowest on the list because the more difficult test was written, and the LLAMA_D, an entirely auditory task, did not correlate with performance on the written task. Abrahamsson and Hyltenstam (2008) utilized both a written and auditory task, so the LLAMA_D might have had some significance (similar to the marginally significant correlation we saw with the auditory task). In the future, perhaps a weighted average could be obtained, or perhaps only one or some of the LLAMA tasks could be administered to participants.

In Conclusion

Our study suggests that individual difference variables like working memory capacity and language aptitude are predictive of native performance on a difficult test. In addition, nonnative English-speakers did not perform significantly differently from natives on such a test. However, in an auditory grammaticality judgment test, easier than the other test, nonnatives struggled more than natives. Though this trend may have been because of our small nonnative sample, it is also possible that, in general, nonnative speakers may study reading and writing more than verbal and auditory tasks, and that the phonological processes necessary for the AGJT are not developed as well during training as during native immersion. Perhaps, due to the nature of the AGJT (which was similar in construction to the Grammaticality Judgment Tests used by previous researchers), participants felt time pressure, and did not take the time to review the sentence in their heads before making a judgment; whereas, on the FCEGT, they had plenty of time to read and reread the sentences in order to determine the correct response. Additionally, despite the fact that Age of Arrival predicts performance on an AGJT, there is no indication that

it is predictive of performance on a more difficult, written test; and Age of Arrival in an English-speaking country may be confounded by total length of time in an English-speaking country.

References

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

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language, 44*. 235-249

Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2009). The role of language aptitude in first language attrition: the case of pre-pubescent attriters. *Applied Linguistics, 31 (3)*, 443-464

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language, 63 (3),* 544-573

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*. 60-99

Lester, M. (2011). *Advanced english grammar for ESL learners.* McGraw-Hill Companies, Inc.

Lusk, D. L., Evans, A. D., Jeffrey, T. R., Palmer, K. R., Wikstrom, C. S., & Doolittle, P. E. (2009). Multimedia learning and the effects of individual differences: mediating the effects of working memory capacity with segmentation. *British Journal of Educational Technology, 40 (4),* 636-651

Meara, P. (2005). *LLAMA Language Aptitude Tests* [Measurement instrument]. Retrieved from http://www.lognostics.co.uk/tools/llama/

Oertel-Knöchel, V., Mehler, P., Thiel, C., Steinbrecher, K., Malchow, B., Tesky, V., Ademmer, K., Prvulovic, D., Banzer, W., Zopf, Y., Schmitt, A., & Hänsel, F. (2014). Effects of aerobic exercise on cognitive performance and individual psychopathology in depressive

and schizophrenia patients. *European Archives of Psychiatry and Clinical Neuroscience*, *264 (7),* 589-604, Feb 2, 2014

Papesh, M. (2014). STM and Working Memory (WM) [PowerPoint slides].

Table 1

Construction	**Example of correct base sentence**
Tense Agreement	Spectators at basketball games throughout the season who happened to look up at the ceiling often saw balls lodged in the rafters, which the court staff had been trying to remove.
Number Agreement	Most experts agree that, though the number of participants in the extremely popular snowshoe baseball has increased with each year, the sport will soon die off into obscurity.
Conjunctions	Since it turned out that the dragon had been slain years ago by the princess herself, the knight found himself with no damsel to rescue and no quest to complete.
Word Order	Those who read the biography of J. R. R. Tolkien soon learn that before the Hobbit became part of the Lord of the Rings mythology, it was a bedtime story that he initially wrote for his children.
Parallel Structures in Lists	In order to help students keep in shape, the campus fitness instructor suggested a variety of exercises, including running laps, throwing a medicine ball, and lifting weights.
Comparatives	When Black Friday sales rolled around, Aunt June bought over $2000 worth of items, 21% more than she did last year, 30% more than Grandma Lois bought, and almost twice as much as Uncle Jim bought.
Relative Pronouns	The Oscars, a televised awards ceremony in which actors and filmmakers are awarded for exemplary performances, is under fire for lacking diversity in its nominee pool, its list of winners, and the Academy in general.
Since/Present Perfect	Since ten years ago, when a clumsy student blew up the chemistry lab, the university has used a variety of buildings to house chemistry classes while awaiting the rebuilding of the old lab.
Ranging From/To	Fresh Food Whole Market has recently introduced a new line of products, ranging from delicious non-homogenized French milk that can be left out for six months if unopened, to vegetable juice without pulp, sugar, or flavors of any kind.

Table 1: Examples of long-format written grammaticality judgment test items, by construction.

