JESLA Journal of the European Second Language Association

Rogers, V. et al. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association,* 1(1), 49–60, DOI: https://doi.org/10.22599/jesla.24

RESEARCH

Examining the LLAMA aptitude tests

Vivienne Rogers, Paul Meara, Thomas Barnett-Legh, Clare Curry and Emma Davie

This study assesses the reliability[1] of the LLAMA aptitude tests (Meara, 2005). The LLAMA tests were designed as shorter, free, language-neutral tests loosely based on the MLAT tests (Carroll & Sapon, 1959). They contain four sub-components: vocabulary acquisition, sound recognition, sound-symbol correspondence and grammatical inferencing. Granena (2013) and Rogers et al. (2016) provided initial results regarding factors which might influence LLAMA test scores. This paper develops this previous work by examining some of issues raised with a larger cohort and focuses on the following research questions.

1. Are the LLAMA tests language neutral?
2. What is the effect of bilingualism on LLAMA test scores?
3. What is the effect of age on LLAMA test scores?
4. How much variance can background factors account for in the LLAMA test results?

Data were collected from 240 participants aged 10–75 for RQ1–3. We found no significant differences in terms of language background (RQ1) but instructed second language learners significantly outperformed monolinguals (RQ2). For RQ3 we found that the younger groups were outperformed by all the other groups.

For RQ4, we investigated how much variance in LLAMA test results six individual background factors could explain. We combined data from Rogers et al. (2016) and this study giving 404 participants in total. Using a multiple regression analysis, we found that prior L2 instruction predicted more of the variance (6%) than any other factor. We suggest that when using the LLAMA tests, researchers should consider controlling for language learning experience.

This study scrutinises the components of the LLAMA tests with a large set of data. We conclude that the results are robust across a range of individual differences but suggest that different norms may be needed for younger age groups and those who have received prior L2 instruction.

Keywords: Aptitude; Second language learning; Testing

1. Introduction

Language-learning aptitude has seen a resurgence of interest in recent years with second language researchers increasingly turning towards aptitude as a factor in explaining individual differences (Wen, Biedroń & Skehan, 2017). Dörnyei and Skehan (2003) suggest a general working definition for aptitude: "there is a specific talent for learning foreign languages which exhibits considerable variation between individual learners" (Dörnyei & Skehan, 2003, p. 590). However, beyond this definition, there is considerable variation among researchers about what components make up language-learning aptitude although they share many common elements. This has given rise to a number of different aptitude tests (e.g. MLAT (Carroll & Sapon, 1959); Pimsleur Aptitude Battery (Pimsleur, 1966); DLAB (Petersen & Al-Haik, 1976); CANAL-FT (Grigornko, Sternberg & Ehrman, 2000); LLAMA (Meara, 2005); HiLAB (Linck et al., 2013)). These tests all have slightly different emphases in what they (claim to) measure and many are not currently available to researchers (see Skehan, 2016, for a fuller discussion).

This study focuses on the free, easily available LLAMA tests given their increasing popularity (over 700 citations on Google scholar since 2013). Before explaining the LLAMA tests in detail, we briefly outline some of the areas investigated in terms of language-learning aptitude. As the tests have been used to address some of these questions, we then report on whether the tests are influenced by some of these factors themselves (e.g. age, bilingualism). We conclude by suggesting that norms are needed for instructed second language learners and also for different age groups.

Swansea University, UK

Corresponding author: Vivienne Rogers (v.e.rogers@swansea.ac.uk)

2. Background on Language Aptitude

Language-learning aptitude research in SLA has generally been founded on the early work by Carroll and Sapon (1959). This approach to language-learning aptitude can be summed up by the following quote by Carroll (1990, p. 26):

> The amount of time a student needs to learn a given task, unit of instruction, or curriculum to an acceptable criterion of mastery under optimal conditions of instruction and student motivation.

For Carroll, aptitude was a relatively stable, unchanging characteristic comprising four sub-components: phonemic coding ability, grammatical sensitivity, inductive language learning ability and associative memory (Carroll, 1973).[2] This approach is epitomised in the Modern Languages Aptitude Test (MLAT) (Carroll & Sapon, 1959). This concept of aptitude has been subject to criticism in terms of how memory is conceptualised (Wen, 2016), the role of implicit learning (De Graaff, 1997), the links with intelligence (Sasaki, 1999) and its stability over time (Kormos, 2013; Ganschow, Fluharty & Little, 1995). For a more in-depth look at the history of language learning aptitude research see Skehan (2016).

