Studies in Second Language Acquisition **41** (2019), 313–336
doi:10.1017/S0272263118000256

Research Article

COGNITIVE APTITUDES AND L2 SPEAKING PROFICIENCY

LINKS BETWEEN LLAMA AND HI-LAB

Gisela Granena*

Universitat Oberta de Catalunya

Abstract

This study investigated the underlying structure of a set of eight cognitive tests from the two most recent language aptitude test batteries: the LLAMA (Meara, 2005) and the Hi-LAB (Linck et al., 2013) to see whether they had any underlying constructs in common. The study also examined whether any of the observed constructs could predict L2 speaking proficiency in terms of complexity, accuracy, or fluency. Participants were 135 college-level students learning Spanish as an L2 in the United States. Results showed that the LLAMA and the Hi-LAB include tests that tap the same constructs. Specifically, the tests from the two batteries loaded onto three different factors, interpreted as "Explicit Aptitude," "Implicit Memory Ability," and "Implicit Learning Ability." The results further showed that Implicit Memory Ability was a significant predictor of L2 speed fluency and interacted with Implicit Learning Ability as a predictor of lexical complexity. This finding suggested that L2 learners with greater Implicit Memory Ability may be better at accessing and retrieving previously learned or known information effortlessly. In the case of lexical complexity, the effect of Implicit Memory Ability depended on the level of Implicit Learning Ability.

LITERATURE REVIEW

LANGUAGE APTITUDE

Language aptitude is a predisposition to learn languages well or rapidly (Carroll, 1981). This special talent for language that can predict how fast a second language (L2) is acquired (i.e., rate of attainment) or how successfully it is acquired (i.e., ultimate attainment) is componential in nature (Carroll, 1962; Skehan, 2002). Despite the

This work was supported by a Spencer Foundation Small Research Grant #201500053 to Gisela Granena and Yucel Yilmaz. The authors would like to thank the Center for Advanced Study of Language (CASL) for providing us with web-delivered versions of the Hi-LAB tests and for scoring and sending us the data, blind of any specific hypothesis about the results.

*Correspondence concerning this article should be addressed to Gisela Granena, Universitat Oberta de Catalunya, Av. Tibidabo 39-43 08035 Barcelona, Spain. E-mail: ggranena@uoc.edu

Copyright © Cambridge University Press 2018

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

consensus on its multicomponential nature, various aptitude models have been proposed, each with different subcomponents and ultimately a different understanding or conceptualization of the construct. One of the first models, and still the most influential, was Carroll's (1962) four-factor aptitude model. Carroll derived aptitude components in a bottom-up fashion after running a series of factor analyses on a battery of approximately 20 verbal tests that were felt to be good predictors of success in intensive foreign language training. The tests were administered to male military personnel between 1953 and 1958. Carroll's job was very clear and that was to help the U.S. Army find and train people who would learn foreign languages with ease in short periods, a need that became obvious with the events of World War II. The main goal of Carroll's project was, therefore, predictive rather than construct validity. More specifically, Carroll investigated the extent to which aptitude could predict initial foreign language learning rate as measured by academic course grades.

Carroll's model included phonetic coding ability (codifying, assimilating, and recalling phonetic material), rote memory (ability to recall words in an unfamiliar language), grammatical sensitivity (identifying grammatical functions of words or phrases in sentences), and inductive language learning ability (figuring out the rules of systematically varying language materials). Three of these four factors gave rise to the Modern Language Aptitude Test battery (MLAT; Carroll & Sapon, 1959). Inductive language learning ability was measured in the initial battery of tests, but it was not measured to any appreciable degree in the final five-test battery (Carroll, 1962).

Similar test batteries to the MLAT are the Defense Language Aptitude Battery (DLAB; Petersen & Al-Haik, 1976), only in use by the U.S. Department of Defense, and the Pimsleur Language Aptitude Test Battery (PLAB; Pimsleur, 1966), for young learners between ages 12 and 17. The DLAB was developed by the Defense Language Institute after the MLAT as a battery for differential prediction, to decide in what language or language family a candidate might do best. It mainly measures the ability to learn the grammatical rules of systematically varying language materials. The PLAB, presented as an alternative to the MLAT for the junior high school level, measures vocabulary knowledge in English, language analytical ability, sound discrimination, and sound–symbol associations.

These early test batteries, which have shaped our understanding of language aptitude for decades, suffer from various shortcomings. First, they were developed for selection purposes and, therefore, had predictive validity as a goal; they were not concerned with revealing the underlying structure of the construct. Second, they are characterized by being heavily weighted in favor of explicit cognitive abilities, especially rote (or explicit associative) memory, and language analytical (or explicit inductive learning) ability. As a result, aptitude has been mostly understood as an ability to consciously reflect on linguistic structure and to memorize (a good predictor of outcomes under methods such as the audio-lingual), rather than as an ability to acquire patterns without awareness.[1]

The impact of this conceptualization on the predictive accuracy of the MLAT in learning contexts different from the ones for which the MLAT was originally conceived for is obvious. In fact, an internal report by the U.S. Department of Health, Education, and Welfare that summarized Carroll's findings already suggested that the MLAT could predict some types of learning better than others: "[R]eading and writing superiority tend to be chiefly predicted by section IV of the MLAT, 'Grammatical Sensitivity;' by time not spent abroad; and third college literature courses" (Markessinis, 1968, p. 28),

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

whereas "active performance is predictable by relatively lower scores on section III of the MLAT, time abroad, occasional or frequent home use of the foreign language; early language study; introductory and intermediate courses; [and] poor rote memory" (Markessinis, 1968, p. 28).

More recent aptitude models lack the strong empirical support the MLAT has in the literature but have been able to provide more detailed conceptualizations of the construct. Skehan (2002) and Robinson (2002) have taken a processing stage approach in which the aptitudes needed for each stage of L2 learning may differ. Skehan (2002) proposed four broad processing stages (noticing and registering input, patterning, controlling analyzed knowledge, and lexicalizing) and matched them with relevant cognitive abilities and processing resources. Robinson's (2002) model included a number of aptitude complexes that are differentially related to language processing under different instructional conditions. The model predicted interactions between individual cognitive abilities and instruction type, with the underlying assumption that matching learners to instructional options will optimize learning.

The strength of Skehan's and Robinson's models lies on the fact that both rely on key issues in SLA theory and pedagogy, such as different processing stages in acquisition and the type of information processing involved in different learning contexts. They also consider the advances in the understanding of human memory systems that have taken place since the MLAT. However, unlike the MLAT, these models have not been developed into test batteries that could be empirically tested. As a result, their predictive validity is largely unknown. A model that was developed into a test battery with the objective of being an alternative to the MLAT is the Cognitive Ability for Novelty in Acquisition of Language-Foreign (CANAL-F; Grigorenko, Sternberg, & Ehrman, 2000). Like Skehan's and Robinson's models, CANAL-F takes acquisition processes as its theoretical base. However, it has not enjoyed the popularity of the MLAT, perhaps due to features such as its duration (i.e., three hours) and format (i.e., paper and pencil), which might have made its administration more challenging.

The two most recent language aptitude test batteries, both computer delivered and, therefore, easier to administer, are the LLAMA (Meara, 2005) and the High-Level Language Aptitude Battery (Hi-LAB; Doughty et al., 2010; Linck et al., 2013). The LLAMA grew out of a series of projects by MA students at the University of Wales working on aptitude and its measurement and, as pointed out by an anonymous reviewer, it has not been properly validated[2] (but see Rogers et al., 2016 and Rogers et al., 2017, for validity-related analyses of the LLAMA). Although largely based on the MLAT, the LLAMA departs from the MLAT in two main ways. First, the tests are language independent and, therefore, suitable for speakers of any language with a Roman alphabet. Second, the battery includes a sound recognition test that is not present in the MLAT. The test measures the ability to recognize patterns in spoken language (i.e., sequence learning ability), an ability involved in the discovery of language structure by way of statistical properties of the input and that, as Meara (2005) argues, can help learners recognize the small morphological variations in a language. Based on the fact that it does not include a study-phase and that it relies on feelings of familiarity involving fast, automatic processes, rather than on conscious recollection involving slow, controlled search processes, Granena (2013a) argued that the learning conditions created by the test are closer to

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

implicit induction (i.e., acquiring patterns unintentionally through exposure) than explicit induction (i.e., figuring out rules and relations).