Table 2

Subject and Type	#	Example	Error Types
Comparative	6	The man thought the pug was (uglier/more ugly) than the greyhound.	Different adjectives have different set correct comp forms
Superlative	6	Rachel made the (civilest/most civil) argument she could think of.	Different adj. have different set correct super forms
Plurals	**12**		
Irregular	4	The man's (feet/foots/feets) hurt after he walked fifteen miles.	Correct, regularized, double-marked
Singular form = plural form	4	The dog chased a dozen (sheep/sheeps).	Correct, double-marked
Regular	4	My aunt has five (cats/cat).	Correct, single
Number agreement	**12**		
Many/Much	4	There (were many/was much) pencils strewn across the floor.	Many for count nouns, much for mass nouns
Less/Fewer	4	She had (less/fewer) trophies than he did.	Less for mass nouns, fewer for count nouns
Few/Little	4	The teacher had (little/few) patience at the end of the day.	Little for mass nouns, few for count nouns
Relative Pronouns	8	This is the jar (where/in which/in whom/that) the cookies are kept.	Who/whom for animate, which/that for inanimate, which/whom are acted upon, that/who act
Pronoun Case	4	She asked that (I/me) buy the book for her.	Incorrect pronoun case
	2	If you are not sure where to go for the ceremony, ask (whoever/whomever) shows up first.	Whoever acts, whomever is acted upon
Word Order	5	I know where (should we go/we should go) for Karen's birthday party.	Incorrect word order
Verb Tense	7	He will/would do the laundry if he had/has time.	Matching vs. nonmatching tense
Subject/Verb agreement	3	The girl (eat/eats) six cookies.	Matching vs. nonmatching verb conjugation
S/V/A: To Be	3	My mother (is/are/be) not home.	

Table 2: Auditory grammaticality test stimulus subjects, examples, and reasoning behind foils.

Table 3

		LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F	WMC
LLAMA_D	Pearson Correlation	0.034				
LLAMA_E	Pearson Correlation	.416**	-0.152			
LLAMA_F	Pearson Correlation	.546**	0.048	.391**		
WMC	Pearson Correlation	.380**	.446**	0.332*	.380**	
Homophone	Pearson Correlation	0.006	0.022	.273*	-0.113	0.073
Syntactical	Pearson Correlation	.377**	0.218	.307*	.387**	.344*
OverallAud	Pearson Correlation	0.04	0.22	0.093	0.2	0.222
GramAud	Pearson Correlation	0.128	0.258	-0.042	0.07	.320*
UnGramAud	Pearson Correlation	-0.033	0.088	0.129	0.179	0.053

Table 3: Correlation matrix of all test measures for native speakers (LLAMA_B through _F, Working Memory Capacity, Homophone structures, Difficult Syntactical Structures, and overall, grammatical, and ungrammatical auditory GJT items). '*' indicates the value is significant at $p < .05$; '**' indicates that the value is significant at $p < .01$.

Table 4

Structure type	Mean	Group
Than	0.5	a
NumAgree	0.538	a
Tense	0.605*	ab
Parallel	0.737*	abc
FromTo	.81*	bc
RelPronoun	0.895*	c
Order	0.92*	c
Conjunction	0.977*	c

Table 4: Native participants' mean scores on Difficult Syntactical Structures, in order from most to least difficult; '*' indicates that the mean is significantly different from chance (.5).

Table 5

Structure type	Mean	Group
ItsIt's	0.895*	a
TheirThereThey're	0.969*	ab
YourYou're	0.988	b

Table 5: Native participants' mean scores on Homophone structures, in order from most to least difficult; '*' indicates that the mean is significantly different from perfect.

Table 6

Structure type	Mean	Group
Superlative	0.82407	a
Comparative	0.84259	ab
RelativePronouns	0.85613	abc
Many/Much	0.8843	abc
Sub/VerbAgree	0.88889	abc
PronounCase	0.9074	abcd
Less/Fewer	0.91	abcde
Few/Little	0.9167	bcde
Tense	0.92824	cde
IrregularPlural	0.96528	de
WordOrder	0.97	de
RegularPlural	0.98	e

Table 6: Native participants' mean scores on grammatical auditory items, in order from most to least difficult. All $p < .05$, all items significantly different from perfect and from chance.