Li (2016) carried out a meta-analysis of 66 studies examining the construct validity of language learning aptitude. He concluded that aptitude was independent of other individual differences like motivation (contra Pimsleur, 1966) and classroom anxiety (indeed Sparks and Patton (2013) suggested that low aptitude may cause classroom anxiety). Li concluded that aptitude was a strong predictor of general proficiency but not of vocabulary learning or L2 writing yet different test sub-components predicted different aspects of learning. This strongly supports a multi-component approach to aptitude.

In terms of memory, Li found that studies showed that executive working memory was more strongly associated with aptitude than phonological short-term memory. However, Linck et al. (2013) argued that phonological short-term memory was of relevance to advanced learners, suggesting that different aspects of memory or aptitude may be relevant at different ages (cf. Abrahamsson and Hyltenstam, 2008; Muñoz, 2014). In an attempt to bring together these different aspects of (working/short-term) memory, Wen (2016) has proposed the "Integrated Approach" in which phonological working memory is a "language learning device" and executive working memory is involved with "language processes" (Wen, 2016, p. 147). This links different types of working memory to different types of aptitude. This approach is along similar lines to Granena (2016), who argued that different types of aptitude are linked to different cognitive styles.

In addition to the general findings regarding aptitude that arise from Li's (2016) meta-analysis, the range of aptitude tests and the variety of assumptions they make about the concept or construct of aptitude is clearly evident. One of these tests is the LLAMA test battery developed by Meara (2005). This test has gained in popularity as it is free, quick to administer and easily available, yet it has not been standardised or validated. In the following sections we outline some of the history and details of the LLAMA tests before turning to some practical, empirical questions that might influence LLAMA test results.[3]

3. LLAMA Aptitude Tests

The LLAMA tests were initially developed as part of a research training program for MA students at Swansea University. They are loosely based on the components that appear in Carroll & Sapon's (1959) Modern Language Aptitude Test (MLAT) but the aim was to take advantage of developments in technology at the time to develop an easier, more appealing user interface.

The 2005 version of the LLAMA tests described in this paper consist of four sub-tests, conventionally referred to as LLAMA_B, LLAMA_D, LLAMA_E and LLAMA_F.

LLAMA_B is the vocabulary learning module of the LLAMA tests. It assesses the users' ability to attach unfamiliar names to unfamiliar objects. Carroll and Sapon's tests assess this ability by asking test-takers to remember a set of paired associates – in the English version of the MLAT, this involves English words paired with Kurdish words. One obvious disadvantage of this approach is that it assumes the English native speaker test-taker is unfamiliar with Kurdish but moreover, that multiple versions of the test are required for different first languages thus adding additional variables when comparing classes with multiple L1s. LLAMA_B solves this problem by presenting test-takers with a set of pictures that do not have obvious names, but can easily be described in any language. This approach breaks away from the paired-associate format, and it allows test-takers a lot of flexibility in the way they approach the vocabulary learning task. **Figure 1** shows an example of type of stimulus used for this task.

There are twenty of these figures in the LLAMA_B test, all displayed simultaneously on screen. Clicking on an object causes its name to be displayed. Test-takers have two

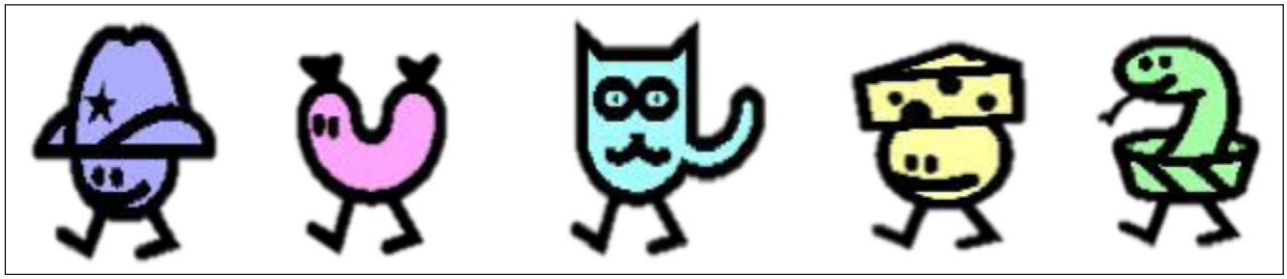

Figure 1: Five of the pictures used in LLAMA_B.