The Hi-LAB (Linck et al., 2013) was created to predict high-level L2 attainment with a focus on interpersonal communication. The battery measures a set of 11 cognitive and auditory perceptual abilities grouped into six different constructs, several of which were not present in the MLAT or in any of the first aptitude test batteries. Two of these constructs, long-term memory retrieval and implicit learning, rely on abilities in the domain of implicit cognitive processes. Long-term memory retrieval is measured by a semantic priming task showing subconscious meaning associations (i.e., the ability to activate semantic networks in an unconscious way), whereas implicit learning is measured by a serial reaction time task showing subconscious pattern-learning (i.e., the ability to pick up on patterns without needing to think about them consciously). Except for processing speed, all the other constructs in the battery (working memory, explicit associative memory, and perceptual acuity) fall well under abilities that engage explicit, attention-driven cognitive processes.

These two aptitude test batteries show that the current understanding of aptitude is that of a multicomponential construct encompassing abilities from both the implicit and explicit cognitive domains. Some of these abilities differ from conventional aptitudes in that they involve cognitive *processes*. The possibility of considering learning and memory processes as a domain of aptitude constructs was, in fact, already entertained by Carroll and Maxwell (1979). Their review acknowledges the promising contribution of experimental cognitive psychology in identifying cognitive processes in perception, learning, memory, problem solving, and other mental activities as cognitive aptitudes. This involves the study of individual differences in task performance as measured by reaction times, error rates, or other indicators.

In sum, the advances in the conceptualization of aptitude have followed calls for new aptitude constructs in educational psychology (e.g., Woltz, 2003) and cognitive psychology (e.g., Kaufman et al., 2010) and are more aligned with developments in our current understanding of how second languages are learned and taught (Long & Doughty, 2009).

LANGUAGE APTITUDE AS A PREDICTOR OF L2 PROFICIENCY

There are multiple factors that account for individual variation in language learning success, but language aptitude is considered one of the best predictors, even if the remaining variation is still considerable and explained by a variety of other factors. Carroll's research (e.g., Carroll, 1963, 1969) showed the power of aptitude as a predictor of course grades in initial foreign language learning (i.e., learning rate). The MLAT yielded correlations in the range of .4 to .6 (Carroll, 1981). Similar predictive validity coefficients were reported for the PLAB and the DLAB. Ehrman and Oxford (1995) also found that aptitude was the strongest predictor of language proficiency (up to .51), among predictors that included learning strategies, learning style, personality, motivation, and anxiety in an intensive language learning course. Similar results were reported by Sparks et al. (1995), in this case with high school foreign language learners. These studies, all of which used the MLAT as a measure, indicate that aptitude is a relevant predictor of academic language proficiency in instructed settings. In fact, a

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

recent meta-analysis (Li, 2016) examining the construct validity of aptitude reported a strong effect size (.49) between aptitude, mostly as measured by the MLAT, and general L2 proficiency when operationalized as course grades or scores on general proficiency tests. The effect size, however, dropped to moderate for specific aspects of L2 learning such as L2 grammar learning (.32), vocabulary, or L2 writing (.34) (Li, 2016).

Most of the research conducted in relation to the predictive validity of aptitude has been done using the MLAT and, therefore, it has followed a traditional conceptualization of aptitude. There is far less research done on the predictive validity of aptitude using more recent conceptualizations and test batteries. Li's (2015) meta-analysis investigating the relationship between aptitude and L2 grammar acquisition, for example, only included three studies that used the LLAMA out of a sample of 33. Li's 2016 meta-analysis investigating aptitude and general L2 proficiency included two out of a sample of 66. These two studies (Abrahamsson & Hyltenstam, 2008; Granena & Long, 2013) were carried out in naturalistic learning contexts with very advanced, near-native L2 speakers. In an instructed setting, Linck et al. (2013) used the Hi-LAB to predict L2 proficiency, operationalized as listening and reading skills as measured by the Defense Language Proficiency Tests. More than 522 L2 learners participated in the study, all members of the U.S. military or personnel from government agencies. The results of the study indicated that listening and reading L2 skills at the most advanced levels were related to phonological short-term memory, explicit associative memory (rote memory), and implicit learning.

These three cognitive abilities were able to distinguish high-attainment L2 learners from control (i.e., mixed-attainment) learners. Phonological short-term memory was a significant predictor of reading skills, while rote memory and implicit learning predicted both listening and reading skills. The strongest relationships identified were between rote memory and listening (.67) and between implicit learning and listening (.51). These findings were considered to give support to the hypothesis that the Hi-LAB would show stronger relations to listening outcomes than reading outcomes and to the fact that it would be a good predictor of spoken language skills, the original aim of the battery.

Despite the evidence provided, Linck et al. (2013) pointed out the need for additional research on the predictive validity of the Hi-LAB constructs, especially in the area of speaking proficiency and subskills such as fluency. The study of the relationship between cognitive factors and linguistic skills in L2 has important theoretical and practical implications. Theoretically, it sheds light on the type of learner internal factors that affect the speed of and success in L2 learning, thus contributing to a more accurate model of SLA. Practically, it provides valuable information that can help tailor instruction to student characteristics, inform curricular design, place or select language learners, or diagnose learning problems.

PURPOSE OF THE STUDY AND RESEARCH QUESTIONS

The present study is part of a broader project that investigated the role of cognitive aptitudes in an instructed L2 learning setting. As part of the project, tests from the two most recent aptitude test batteries available in SLA were administered (i.e., LLAMA and Hi-LAB). The current study reports on the results of an exploratory factor analysis that examined the underlying structure of the Hi-LAB and LLAMA tests employed in the

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

project and the extent to which the observed underlying constructs could predict L2 speaking proficiency in terms of complexity, accuracy, and fluency (CAF).

This study was guided by the following two research questions, which aimed to find out, first, whether there were any constructs in common underlying the tests in the LLAMA and the Hi-LAB aptitude test batteries and, second, whether the constructs identified could predict L2 Spanish speaking proficiency in terms of CAF.

1. Are the LLAMA and Hi-LAB subtests measuring any common constructs?
2. Are any of these constructs good predictors of L2 speaking proficiency in terms of complexity, accuracy, and/or fluency?

METHODOLOGY

PARTICIPANTS

A total of 135 college-level learners of L2 Spanish took part in this study. The requirements to participate were (a) to be a native speaker of English, and (b) to have an intermediate level of Spanish (the equivalent of a 200-level course). All of them responded to a flyer advertising the study and were further screened by means of a background questionnaire. Ninety were female (66.7%) and 45 were male (33.3%). Their mean age at testing was 20 (SD = 4.3) and their mean age of onset was 13 (SD = 4.4). Six had been in Spanish-speaking countries for stays longer than one month and shorter than a year (2, 3, 4, 6, 9, and 10 months). The rest had never stayed abroad for purposes other than travel, and those stays ranged between three days and four weeks. Participants were enrolled in a variety of majors. Twelve were studying Spanish as a minor.