Table 7

		AoA	AoE	LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F	WMC
LLAMA_B	Pearson Correlation	-0.59	-0.491					
LLAMA_D	Pearson Correlation	-0.075	-0.383	-0.087				
LLAMA_E	Pearson Correlation	-0.103	-0.622	0.192	0.505			
LLAMA_F	Pearson Correlation	0.594	0.163	-0.487	0.052	-0.019		
WMC	Pearson Correlation	-0.408	-0.408	0.790*	-0.034	0.377	-0.013	
Homophone	Pearson Correlation	0.492	0.435	0.061	-0.196	-0.244	0	0.068
Syntactical	Pearson Correlation	-0.085	-0.311	-0.063	0.224	0.542	0.036	0.373
OverallAud	Pearson Correlation	-0.777**	-0.533	0.303	0.267	0.379	-0.365	0.445
GramAud	Pearson Correlation	-0.706*	-0.621	0.257	0.6	0.364	-0.242	0.367
UngramAud	Pearson Correlation	-0.671*	-0.353	0.275	-0.047	0.312	-0.385	0.412

Table 7: Correlation matrix of all test measures for nonnative speakers (Age of Arrival, Age of Exposure, LLAMA_B through _F, Working Memory Capacity, Homophone structures, Difficult Syntactical Structures, and overall, grammatical, and ungrammatical auditory GJT items). '*' indicates the value is significant at $p < .05$; '**' indicates that the value is significant at $p < .01$.

Table 8

Itemtype	Mean	Group
Parallel	0.5	a
Numagree	0.5	a
Than	0.55	ab
Fromto	0.6	ab
Tense	.6125*	a
Order	0.7	ab
Relpron	.85*	ab
Conjunction	.95*	b

Table 8: Nonnative participants' mean scores on syntactical items in order from most to least difficult; '*' indicates different from chance.

Table 9

Itemtype	Mean	Group
Less/Fewer	0.433	a
WordOrder	0.62	**
RelPronouns	0.631	ab
Tense	0.667	abc
Comparatives	0.68	abc
Superlatives	0.7	abc
IrregPlurals	0.721	abc
PronounCase	0.817	abc
Few/Little	0.875	bc
Many/Much	0.925	c
SV Agreement	0.933	**
RegPlurals	0.963	c

Table 9: Nonnative participants' average scores on ungrammatical auditory items, in order from most to least difficult; '*' indicates different from chance.

**Word Order does not differ from any other item, and Subject/Verb Agreement only differs from Less/Fewer and Tense.

Table 10

Itemtype	Mean	Group
Few/Little	0.75	a
RelPronouns	0.762	a
Tense	0.788	a
Superlatives	0.8	a
PronounCase	0.8	a
Comparatives	0.817	a
IrregPlurals	0.85	a
Less/Fewer	0.867	a
SV Agree	0.875	a
WordOrder	0.9	a
Many/Much	0.9	a
RegPlurals	0.925	a

Table 10: Nonnative participants' average scores on grammatical auditory items, in order from most to least difficult; '*' indicates different from chance.