minutes to examine all 20 objects and learn their names. The program places no constraints on how they do this, so test-takers can adopt a number of different strategies to complete the task. At the end of the learning phase, LLAMA_B moves to a testing phase: the program displays the name of each of the 20 objects, and test-takers have to identify the object by clicking on it. Five points are scored for each object correctly identified, and there is no correction for guessing. This means that LLAMA_B scores range from 0–100, and the expected score for random guessing is 5.[4]

LLAMA_D is a new test that does not appear in MLAT. It is based on a suggestion that a core skill in language learning is the ability to recognise repeated sounds in spoken language: basically, a learner who is able to recognise repeated stretches of sound is more likely to notice small variations in speech, and this makes it easier for them to isolate the individual words and variants of these words that signal morphology. To this extent, it can be considered a measure of implicit learning. The test works in two phases. In Phase 1 the test-taker hears a series of short sound clips in an unfamiliar language. In Phase 2 the test-taker hears another set of sound clips. Some of these are new, but some are repeated from Phase 1. For each clip, the test-taker has to indicate whether they have heard it in Phase 1 before or not. Five points are scored for each correct answer, and test-takers are penalised for guessing. The entire test takes about five minutes. It generally gets positive comments from users but appears to be very hard in that very few test-takers score highly on it.[5]

LLAMA_E is an adaptation of MLAT's sound-symbol correspondence task. The test interface consists of a series of 24 labelled buttons in a Roman alphabet, but one that uses these familiar symbols in an unfamiliar way (see **Figure 2**). Clicking a button plays the syllable that is represented by the label. Test-takers have two minutes to explore this interface. The programme then moves to its test phase. In this phase, test-takers hear a complex two-syllable "word" and have to decide which of two spellings is correct. Five points are scored for each correct answer, and five points are deducted for an incorrect answer.

LLAMA_F is a grammar inferencing test. The presentation phase of the program shows the test-taker a series of pictures depicting shapes and objects, and a short sentence in an artificial language which describes each picture. An example is shown in **Figure 3**. The test-taker is expected to work out how the descriptions relate to the pictures. From this, they should be able to intuit some of the grammatical and morphological features of the language: word order, gender, singular, dual and plural numbers, conjugating prepositions, and so on. Test-takers have five minutes to explore this data set.[6] Then they are presented with a new set of pictures that incorporate new elements. Each picture is accompanied by two sentences which might describe it, and test-takers indicate which is the correct description. They should be able to do this if they have internalised the grammatical rules evidenced in the presentation phase. Five points are awarded for a correct answer and five points deducted for an incorrect choice.[7]

4. Methodology

4.1. Research questions and hypotheses

This study arose out of limitations from our previous study investigating the factors which might influence LLAMA test performance (Rogers et al., 2016). In that

Figure 2: The syllabary used in LLAMA_E.

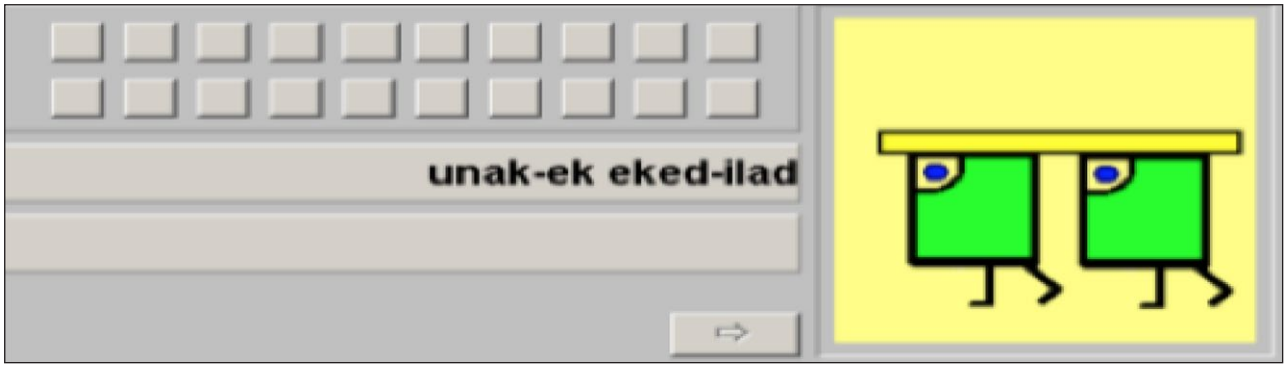

Figure 3: An example of the stimuli used in the presentation phase. The pictures are designed to highlight key grammatical features.