COGNITIVE TESTS

A total of eight tests measuring different cognitive abilities were administered. Four belonged to the LLAMA aptitude test battery (Meara, 2005): LLAMA_B, LLAMA_D, LLAMA_E, and LLAMA_F. Four belonged to the Hi-LAB (Doughty et al., 2010; Linck et al., 2013): Paired Associates, Letter Span, Available Long-Term Memory (ALTM) Synonym, and Serial Reaction Time Task. Three of these tests were selected based on the findings by Linck et al. (2013), which showed that, out of the 11 Hi-LAB tests, Paired Associates, Letter Span, and Serial Reaction Time Task were significant predictors of general L2 proficiency. The fourth test, ALTM Synonym, a priming test, was added to the current study as a novel and underinvestigated measure of aptitude. Each of the tests is described in the following text.

LLAMA_B

LLAMA_B is a vocabulary learning test that measures the ability to learn new words. It is loosely based on the paired-associates test in the MLAT and, therefore, a measure of explicit associative memory. The 20 words to be learned, taken from a Central American language, are presented visually and linked to a target image. Participants are given two

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

minutes to study the word–picture associations. In the testing phase, the program displays one of the words and participants have to identify the correct picture on the screen.

LLAMA_D

LLAMA_D is a sound recognition test that measures the ability to recognize sound sequences. Participants listen to a string of 10 sound sequences that are computer generated and based on words from a British Columbian Indian language. There is no study phase and stimuli are only played once. Then, participants complete a recognition test where they have to discriminate between old and novel items.

LLAMA_E

LLAMA_E is a sound-symbol association test that measures the ability to form novel sound-symbol associations. Similar to the MLAT phonetic script subtest, it requires participants to work out relationships between sounds (i.e., recorded syllables), 24 in total, and a written representation of those sounds in an unfamiliar alphabet. Participants are given two minutes to study the associations. In the testing phase, they hear a combination of two syllables and make a choice between two possible symbol representations.

LLAMA_F

LLAMA_F is a grammatical inferencing test that measures the ability to infer or induce the rules of an unknown language (i.e., explicit inductive learning ability). Participants see a set of pictures and sentences in an unfamiliar language describing the pictures, 20 in total, and are asked to work out the rules that operate in the language. In the study phase, participants have five minutes to click on a series of buttons arranged on the screen. For each button, a picture and a sentence are displayed. In the testing phase, participants see a picture and two sentences and have to choose the sentence they consider correct.

Paired Associates

This is a test that measures explicit associative memory, or the extent to which participants can remember novel associations. Participants' task consists in remembering pairs of words, each composed by an English word and a nonword. The test includes two phases. In the study phase, participants see each pair five times for five seconds each time. In the testing phase, participants are given the nonword and are asked to type the corresponding English word.

Letter Span

This test measures phonological short-term memory. Participants see lists of letters on the screen that they have to recall in the order they were presented. The letters are presented one by one for 900 ms each. The test includes a total of 21 lists of lengths from three to nine. There are three lists of each length, all including consonants, and they are presented in a pseudo-random order. Participants respond by clicking onscreen buttons.

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

ALTM Synonym

This test measures associative priming, or the extent to which the activation of one concept in participants' lexicon will activate related concepts, thus indicating facilitation of lexical processing. The test includes two tasks. In the priming task, participants listen to a list of five words. Then, they are shown two topic words, one of which is a synonym for two of the words in the list and one of which is a synonym for the other three words in the list. For example, a participant can hear *moist smart damp brilliant soggy*, where three words are synonyms of *wet* and two words are synonyms of *intelligent* (Was & Woltz, 2007). Participants' task consists in indicating the word that had more synonyms in the list by answering a question such as "Were there more words meaning wet or intelligent?" In the comparison task, participants see pairs of words, and their task is to indicate whether the meanings of the words are similar or different. There are 18 sets in the comparison task, each of which comes after a priming list from the priming task. Each set includes 12 word pairs, four of which are unscored and eight of which are scored. Nine of the sets are primed and nine are unprimed. In primed sets, one or both words in the pairs are synonyms of one of the two topic words from the priming list (e.g., *wet* or *intelligent*). In unprimed sets, none of the words are synonyms of the two topic words from the preceding priming list.

Serial Reaction Time

This test measures sequence learning and has been proposed as a measure of implicit learning (Willingham, Nissen, & Bullemer, 1989). Participants see four boxes arranged horizontally on the screen indicating the four positions where an asterisk may appear. Participants' task consists in pressing the button that corresponds to the location of the asterisk on a response box. The test includes six blocks of 96 trials with an intertrial interval of 500 ms. Asterisks appear in a pseudo-random order in the first and last blocks. In the rest of blocks, asterisks follow a repeating sequence of length 12.

LANGUAGE TEST

Oral Production Task

A picture taken from a spot-the-difference task from *New English File Intermediate* by Oxford University Press was used as a picture description task. The picture shows a busy street full of people and cars doing various things. Seven of the ongoing scenes were marked with letters A–G, and participants were asked to describe the picture in as much detail as possible paying special attention to the seven activities marked with a letter.

PROCEDURE

Each participant was tested individually at a research lab. The oral production task was administered first. The cognitive tests were administered next on a desktop computer in a randomized order using E-prime 2.0 (Psychology Software Tools, 2011).

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

DATA CODING

LLAMA_B

An accuracy score is automatically computed by the LLAMA software program for every participant out of a maximum of 100. A correction for guessing is automatically done by the software. The internal consistency of the test ($k = 20$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .81.

LLAMA_D

An accuracy score is automatically computed by the LLAMA software program for every participant out of a maximum of 75. A correction for guessing is automatically done by the software. The internal consistency of the test ($k = 30$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .50.

LLAMA_E

An accuracy score is automatically computed by the LLAMA software program for every participant out of a maximum of 100. A correction for guessing is automatically done by the software. The internal consistency of the test ($k = 20$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .72.

LLAMA_F

An accuracy score is automatically computed by the LLAMA software program for every participant out of a maximum of 100. A correction for guessing is automatically done by the software. The internal consistency of the test ($k = 20$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .66. After deleting the last item on the test, Cronbach's alpha was .69.

Paired Associates

An accuracy score was computed by adding the total number of correctly recalled English words. The internal consistency of the test ($k = 20$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .87.

Letter Span

An accuracy score was computed by adding the total number of letters participants recall in the right order for a total of 126 (21 lists, three lists of each length from three to nine). The internal consistency of the test ($k = 21$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .76.

ALTM Synonym

A rate score combining response time and accuracy for each comparison was computed. The score shows correct responses per minute, and it results from dividing the number of

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

correct responses in a set by the total response time within that set. Separate scores for same and different comparisons within each of the nine primed sets are computed. These rate scores within each primed set are then regressed on the corresponding unprimed set. This yields a residual priming score that captures variability in the rate score of primed sets (i.e., dependent variable) that is left after accounting for the variability explained by the rate score of unprimed sets (i.e., independent variable). The 18 residual difference scores resulting from same and different comparisons within the nine primed sets are added to create the final score. A positive score means more priming and a negative score, less priming. The internal consistency of the test ($k = 18$) for the participant sample ($N = 135$), according to Cronbach's alpha, was .70.

Serial Reaction Time

A facilitation score was derived as the difference in median reaction times in the last sequential block (block 5) and the last random block (block 6). The higher the score is, the more facilitation, and, therefore, better sequence learning. The internal consistency of the test, using split-half reliability, was .79.

Oral Production Task

The audio files were transcribed using Praat software, Praat software package (version 5.3.67; Boersma & Weenink, 2014). Recordings were segmented into Analysis of Speech units (AS-units; Foster, Tonkyn, & Wigglesworth, 2000). A "Level 3" analysis, as defined by Foster et al. (2000), was adopted. This excluded minor utterances (e.g., *yeah*, *thanks*), verbatim echoes, and certain turn-initial units without a verb, such as elliptical responses. Foster et al. (2000) have shown that using the AS-unit is the optimal way of dividing transcribed oral data into analyzable units. As defined by Foster et al. (2000, p. 365), an AS-unit is "a single speaker's utterance consisting of an independent clause, or a sub-clausal unit, together with any subordinate clause(s) associated with either."