Figure 1

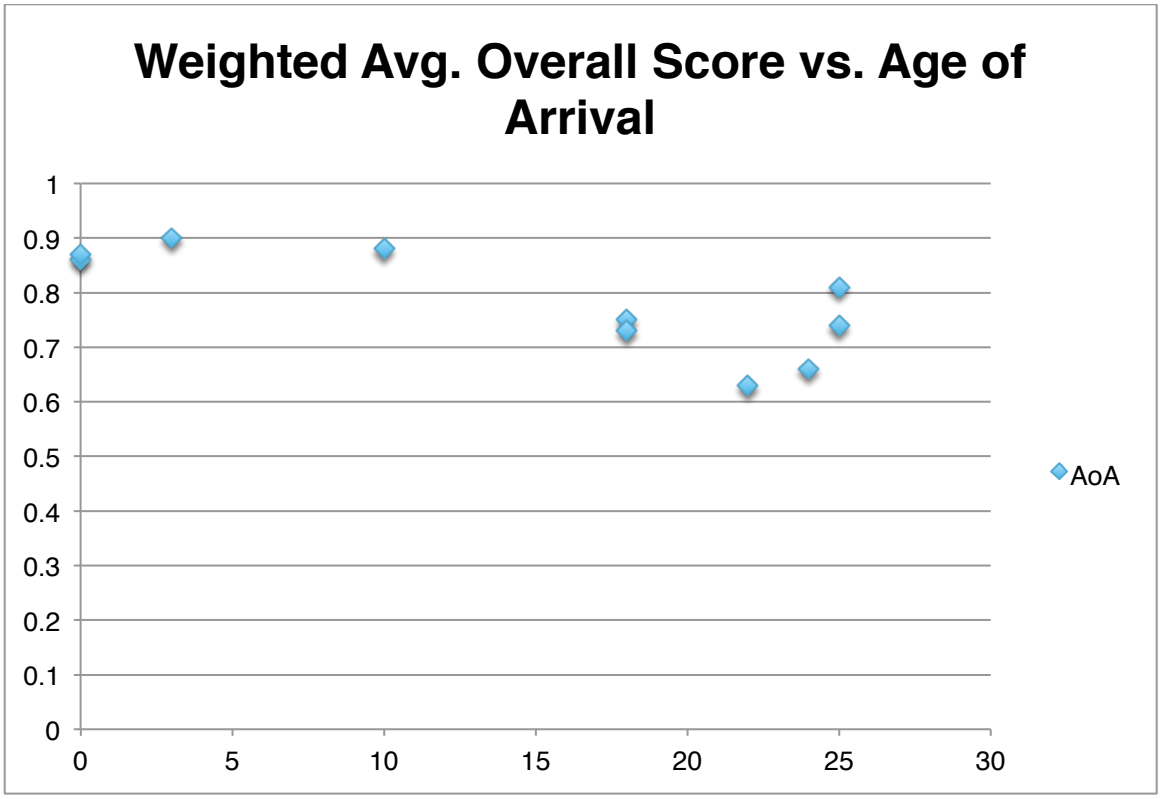

Figure 1: Nonnative participants' weighted average overall scores on the AGJT vs participants' Age of Arrival.

Figure 2

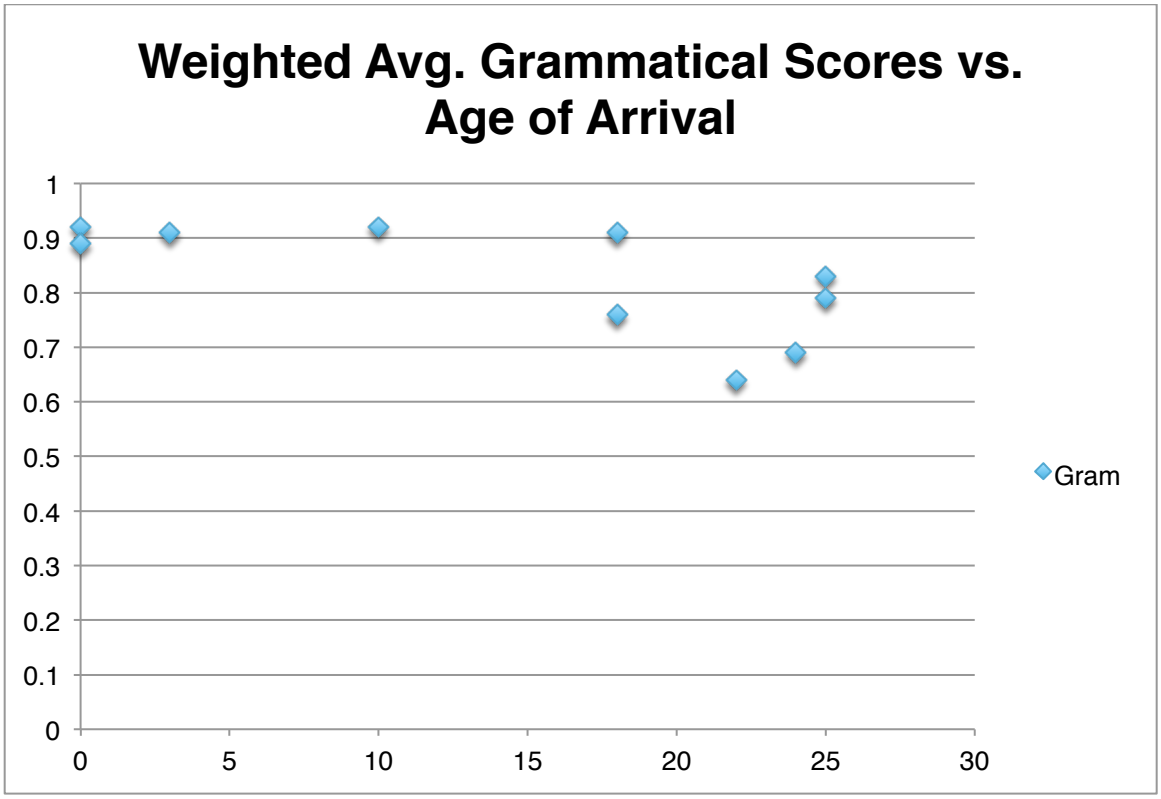

Figure 2: Nonnative participants' weighted average scores on grammatical items of the AGJT vs participants' Age of Arrival