study, we tested 229 participants aged 10–75 from a range of typologically distinct L1s, with various education levels. We found no effect for L1 or gender, but we did find an effect of education level on three subcomponents of the LLAMA tests (B, E & F). The younger participants were outperformed by the adults on LLAMA_E (sound-symbol correspondence). These findings provided an important early step in validation of the LLAMA tests but raised a number of issues, particularly in terms of their suitability for use with younger participants and the role of education in test results. However, many of these sub-groups had low participant numbers thus limiting their generalisability. These limitations led us to address the following questions.

1. Are the LLAMA tests language neutral?
2. What is the effect of monolingualism, early bilingualism and instruction in a L2 on LLAMA test scores?
3. What is the effect of age on LLAMA test scores?
4. How much variance do key background factors (e.g. age, gender, L1, L2 status) account for in the LLAMA test results?

As previously noted, one of the reasons for developing the LLAMA tests was to enable it to be used with a range of L1s. However, two of the tests, LLAMA_B and LLAMA_F, use words written in a Roman script. LLAMA_E also has letters and numbers from a Roman script. This led us to the first research question. Several studies suggest the degree of distance between an L1 and an L2 plays a fundamental role in word processing and retention in an L2 (Gholamain & Geva, 1999; Green & Meara, 1987; Hamada & Koda, 2008). If the language script of the L1 can influence the acquisition of the L2, then the question arises if the L1 script of the learner influences their aptitude scores.

In Rogers et al. (2016), we looked at this question but had a small sample size (n = 14) and grouped Arabic and Chinese native speakers together as a non-Roman script group. This was less than satisfactory due to the differences between Chinese as morphosyllabic (Tolchinsky, Levin, Aram & McBride-Chang, 2012, p. 1598) or logographic (Crystal, 1987, p. 200) script and Arabic, which as a consonant alphabetic script shares a common Semitic ancestor with Roman scripts (Sampson, 1985, p. 77). This current study addresses this limitation by comparing these two groups to each other and to L1 English participants. This enables us to formulate two hypotheses. The first is that the L1 English group will outperform the other two groups in LLAMA_B and LLAMA_F, as having the words in a Roman script will increase the processing load for the other two groups (Tan et al., 2003). Our second hypothesis is that the Arabic group will outperform the Chinese group as Arabic is an alphabetic system. We may also see an effect for LLAMA_E with the Chinese group as it contains a combination of words and letters as Akamatsu (1999) found that manipulating the way words were presented (use of capital letters, etc.) affected ESL speakers of L1 logo-graphic languages more than other ESL learners.

Our second research question asks if having a second language or being bilingual would account for differences in LLAMA test performance and is motivated by previous research suggesting that aptitude can be trained (e.g. Grigorenko et al., 2000; McLaughlin, 1990; Sternberg & Grigorenko, 2002) or changed due to experience (e.g. Hyltenstam, 2016; Kormos, 2013; Safar & Kormos, 2008; Sawyer, 1992; Sparks et al., 1995; Thompson, 2013). Our previous study did not find any significant differences in a post-hoc analysis of reported language experience. However, this did not take into account the level of language proficiency. This study specifically targets bilinguals (two L1s before age five) and instructed L2 learners in comparison with monolinguals. Our first hypothesis is that following Sparks et al. (1995), the instructed L2 group will outperform the other groups on the explicit measures (LLAMA_B, LLAMA_E and LLAMA_F) as they will have developed strategies for learning vocabulary (LLAMA_B) and grammar or pattern detection (LLAMA_E & LLAMA_F).[8] We do not expect a difference between the bilinguals and the monolinguals, as they will not have been instructed in any language-learning strategies. Our second hypothesis is that due to purported bilingual cognitive advantage effects (Bialystok, Luk & Kwan, 2005; Kaushanskaya & Marian, 2009), the bilingual participants will outperform the monolingual group due to their greater language awareness.