Complexity was measured syntactically and lexically, based on Skehan's (2009) recommendation to supplement CAF measures with measures of lexical use. Syntactic complexity was measured by calculating the mean number of clauses per AS-unit (Foster et al., 2000). Lexical complexity was measured in terms of lexical diversity (Read, 2000; Skehan, 2003) by means of the Guiraud Index of lexical richness (Guiraud, 1954), which includes a mathematical transformation of the Type-Token Ratio. The Guiraud Index is a widely used measure, which is calculated by dividing the number of word types in a speech sample by the square root of the number of word tokens produced. The square root is introduced to reduce the effect of differences in text length (Vermeer, 2000).

Accuracy was measured by calculating the mean number of errors per AS-unit (Michel, Kuiken, & Vedder, 2007). Errors included syntactic, morphological, or lexical choice errors. This measure was preferred to the percentage of error-free clauses (Foster & Skehan, 1996) based on its greater ability to discriminate among L2 learners with very few error-free clauses.

Finally, fluency was measured as pruned speech rate per minute (i.e., Total number of syllables excluding filled pauses, reformulations, false starts and repetitions/Total

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

seconds of speech × 60). This measure is reported to be a fluency-sensitive measure (e.g., Freed, 2000). Pruned items (i.e., those removed from analysis) included filled pauses, word fragments, repeated words, false starts, replacements, and explicit editing items (Ishikawa, 2008).

DATA ANALYSIS

The first stage in data analysis involved performing a principal components analysis (PCA) on the eight cognitive measures administered in the study: the four LLAMA subtests (LLAMA_B, LLAMA_D, LLAMA_E, and LLAMA_F) and four Hi-LAB subtests (Paired Associates, Letter Span, ALTM Synonym, and Serial Reaction Time). The resulting factor scores were saved as variables.

In the second stage of the analysis, factor scores were used as continuous predictor variables in univariate general linear models. Custom models with factor scores as covariates, including all covariate interaction terms, were computed to better understand any complex relationships among the variables. Separate analyses were conducted for each of the dependent variables (accuracy, fluency, syntactic complexity, and lexical complexity). *Z*-scores were used to standardize the data.

RESULTS

This study had two main goals. First, to explore whether the LLAMA and Hi-LAB test batteries had any underlying constructs in common, and, second, to examine whether any of these constructs were good predictors of L2 speaking proficiency. Table 1 displays the descriptive statistics of the eight cognitive measures,[3] while Table 2 shows the descriptive statistics of the language task in terms of CAF.

Table 3 displays the correlation matrix for all the measures in the study, including predictors and outcomes.

A PCA was first conducted on the eight cognitive tests, to determine any factors on the basis of factor loadings. An oblique rotation method was chosen due to the theoretical expectation that any underlying factors might be correlated. The Kaiser-Meyer-Olkin

TABLE 1. Means, standard deviations, and maximum and minimum values for the eight cognitive tests

	Mean	*SD*	Min.	Max.
LLAMA_B	62.68	22.06	15	100
LLAMA_D	26.22	15.26	0	75
LLAMA_E	70.70	27.0	0	100
LLAMA_F	46.19	24.68	0	100
Paired Associates	11.60	5.23	0	20
Letter Span	88.61	12.69	59	116
ALTM Synonym	−1.66	4.60	−14	10
Serial Reaction Time	32.58	36.86	−90	146

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

TABLE 2. Means, standard deviations, and maximum and minimum values for the language task in terms of CAF

	Mean	*SD*	Min.	Max.
Syntactic complexity	.25	.30	.00	1.67
Lexical complexity	4.64	.75	2.73	6.56
Accuracy	1.54	1.69	.13	.15
Fluency	58.58	29.33	3.72	161.96

(KMO) measure of sampling adequacy (a measure of how suited data are for factor analysis) was .77. Conventionally, values that are higher than .60 suggest that the proportion of common variance among the variables is low (partial correlations should not be very large if distinct factors are expected to emerge). Bartlett's test of sphericity was significant ($p < .001$), indicating that the relationships among the variables were not due to chance. The analysis revealed three factors with eigenvalues greater than 1.0, which accounted for 59.62% of the total variance. The first component accounted for 28.02% of the total variance, while the second and third components accounted for 17.05% and 14.55% of additional variance. According to the rotated solution, five tests loaded on the first component with loadings greater than .44[4]: Letter Span (.55), Paired Associates (.79), LLAMA_B (.75), LLAMA_E (.64), and LLAMA_F (.65). Two tests loaded on the second component with loadings greater than .4: ALTM Synonym (.76) and LLAMA_D (.77). Finally, Serial Reaction Time loaded strongly on the third component (.85) (see Table 4).[5]

The next step in the analysis consisted in using the saved factor scores as predictors of L2 speaking proficiency. Separate univariate general linear models were computed for each speaking proficiency dimension with the three factors as continuous predictor variables. Custom models that included the three predictors and all possible interactions among predictors were used. The general linear model for accuracy accounted for 6% of the total variance ($R^2 = .06$). Factors 1, 2, and 3, labeled Explicit Learning Ability, Implicit Memory Ability, and Implicit Learning Ability, respectively, were not statistically significant: Factor 1, $F(1, 121) = .689$, $p = .408$, $\eta^{26} = .005$; Factor 2, $F(1, 121) = 2.370$, $p = .126$, $\eta^2 = .020$; and Factor 3, $F(1, 121) = 1.376$, $p = .243$, $\eta^2 = .010$. The interactions among predictors were not statistically significant either: Factors 1 and 2, $F(1, 121) = 1.437$, $p = .233$, $\eta^2 = .011$; Factors 1 and 3, $F(1, 121) = .680$, $p = .411$, $\eta^2 = .005$; Factors 2 and 3, $F(1, 121) = .024$, $p = .876$, $\eta^2 = .000$; and Factors 1, 2, and 3, $F(1, 121) = .542$, $p = .463$, $\eta^2 = .004$.

The general linear model for fluency accounted for 8% of the total variance ($R^2 = .08$). In this case, interactions among predictors were not statistically significant: Factors 1 and 2, $F(1, 121) = .070$, $p = .792$, $\eta^2 = .001$; Factors 1 and 3, $F(1, 121) = 2.885$, $p = .092$, $\eta^2 = .023$; Factors 2 and 3, $F(1, 121) = .321$, $p = .572$, $\eta^2 = .003$; and Factors 1, 2, and 3, $F(1, 121) = .030$, $p = .864$, $\eta^2 = .000$. Factors 1 and 3 were not statistically significant either: $F(1, 121) = .494$, $p = .483$, $\eta^2 = .004$ and $F(1, 121) = 1.241$, $p = .267$, $\eta^2 = .009$, respectively. However, Factor 2, the factor labeled Implicit Memory Ability, had a statistically significant effect on fluency scores, $F(1, 121) = 4.308$, $p = .040$, $\eta^2 = .034$. Although this effect was statistically significant, it was associated with a small effect size,