Figure 3

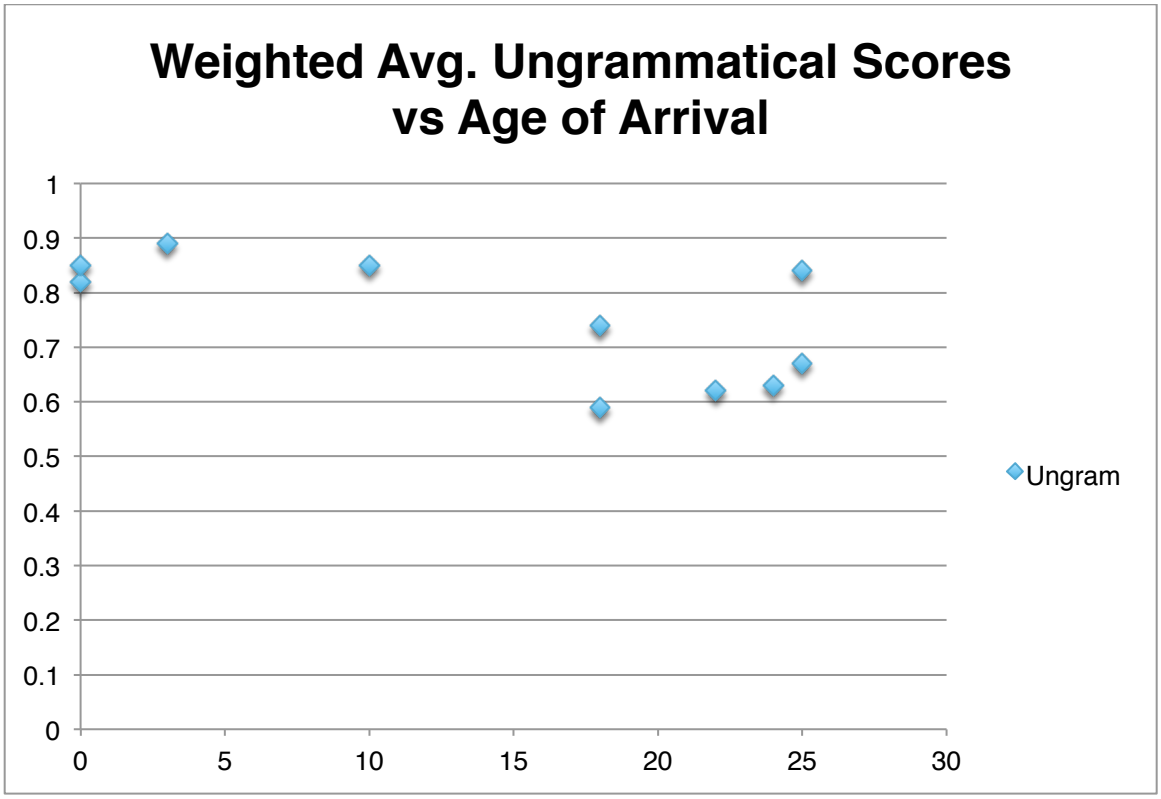

Figure 3: Nonnative participants' weighted average scores on ungrammatical items of the AGJT vs participants' Age of Arrival.

Appendix 1

Demographics Questionnaire

Major:
Current age:
Gender:
First language:
Years studied in high school:
Semesters studied in college:
Rate your fluency: 1 2 3 4 5 (1 being lowest and 5 being highest)
Second language:
Years studied in high school:
Semesters studied in college: Age you first started learning L2?
Rate your fluency: 1 2 3 4 5
Additional languages, and years studied:

_____
_____
_____

Where have you lived in the past? For how long? What language did you primarily speak and hear?

Location	**Length of time (or ages)**	**Primary language**

Where have you visited in the past? For how long? How frequently? What language?

Location	**Length of time**	**Frequency (and # years)**	**Primary language**

What percent of your daily speech/conversation is in English? _____
Do you like to read?
Yes/ Somewhat/ No
How many books have you read in the past year? _____

Appendix 1, cont.

How often do you read for class?

Less than once a month/Once a month/ Once a week/2-3 times a week/4-5 times a week

How often do you write for class?

Less than once a month/Once a month/ Once a week/2-3 times a week/4-5 times a week

How often do you write letters or email?

Less than once a month/Once a month/ Once a week/2-3 times a week/4-5 times a week

Do you use Facebook or other social networking sites? Yes/No

If so, how often? (Daily, weekly, more than once a day, a few times a week, etc.)

Do you read the news? Yes/No

How often? (Daily, weekly, more than once a day, a few times a week, etc.)

Would you say you identify more closely with an individualistic culture (more Western) or a community-based culture (more Eastern)? _____