Our third research question arose as the question of aptitude and age of onset has been contested in the literature (e.g. Abrahamsson & Hyltenstam, 2008; Muñoz, 2014). Although the LLAMA tests were not originally designed for use with children, it seems appropriate to investigate the use of these tests with younger populations. While many researchers have used aptitude tests retrospectively, i.e. tested adults who started learning another language at a young age, there has also been a trend to test younger participants to determine how their aptitude can predict subsequent language results. This conflates the age of onset with the age of testing.[9] Here we focus on the issue of age at testing to establish if the LLAMA tests can be used with younger learners in the same way as with adolescents and adults. To address this third research question, we examined the test results from the vocabulary test (LLAMA_B) and the sound recognition measure (LLAMA_D) from three groups of learners. Group 1 comprises 10–11-year-olds, Group 2 comprises 20–21-year-olds and Group 3 comprises adults over the age of 30.[10] These age groups were chosen to examine a range of ages, including both younger and older participants who are long past any possible critical period and are cognitively mature. The decision to concentrate on these two tests was both principled and practical. For principled reasons, we wished to investigate areas in which these tests may be used to see differences in ages following cognitive development or critical/sensitive period hypotheses views (Bley-Vroman, 1989, 2009; Patkowski, 1980) and also in which we would not expect there to be age-related differences. Practically, we were constrained by the amount of time available with each child and so could not administer the whole LLAMA test battery.

We make two contrasting hypotheses for LLAMA_B (vocabulary learning). Our first hypothesis is that we would not see any differences between the groups as we continue to learn new words throughout our lives and so cognitive development or critical-period effects should not be evident. Our second hypothesis follows work by Miralpeix (2006, 2009) that older participants (over 11) would outperform the younger learners due to their increased cognitive advantages and maturity.[11] Our third hypothesis relates to LLAMA_D. As it is purported to be a measure of implicit learning (Granena, 2013) and if younger learners are claimed to make greater use of implicit learning[12] in comparison to adult learners (DeKeyser, 2000), then we would predict that the younger learners would outperform the older learners.

For the fourth research question, we have combined these results with Rogers et al. (2016) as the data were collected under similar conditions, with similar background questionnaires giving a total of 404 participants. This allows us to carry out a more powerful statistical analysis to consider the effects of various individual background variables on LLAMA test performance. These variables are age, gender, L1 script, L2 status, highest education qualification and whether or not the participant regularly plays logic puzzles. The rationale for the inclusion of these variables will be discussed in more detail in the results section.

4.2. Tasks and administration

The four sub-components were administered to all participants over the age of 18. They were administered on Windows computers, either on an individual basis or in larger computer classrooms. The latter were drop-in sessions advertised to the students at a UK university. Tests were scored automatically by the programs (as outlined above) and the results noted on a piece of paper.

Participants also took a background questionnaire. This was computer based using our university's Lime Survey software. Participants were given a URL and asked to give the same name that was on their LLAMA results sheet to allow for subsequent matching. Unfortunately, not all participants did so and their data were discarded. The questionnaire software automatically coded the results and it was imported into SPSS v. 20 for analysis.

Before participants took the LLAMA tests or completed the questionnaire, they were briefed on the nature of the research project,[13] given an information sheet and asked to complete a consent form. For participants under the age of 18, a parental consent form was given with the information sheet. These had to be returned before data collection from the children could be carried out. A simplified background questionnaire was given to the children under 18 in paper format. For the 10–11-year-olds, data-collection time was restricted to a maximum of 30 minutes, so as not to place an undue burden on the children.

4.3. Participants

Participants were recruited either through university-wide emails and posters or individually by some of the research team. No participants were paid for their time and, therefore, represent a generally opportunistic sample.

Data were collected from a total of 240 participants (128 female and 112 male). Participants' ages ranged from 10–75, but 148 were aged between 18–24, with a total of 211 over the age of 18. This was due to the majority of data collection taking place with undergraduate students as outlined above.

Of the participants over 18, we also had a range of educational backgrounds: 14 had left school at the end of compulsory education (aged 16), 112 had obtained qualifications at age 18 (A-level or equivalent), 70 already had an undergraduate degree and 13 had postgraduate qualifications. Again from the participants over 18, we had a range of prior language experience: 142 participants had learnt another language in school, 46 were monolingual English native speakers and 23 were bilingual speakers. Bilingualism was self-reported but defined as having acquired both languages before the age of five. In addition to English native speakers (n = 136), we had 56 L1 Chinese, 32 L1 Arabic and fewer than five each of German, Japanese, Welsh, Greek and Polish.

5. Results and Discussion

In this section, we present the results relating to each research question in turn before discussing them in terms of the hypotheses outlined previously.