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

TABLE 3. Correlation matrix with predictors and outcome measures

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. PA	_											
2. Letter Span	.35** (.000)	_										
3. ALTM	.15 (.079)	.24** (.005)	_									
4. SRT	.03 (.725)	-.10 (.256)	-.04 (.633)	_								
5. LLAMA_B	.46** (.000)	.23** (.008)	.18* (.032)	.14 (.095)	_							
6. LLAMA_D	.12 (.167)	.19* (.025)	.23** (.007)	.10 (.240)	.23** (.009)	_						
7. LLAMA_E	.34** (.000)	.28** (.001)	.14 (.095)	.04 (.633)	.32** (.000)	.19* (.031)	_					
8. LLAMA_F	.35** (.000)	.23** (.008)	.18* (.032)	.14 (.097)	.46** (.000)	.27** (.002)	.26** (.002)	_				
9. Syntactic Complexity	.10 (.243)	.17* (.044)	.05 (.601)	-.02 (.863)	.15 (.083)	.16 (.057)	.08 (.355)	-.05 (.549)	_			
10. Lexical Complexity	.01 (.906)	.17* (.046)	.09 (.318)	-.07 (.472)	.04 (.670)	.17* (.047)	-.01 (.875)	-.10 (.270)	.49** (.000)	_		
11. Accuracy	.03 (.774)	-.16 (.068)	.01 (.933)	.12 (.170)	-.01 (.932)	-.19* (.032)	-.15 (.094)	-.01 (.984)	-.26** (.002)	-.42** (.000)	_	
12. Fluency	-.05 (.562)	.18* (.038)	.21* (.016)	-.06 (.468)	.02 (.794)	.02 (.863)	-.04 (.679)	-.18* (.042)	.61** (.000)	.65** (.000)	-.41** (.000)	_

*Correlation is significant at the 0.05 level (two-tailed).
**Correlation is significant at the 0.01 level (two-tailed).

TABLE 4. Loadings from exploratory factor analysis (PCA)

Component			
	1	2	3
LLAMA_B	**.75**	.28	.23
LLAMA_D	.12	**.77**	.22
LLAMA_E	**.64**	.20	-.07
LLAMA_F	**.65**	.38	.28
Paired Associates	**.79**	.10	-.10
Letter Span	**.55**	.37	**-.45**
ALTM Synonym	.10	**.76**	-.23
Serial Reaction Time	.08	.06	**.85**

Note: Loadings greater than .4 appear in bold.

accounting for 3% of the total variance, as also evidenced by its standardized beta coefficient of .19, indicating a weak positive relationship. Specifically, for every increase of one standard deviation in Implicit Memory Ability, fluency was predicted to increase by .19 standard deviations, holding all other variables constant.

The general linear model for syntactic complexity accounted for 6% of the total variance ($R^2 = .06$). This model did not yield any statistically significant results. Factors 1, 2, and 3 were not significant: $F(1, 121) = 1.926$, $p = .168$, $\eta^2 = .015$; $F(1, 121) = 1.585$, $p = .210$, $\eta^2 = .012$; and $F(1, 121) = .055$, $p = .816$, $\eta^2 = .000$, respectively. The interactions among predictors were not statistically significant either: Factors 1 and 2, $F(1, 121) = .143$, $p = .706$, $\eta^2 = .001$; Factors 1 and 3, $F(1, 121) = .197$, $p = .658$, $\eta^2 = .002$; Factors 2 and 3, $F(1, 121) = .608$, $p = .437$, $\eta^2 = .005$; and Factors 1, 2, and 3, $F(1, 121) = 1.218$, $p = .272$, $\eta^2 = .010$.

Finally, the general linear model for lexical complexity accounted for 11% of the total variance ($R^2 = .11$). In this model, Factors 1, 2, and 3 were not statistically significant: $F(1, 121) = .000$, $p = .986$, $\eta^2 = .000$; $F(1, 121) = 3.362$, $p = .069$, $\eta^2 = .027$; and $F(1, 121) = .299$, $p = .585$, $\eta^2 = .002$, respectively. The interactions between Factors 1 and 2, Factors 1 and 3, and Factors 1, 2, and 3 were not significant either: $F(1, 121) = .555$, $p = .458$, $\eta^2 = .005$; $F(1, 121) = 1.882$, $p = .173$, $\eta^2 = .015$; and $F(1, 121) = .015$, $p = .901$, $\eta^2 = .000$, respectively. However, the interaction between Factors 2 and 3, the factors interpreted as Implicit Memory and Implicit Learning Ability, reached statistical significance: $F(1, 121) = 4.374$, $p = .039$, $\eta^2 = .035$. This interaction indicated that the effect of the Implicit Memory Ability factor on lexical complexity depended upon the value of the Implicit Learning Ability factor. The size of this interaction was small, explaining 4% of the total variance. Its standardized beta of .20 showed a weak positive relationship with lexical complexity. Specifically, the standardized beta value indicated that the contribution of the interaction accounted for shift in lexical complexity of .20 standard deviations.

To explore the nature of this interaction more in detail, separate models were run predicting lexical complexity with Factor 2 (Implicit Memory) at different levels of Factor 3 (Implicit Learning) to assess whether the regression slopes for each of these

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

individual Factor 3 groups are statistically significant. Cohen et al.'s (2003) recommendation to use values corresponding to +1 standard deviation, mean, and -1 standard deviation was followed. Within the group with low- or mid-level Implicit Learning Ability (i.e., below 1 standard deviation or between -1 standard deviation and +1 standard deviation), Implicit Memory Ability was not significantly predictive of lexical complexity: $B = .027$, $t(96) = .258$, $p = .797$ and $B = .051$, $t(104) = .508$, $p = .612$. However, Implicit Memory Ability was significantly predictive of lexical complexity within the group with high Implicit Learning Ability (i.e., above 1 standard deviation), $B = .657$, $t(21) = 3.699$, $p = .001$. These results indicated that Implicit Memory Ability mattered for the development of lexical complexity for L2 learners with high Implicit Learning Ability.

To summarize, the data analysis showed that the eight cognitive tests examined, four from the LLAMA and four from the Hi-LAB batteries, were measuring three underlying latent constructs and that one of the factors, Factor 2, was a significant predictor of L2 speaking proficiency, particularly of L2 fluency as measured by pruned speech rate per minute, a measure of speed fluency (Tavakoli & Skehan, 2005) indicating the extent to which a speaker has a good balance of speed and the ability to keep talking. In addition, Factor 2 interacted with Factor 3 as a predictor of L2 lexical complexity operationalized as lexical diversity.

DISCUSSION

The current study had two objectives. First, to explore possible relationships between the Hi-LAB and LLAMA aptitude tests by examining whether the two batteries had any underlying constructs in common, and, second, to investigate whether any of these constructs were good predictors of L2 speaking proficiency in the context of 135 college students learning Spanish as an L2.

RESEARCH QUESTION 1: COMPARABILITY OF APTITUDE TESTS

A PCA run on eight aptitude tests, the four LLAMA tests and four Hi-LAB tests, identified three underlying constructs that accounted for 59.62% of the total variance in the data. The first factor, which accounted for 28.02% of the total variance, received significant loadings from five of the eight tests, three LLAMA tests (LLAMA_B, LLAMA_E, and LLAMA_F) and two Hi-LAB tests (Paired Associates and Letter Span). These tests have in common the fact they are measuring explicit, attention-driven cognitive abilities involving conscious and reflective learning processes. LLAMA_B and Paired Associates are both adaptations of Carroll and Sapon's (1959) paired associates measure and require learning associations between a nonword and either a picture or another word. The central executive resources engaged in such a task, which are necessary for guiding attention, strategy selection, planning, decision making, and control and coordination of the processes involved in short-term storage (Van der Linden et al., 1999), should also be engaged in the other three tests that loaded on Factor 1. The Letter Span, measuring phonological short-term memory, requires participants to recall lists of letters, and, therefore, relies on the central executive to temporarily store information, while needed for reasoning and processing, and to call up such information

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

(Kane & Engle, 2002). LLAMA_E and LLAMA_F both involve working out relations in datasets and explicitly instruct test takers to "work out the relationship" and "work out the grammatical rules," respectively. They have in common with paired associates and short-term memory tasks the fact that the rehearsing of testing materials is possible, as well as opportunities for strategy use and problem-solving techniques. Explicit associative learning (i.e., the learning of associations acquired consciously and intentionally) is likely to play a role in all of them. Based on the fact that the five tests loading on the first component are measuring cognitive abilities that involve explicit cognitive and memory processes, Factor 1 was labeled "Explicit Aptitude."

The second and third factors, which accounted for 17.05% and 14.55% of additional variance, respectively, received loadings from tests that have been argued to tap implicit cognitive processes (e.g., Granena, 2013a, 2013b; Kaufman et al., 2010; Woltz, 2003). ALTM Synonym and LLAMA_D loaded on the second factor, while Serial Reaction Time loaded on the third factor. The ALTM Synonym Task includes a semantic priming component that measures implicit (or procedural or nondeclarative) memory, a type of memory that is not available to consciousness, but that is responsible for the improvement or change in the identification, production, or classification of a stimulus as a result of a prior encounter with the same or a related stimulus (Tulving & Schacter, 1990). LLAMA_D, however, is a recognition task where test takers have to discriminate between old and novel items. Recognition tasks are supposed to involve episodic (explicit) memory retrieval. However, recognition can be based on the assessment of stimulus familiarity (Yonelinas, 2002). Recognition in LLAMA_D could be familiarity based, that is, recognition based on the sensation that an item has been seen before without explicitly remembering any detailed contextual information. Familiarity-based recognition memory judgments have been shown to rely on the same process that supports conceptual implicit memory (e.g., Wang & Yonelinas, 2012). Given that both LLAMA_D and the ALTM Synonym Task could be measuring implicit memory (i.e., unconscious retention), Factor 2 was labeled "Implicit Memory Ability."[7]

It should be noted that the reliability of one of the two measures that loaded onto Factor 2, LLAMA_D, was .50. This reliability is considered "poor" according to George and Mallery (2003) and it would be "unacceptable" below .50, according to the same standards. However, Nunnally (1967) maintained that even reliabilities of .60 or .50 might be acceptable in theoretical studies. Hair et al. (2006) argued that reliability may decrease to .60 and still be acceptable in exploratory studies and research in the social sciences, while Aron and Aron (1999) proposed that values of .60 or even lower could be adequate in psychology research. Although there is no question that values exceeding .7 should be preferred (Plonsky & Derrick, 2016, proposed .74 as a threshold for a minimally acceptable estimate), there is evidence suggesting that the reliability of implicit learning tasks is typically lower than the reliability of explicit learning tasks because they place fewer constraints on performance processes, which results in noisier data (e.g., Dienes, 1992; Kaufman et al., 2010; Reber, Walkenfeld, & Hernstadt, 1991; Robinson, 1997). For example, Reber et al. (1991) and Robinson's (1997) replication study reported reliabilities of .51 and .52, respectively. That notwithstanding, the low reliability of LLAMA_D in terms of its internal consistency means that measurement error is present and may have affected the results of the study, undermining the claims made. Therefore, caution should be exercised when generalizing the results obtained in relation to LLAMA_D.

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

The only task that loaded positively and significantly on the third component, the Serial Reaction Time Task, has been proposed as a measure of implicit inductive learning ability (Nissen & Bullemer, 1987), that is, of an individual's ability to learn a pattern or rule through simple exposure and without the intent to learn the pattern. Given the fact that Serial Reaction Time was the test with the strongest loading on the third factor (.85) and that the only other significant loading was a negative loading from the Letter Span Test (-.45), Factor 3 was labeled "Implicit Learning Ability." A possible explanation for the negative loading of the Letter Span Test could be the different nature of the cognitive processes that characterize short-term memory and implicit learning. As a component of working memory, short-term memory engages attention-driven memory processes. While selective attention is required for implicit learning, intention to learn or executive processing resources that guide the focus of attention are not. The inverse relationship between short-term memory and the hypothesized "Implicit Learning Ability" factor could be interpreted as a competition of implicit and explicit cognitive processes in implicit learning such that restricting the contribution of the explicit processes involved in short-term memory results in an enhancement of implicit learning. The claim that implicit and explicit systems operate in a push-pull manner is defended by dual-process theories of cognition (e.g., Jacoby, 1991). However, it should be noted that the same relationship was not found in the case of Factor 1, "Explicit Aptitude." In this case, there was no association, neither positive nor negative, between Serial Reaction Time task performance and the factor (.08), perhaps suggesting that explicit cognitive processes do not require restricting the contribution of implicit processes to operate. In terms of cognitive aptitudes, this result could suggest that individuals with better implicit learning ability are those with lower executive functions, but further research is necessary to confirm this claim.

These results suggest that the tests in the LLAMA and the Hi-LAB may have some underlying constructs in common, mainly, one measuring a combination of explicit learning and memory abilities and one measuring abilities in the implicit domain, specifically implicit memory. The constructs in the two batteries, however, do not completely overlap. According to the results, the Hi-LAB is measuring implicit learning in addition to implicit memory, while the LLAMA is not. This shows differences in the conceptualization of aptitude underlying the two batteries, and it is consistent with the broader scope of the Hi-LAB, a battery of 11 tests developed to predict high-level proficiency and that measures abilities such as executive functioning and processing speed, which the LLAMA, largely based on the MLAT, was not designed to tap. The fact that both batteries are hypothesized to predict L2 learning is not in conflict with the lack of a full overlap between their underlying constructs. It may indeed affect their predictive power in particular learning contexts or for particular aspects of L2 learning, but this is precisely one of the challenges that aptitude research has ahead, understanding the role of subcomponents of aptitude for particular learning outcomes.

The exploratory factor analysis performed in this study further showed that cognitive abilities in the implicit domain are not unitary and that, at least, a distinction between implicit memory and implicit learning is necessary. Although there is no consensus in the cognitive psychology literature, one position in memory research is that implicit memory and implicit learning do not overlap when implicit memory is used to refer to a particular nondeclarative memory phenomenon, more specifically to the effects that prior

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

experiences have on participants' current performance (e.g., Buchner & Wippich, 1998; Reber, 2013; Seger, 1994). Implicit learning, however, refers to the incidental acquisition of knowledge about the structural properties of the relations between objects or events (Buchner & Wippich, 1998). Also, whereas implicit memory studies tend to be interested in the retrieval stage of memory, implicit learning studies are primarily concerned with the encoding stage (i.e., the acquisition of knowledge). This allows for the possibility that implicit memory and implicit learning are two different aptitude constructs resulting in nonoverlapping individual variation.

RESEARCH QUESTION 2: APTITUDE AND L2 SPEAKING PROFICIENCY

Regarding the second research question, which asked whether any of the constructs identified would be good predictors L2 speaking proficiency, the results showed that the second component, the factor labeled "Implicit Memory Ability," with significant loadings from ALTM Synonym and LLAMA_D, was able to predict L2 speaking proficiency in terms of a subconstruct of fluency measured as pruned speech rate per minute. Implicit memory is a major form of long-term memory that does not require conscious thought. While implicit learning mostly concerns the encoding of input (i.e., the acquisition of new information), implicit memory mostly concerns the retention and, especially, the retrieval of information. It involves effortless and incidental retrieval of previously learned information that was acquired either intentionally or incidentally (Reber, 2013). Some individuals are better at spreading activation in implicit memory and, as a result, at making use of, or retrieving, information more efficiently (Woltz, 2003). The fact that implicit memory ability explains differences in L2 speed fluency suggests that those learners with greater implicit memory ability are more able to recall words in a foreign language quickly and automatically. They may also be more able to build meaningful associations between words and to map meaning to form.[8]

The ALTM Synonym measure, however, was not able to discriminate between high L2 attainment and mixed L2 attainment in Linck et al. (2013). This could be due to the fact that Linck et al. (2013) examined predictability of reading and listening skills, whereas the present study focused on speaking performance. Therefore, it could be that Implicit Memory Ability is more relevant for productive than receptive skills because production requires actual language retrieval. Learners with greater Implicit Memory Ability would be better at retrieving previously learned or known information effortlessly and incidentally.