5.1. Research question 1: language neutrality

The first question examined the role of L1 script in LLAMA tests results and whether they could be considered language neutral. To investigate this, we examined three groups: L1 Arabic speakers, L1 Chinese speakers and L1 English speakers as shown in **Table 1**. All participants over age 18 took all four tests, giving a total of 195 participants.

The results of a one-way between groups ANOVA show that there were significant differences between the groups for all of the LLAMA sub-components except LLAMA_D:

Table 1: Results of LLAMA tests according to L1 script.

		LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F
English	M	45.28	27.94	68.32	36.40
(n = 107)	s.d.	21.608	16.653	29.065	24.618
Chinese	M	55.89	31.16	56.34	46.96
(n = 56)	s.d.	27.288	24.458	28.034	25.984
Arabic	M	53.75	34.38	62.19	49.06
(n = 32)	s.d.	24.163	15.748	25.207	24.933

LLAMA_B F (2, 192) = 4.212, p = 0.016; LLAMA_E F(2, 192) = 3.389, p = 0.036 and LLAMA_F F(2, 192) = 5.000, p = 0.008 but not for LLAMA_D F(2, 192) = 1.563, p = 0.212.

Results from a post-hoc Games-Howell test (unequal variances) showed that for LLAMA_B, the L1 Chinese group (M = 55.89, s.d. = 27.288) significantly outperformed (p = 0.035) the L1 English group (M = 45.28, s.d. = 21.608). There were no significant differences between the L1 Arabic and either L1 Chinese group or L1 English group. For LLAMA_E, again there was a significant difference between the L1 Chinese and L1 English groups (p = 0.032) but this time the L1 English group (M = 68.32, s.d. = 29.065) outperformed the L1 Chinese group (M = 56.34, s.d. = 28.034). Again there were no significant differences between the L1 Arabic and either L1 Chinese group or L1 English group. For LLAMA_F, there were significant differences between the L1 English group (M = 36.40, s.d. = 24.618) and both the L1 Chinese group (M = 46.96, s.d. = 25.984) and the L1 Arabic group (M = 49.06, s.d. = 24.933). The L1 English group performed significantly worse than both the L1 Chinese (p = 0.036) and L1 Arabic (p = 0.038) groups. There was no significant difference between the L1 Chinese and L1 Arabic groups.

We were concerned that perhaps the three groups were not comparable as many of the L1 English group were monolinguals. **Table 2** shows the results of the participants over 18 who reported having studied another language. As **Table 2** shows, this reduces the L1 English group from n = 107 to n = 48. It also reduces the L1 Arabic group to 30, as two participants were bilingual with English and had not studied another language.

The results of a one-way between groups ANOVA for these L2 groups show that there were no significant differences on any of the LLAMA sub-components: LLAMA_B F (2, 131) = 0.263, p = 0.769; LLAMA_D F (2, 131) = 0.986, p = 0.376; LLAMA_E F (2, 131) = 3.021, p = 0.052; LLAMA_F F (2, 131) = 0.714, p = 0.492. While none of these results show overall significant differences, the results for LLAMA_E approach significance. This is due to differences between the L1 Chinese group (M = 56.34, s.d. = 28.034), who scored lower than the L1 English group (M = 69.90, s.d. = 29.867). This is in line with the findings of Akamatsu (1999) regarding the extra difficulties faced by speakers of logographic languages (like Chinese) when Roman alphabet text is manipulated.[14]

In terms of our hypotheses for this question, our first hypothesis predicted that the L1 English group (Roman script) would outperform the other groups. As shown in both **Tables 1** and **2**, this is not the case regardless of whether the role of language instruction experience is considered or not. Our second hypothesis suggested that L1 Arabic participants would outperform the L1 Chinese group as Arabic is a consonant alphabetic language. This hypothesis was also not supported by the data; there were no differences between the groups in **Table 2**. This suggests that the LLAMA tests are indeed language neutral as there were no differences between groups once other factors (e.g. L2 instruction) were controlled for. This result follows Granena (2013), who also found no difference between her 187 Chinese, English and Spanish subjects. If the LLAMA tests can be used across participants of different language backgrounds and language pairings, as these results suggest, then this opens up aptitude testing to a much wider audience. Most of the existing aptitude tests are designed for homogeneous groups and require multiple versions for different L1s (e.g. MLAT).