The fact that the speaking task employed did not focus on language forms and, instead, involved spontaneous, real-time performance could further explain that explicit cognitive aptitudes did not predict L2 outcomes. However, this does not explain the lack of a relationship between the Implicit Learning Ability factor and L2 proficiency. Linck et al. (2013) found that Serial Reaction Time was a significant predictor of L2 proficiency in their study and argued that implicit learning played an important role in achieving high-level L2 attainment. One possibility for this discrepancy is that this study included intermediate-level L2 learners, or the equivalent of a 200-level language course, whereas the participants in Linck et al. (2013) were, at least, at an ILR level 4 on the Defense Language Proficiency Tests (DLPT; Defense Language Institute Foreign Language Center, 2009), that is, between an advanced professional and a functionally native

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

proficiency level. Implicit Learning Ability may be a significant predictor of L2 proficiency at advanced levels, but not at intermediate language levels.[9] Similarly, more explicit cognitive abilities such as analytical ability could be relevant at initial stages, as hypothesized by Linck et al. (2013), and as amply shown by studies from the *MLAT Manual* carried out to establish the predictive validity and percentile norms of the MLAT battery (Carroll & Sapon, 2002).

The results of this study also revealed a significant interaction between two of the factors (Implicit Memory and Implicit Learning Ability) as predictors of lexical complexity. Lexical complexity has been defined by Bulté and Housen (2012) as "the degree of elaboration, the size, the breadth, width or richness of the L2 learner's [lexical knowledge]" (p. 25). In this study, it was operationalized as lexical variety and used as a measure of the range of an L2 learner's productive vocabulary. The significant interaction found suggests that those learners with a broader productive vocabulary were those with higher implicit memory and implicit learning abilities. Specifically, the nature of the interaction indicated that the effects of implicit memory ability on this aspect of vocabulary acquisition do not begin to show up unless implicit learning ability is high. This suggests that as L2 learners become more able to acquire more vocabulary incidentally, their implicit memory starts playing a more important role allowing them to access and retrieve words more easily.

Although this study found statistically significant results, the effect sizes obtained by means of eta squared (η^2) were small, indicating that the effects observed could only explain between 3% and 4% of the total variance. Language aptitude has generally yielded correlation coefficients in the .40 to .60 (Carroll, 1981, p. 93). This means that language aptitude has been able to account for approximately 16% to 36% of the variance as an individual difference. For specific aspects of L2 learning such as L2 grammar learning, vocabulary, or L2 writing, the effect sizes have been shown to be smaller (e.g., .32 or 10% and .34 or 11%, respectively; Li, 2016). It is likely that other factors not considered in this study can account for greater variability in L2 proficiency. Future research is necessary, particularly research investigating cognitive aptitudes proposed by more recent conceptualizations of aptitude.

CONCLUSION

This study, part of a broader project on the role of cognitive aptitudes in an instructed setting, investigated the underlying structure of the two most recent aptitude test batteries in the SLA field (the LLAMA and the Hi-LAB) to see whether they had any constructs in common. The study also examined whether any of the observed underlying constructs could predict L2 speaking proficiency in terms of CAF. Results showed that the LLAMA and the Hi-LAB include tests measuring the same constructs. Specifically, the tests from the two batteries loaded onto three different factors, which were interpreted as "Explicit Aptitude," "Implicit Memory Ability," and "Implicit Learning Ability." The results further showed that Implicit Memory Ability was a significant predictor of L2 fluency and that it interacted with Implicit Learning Ability as a predictor of lexical complexity. This suggested that L2 learners with greater Implicit Memory Ability might be better at retrieving previously learned or known information effortlessly and incidentally. Implicit Memory Ability was further associated with having richer, more varied vocabulary

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

provided that Implicit Learning Ability was also high, possibly allowing learners to acquire a wider range of words available for retrieval and use in oral production.

LIMITATIONS AND FURTHER RESEARCH

This study had several limitations. First, the study did not include multiple CAF measures that could allow for more fine-grained analyses in terms of CAF subconstructs. They did not target specific language aspects or target structures either. Further research should look at the predictive validity of the factors identified in carefully controlled experimental instructed SLA studies, especially given the small effect sizes found in the current study. As an anonymous reviewer rightly pointed out, there is also a need to examine areas of overlap between different aptitude tests, not only Hi-Lab and LLAMA, as a way of further refining the aptitude construct in a bottom-up fashion, as measured by currently available instruments. Also, implicit memory is typically measured by means of priming tasks. However, it can take different forms and some researchers in SLA (e.g., Brill-Schuetz & Morgan-Short, 2014; Faretta-Stutenberg & Morgan-Short, 2018; Morgan-Short et al., 2014) have focused on individual differences in procedural memory, a subset of implicit memory in charge of sensory-motor habits or automatic skills. Future research should investigate the extent to which individual differences in implicit memory subtypes overlap and whether they are equally good predictors of L2 outcomes. Finally, the L2 learners in this study were intermediate-level learners and the measures that predicted proficiency differed from those that predicted advanced-level proficiency in Linck et al. (2013). Future studies looking at aptitude as a predictor of general L2 proficiency should consider including different proficiency groups (beginners, intermediate, and advanced) in the same study to gain a better understanding of the role of aptitude in L2 learning.

NOTES

[1]An anonymous reviewer pointed out that this begs the question whether any patterns can be acquired without conscious awareness (as in noticing). Indeed, this issue remains a matter of debate even in cognitive psychology, with some authors concluding that awareness systematically accompanies human learning (Shanks & St. John, 1994) and others that unconscious learning is a fundamental process in human cognition (Reber, 1993). Without denying that implicit and explicit processes typically function together in complex learning tasks, lack of awareness is understood in this paper as defined by Jiménez (2002): 1) lack of conscious intention to learn, 2) lack of conscious awareness of the fact that we are learning, and 3) lack of conscious attribution of any noticed changes to the effects of learning or to the original learning event.

[2]The same reviewer added that the LLAMA tests are not to be used in high-stake situations such as the ones for which other batteries (e.g., MLAT, Hi-LAB) were intended. It is the reviewer's impression that the LLAMA tests are being used somewhat uncritically and reminds SLA researchers of the need to acknowledge the limitations of the battery. Finally, the reviewer mentioned that the current Windows version of the LLAMA tests will be soon decommissioned in favor of a Web-based system.

[3]An anonymous reviewer expressed his concern about the minimum score values for LLAMA_D, LLAMA_E, and LLAMA_F. Specifically, there were 9 participants that scored zero on LLAMA_D, 6 participants that scored zero on LLAMA_E, and 10 participants that scored zero on LLAMA_F. From my own experience with the LLAMA tests, it is not unusual to find participants scoring zero in large samples such as the one used for this study (>100). It may be, however, a reason for concern, as expressed by the reviewer, and it may indicate the need for future studies to ensure that participants understood test instructions correctly when moving from the study phase to the testing phase of the LLAMA tests. The participants that scored zero in this study were not the same for the three tests, suggesting that their performance was not consistently low due to

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

uncontrolled factors such as distraction. Given that the distribution of the scores was normal for all the LLAMA tests according to Kolmogorov-Smirnov (*K-S*) tests ($p > .05$) the zero scores were kept in the analysis.

[4]Factor loadings (weights) with absolute values greater than .3 or .4 are usually judged to be substantial or "significant" (Hair, Black, Babin, & Anderson, 2006).

[5]An anonymous reviewer suggested removing LLAMA_D from the analysis on the basis of its low reliability. Removing LLAMA_D from the PCA did not change the underlying factor structure. The analysis identified three factors explaining 64.1% of the total variance. Factor 1 accounted for 34.94% of the variance and received significant loadings from the Letter Span (.54), Paired Associates (.78), LLAMA_B (.71), LLAMA_E (.67), and LLAMA_F (.61) tests. Factor 2 accounted for 16.56% of the variance and received significant findings from the Serial Reaction Time (.82) and the Letter Span (−.41) tests. Finally, Factor 3 accounted for 12.56% of the variance and received significant loadings from the ALTM synonym task (.96).

[6]Effect size was calculated as eta squared (η^2), indicating the percentage of variance explained in the dependent variable. Eta squared was derived by dividing the sum of squares for the effect of interest by the total sum of squares. The magnitude of the effect was interpreted according to the standards of Cohen (1988, 1992): small effect as .01 (r = .10), medium effect as .09 (r = .30), and large effect as .25 (r = .50).

[7]An anonymous reviewer disagreed with the possibility that LLAMA_D and the ALTM synonym task could be measures of the same construct and with the explanation provided. The reviewer suggested labeling the factor "Auditory Short-Term Memory," given that both tests involve auditory input and that participants need to listen carefully to the stimuli provided. I have attempted to provide a tentative explanation based on the available literature on priming, including Linck et al.'s (2013) description of the ALTM synonym task, and my own interpretation of the construct, which, of course, will have to be further investigated.

[8]Further evidence for the role of implicit memory in language fluency can be found in the literature on language disorders, such as non-fluent and fluent aphasia. These are impairments that have been related to damage to the brain structures underlying procedural memory (Ullman, 2003).

[9]An anonymous reviewer clarified that, although ILR level 4 was sufficient for membership in the high proficiency group in Linck et al. (2013), this was not the only criterion used, because very few individuals actually take the upper-level DLPT that is required to earn a score above ILR level 3. As a result, the study also included individual

REFERENCES

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

Aron, A., & Aron, E. (1999). *Statistics for psychology* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (Version 5.3.76). [Computer software]. Retrieved from http://www.praat.org/.

Brill-Schuetz, K. A., & Morgan-Short, K. (2014). The role of procedural memory in adult second language acquisition. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 260–265). Quebec City, QC: Cognitive Science Society.

Buchner, A., & Wippich, W. (1998). Differences and commonalties between implicit learning and implicit memory. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 3–46). Thousand Oaks, CA: Sage Publications.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam, The Netherlands, and Philadelphia, PA: Benjamins.

Carroll, J. B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87–136). Pittsburgh, PA: University of Pittsburgh Press.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*, 723–733.

Carroll, J. B. (1969). What does the Pennsylvania foreign language research project tell us? *Foreign Language Annals*, *3*, 214–236.

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

Carroll, J. B., & Maxwell, S. E. (1979). Individual differences in cognitive abilities. *Annual Review of Psychology*, *30*, 603–640.

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

Carroll, J. B., & Sapon, S. (1959). *Modern language aptitude test: Form A*. New York, NY: Psychological Corporation.

Carroll, J. B., & Sapon, S. (2002). *Manual for the MLAT*. Bethesda, MD: Second Language Testing.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Defense Language Institute Foreign Language Center. (2009). Defense language proficiency testing system 5 framework. Retrieved from http://www.dliflc.edu/file.ashx?path=archive/documents/Framework_Document_Sep_10_09.pdf.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, *16*, 41–79.

Doughty, C., Campbell, S., Bunting, M., Mislevy, M., Bowles, A., & Koeth, J. (2010). Predicting near-native L2 ability: The factor structure and reliability of Hi-LAB. In M. T. Prior, Y. Watanabe, & S.-K. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum: Exploring SLA perspectives, positions, and practices* (pp. 10–31). Somerville, MA: Cascadilla Proceedings Project.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *Modern Language Journal*, *79*, 67–89.

Faretta-Stutenberg, M., & Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second Language Research*, *34*, 67–101.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299–323.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). Ann Arbor, MI: University of Michigan Press.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Boston, MA: Allyn & Bacon.

Granena, G. (2013a). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–129). Amsterdam, The Netherlands: John Benjamins.

Granena, G. (2013b). Individual differences in sequence learning ability and SLA in early childhood and adulthood. *Language Learning*, *63*, 665–703.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, aptitude and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*, 311–343.

Grigorenko, E., Sternberg, R., & Ehrman, M. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal*, *84*, 390–405.

Guiraud, P. (1954). *Les charactères statistiques du vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson, Prentice Hall.

Ishikawa, T. (2008). The effects of task demands of intentional reasoning on L2 speech performance. *The Journal of Asia TEFL*, *5*, 29–63.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.

Jiménez, L. (2002). Attention in probabilistic sequence learning. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 43–67). Amsterdam, The Netherlands: John Benjamins.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working memory capacity, executive attention, and general fluid intelligence: An individual difference perspective. *Psychonomic Bulletin and Review*, *9*, 637–667.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jimenez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–340.

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

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

Li, S. (2016). The construct validity of language aptitude. A meta-analysis. *Studies in Second Language Acquisition*, *38*, 801–842.

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

Long, M. H., & Doughty, C. (2009). *The handbook of language teaching*. Malden, MA: Wiley-Blackwell.

Markessinis, J. (1968). *Summary of Dr. John B. Carroll's "The foreign language attainments of language majors in the senior year"*.Washington, DC: Peace Corps.

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

Michel, M., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics*, *45*, 241–259.

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*, 56–72.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Petersen, C., & Al-Haik, A. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement*, *36*, 369–380.

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (PLAB)*. New York, NY: Psychological Corporation.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, *100*, 538–553.

Psychology Software Tools. (2011). E-Prime™ [Computer software]. Pittsburgh, PA: Psychology Software Tools.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. London, UK: Oxford University Press.

Reber, A. S., Walkenfeld, F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 888–896.

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*, 2026–2042.

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

Robinson, P. (2002). Individual differences in intelligence, aptitude and working memory during adult incidental second language learning: A replication and extension of Reber, Walkenfeld, and Hernstadt (1991). In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211–266). Amsterdam, The Netherlands: Benjamins.

Rogers, V. E., Meara, P., Aspinall, R., Fallon, L., Goss, T., Keey, E., & Thomas, R. (2016). Testing aptitude. *EUROSLA Yearbook*, *16*, 179–210.

Rogers, V. E., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *JESLA*, *1*, 49–60.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163–196.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral & Brain Sciences*, *17*, 367–447.

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

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1–14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.

Sparks, R. L., Ganschow, L., & Patton, J. (1995). Prediction of performance in first year foreign language courses. *Journal of Educational Psychology*, *87*, 638–655.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). Amsterdam, The Netherlands: Benjamins.

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*, 301–306.
Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231–270.
Van der Linden, M., Collette, F., Salmon, E., Delfiore, G., Degueldre, C., Luxen, A., & Franck, G. (1999). The neural correlates of updating information in verbal working memory. *Memory*, *7*, 549–560.
Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 65–83.
Wang, W. C., & Yonelinas, A. P. (2012). Familiarity is related to conceptual implicit memory: An examination of individual differences. *Psychonomic Bulletin & Review*, *19*, 1154–1164.
Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language*, *56*, 86–102.
Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1047–1060.
Woltz, D. J. (2003). Implicit cognitive processes as aptitudes for learning. *Educational Psychologist*, *38*, 95–104.
Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

https://doi.org/10.1017/S0272263118000256 Published online by Cambridge University Press