5.2. Research question 2: L2 status

The second research question asked if bilingualism, monolingualism or instructed second language learning would impact on LLAMA scores. We divided the participants into three groups based on their answers in our background questionnaire; monolinguals, bilinguals (prior to age five) and instructed L2 learners. We compared the results of participants over the age of 18 who had completed all four of the LLAMA tests' sub-components (n = 211). The results are given in **Table 3**.

Table 2: Results of LLAMA tests according to L1 script for L2 learners.

		LLAMA B	LLAMA D	LLAMA E	LLAMA F
English	M	52.40	28.33	69.90	42.19
(n = 48)	s.d.	20.499	15.890	29.867	27.789
Chinese	M	55.89	31.16	56.34	46.96
(n = 56)	s.d.	27.288	24.458	28.034	25.984
Arabic	M	54.17	34.83	62.33	49.00
(n = 30)	s.d.	24.917	15.838	24.835	25.643

Table 3: Results of LLAMA tests according to L2 status.

		LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F
L2-er	M	53.24	30.85	63.31	45.25
(n = 142)	s.d.	24.234	19.902	28.434	27.310
Monolingual	M	39.57	25.65	65.11	31.20
(n = 46)	s.d.	20.759	17.720	28.800	20.033
Bilingual	M	42.39	32.83	66.52	38.260
(n = 23)	s.d.	22.303	14.834	30.243	25.876

The results of a one-way between groups ANOVA for these L2 status groups show that there were significant differences on two of the LLAMA sub-components: LLAMA_B F (2, 208) = 7.032, p = 0.001 and LLAMA_F F (2, 208) = 5.366, p = 0.005 but not for LLAMA_D $F(2, 208) = 1.604$, $p = 0.204$ or LLAMA_E $F(2, 208) = 0.164$, $p = 0.849$. Post-hoc Games-Howell (unequal variances) tests showed that for LLAMA_B the L2-er group (M = 53.24, s.d. = 24.234) significantly outperformed (p = 0.001) the monolingual group (M = 31.20, s.d. = 20.759). There were no significant differences between the bilingual group (M = 42.39, s.d. = 22.303) and either of the other groups. For LLAMA_F the situation is the same as the L2 group (M = 45.25, s.d. = 27.310) significantly outperformed (p = 0.001) the monolingual group (M = 31.20, s.d. = 20.033). Again there were no significant differences between the bilingual group (M = 38.260, s.d. = 25.876) and either of the other two groups.

Earlier, following Sparks et al. (1995), we hypothesised that the instructed L2 group would outperform the other two groups on explicit measures LLAMA_B, LLAMA_E and LLAMA_F, as they would have developed strategies for learning vocabulary and grammar/pattern recognition. This hypothesis was partially confirmed. There were overall effects of group on both LLAMA_B and LLAMA_F, but significant differences were only found between the instructed L2 group and the monolinguals – not the bilinguals. The instructed L2 group did outperform the bilinguals in both LLAMA_B and LLAMA_F, but this did not reach significance.

Our second hypothesis suggested that if bilinguals have a cognitive advantage, then we would expect them to outperform the monolingual group. This hypothesis was not confirmed statistically; there were no significant differences between the bilinguals and the monolinguals. However, the bilinguals did perform better than the monolinguals.

Granena (2013) found that LLAMA_B, E and F all weighted on the same component and suggested that these measured more explicit aspects of language-learning aptitude. In this respect, it is perhaps not surprising the instructed L2 learners perform best on these measures, as learning vocabulary and grammar rules are core elements of much L2 classroom instruction. To this extent, the idea of a training effect in aptitude testing (Grigornko et al., 2000; Kormos, 2013) is perhaps not a surprise. However, whether this suggests that aptitude itself is trainable or whether it is test performance that is affected remains an open question and one that would be difficult to empirically address. Nayak, Hansen, Krueger and McLaughlin (1990) suggest that multilingual learners are more adept at using strategies in taking the tests rather than being more successful overall, and this may be the case with our participants as well.[15]

This question of training, however, does lead to certain methodological consequences. It appears that irrespective of whether you regard aptitude as stable or trainable, the LLAMA tests seem to be influenced by prior experience or training (instruction). This leads us to suggest the caveat that when using the LLAMA tests (particularly B and F), researchers should be aware of the language-learning background of their participants. By this we mean that in situations with a mix of participants with no prior L2 instruction experience (L2-ers) and those who have had instruction (L3-ers), then we would anticipate that the learners with prior instruction would outperform the others and therefore their results cannot be taken as a whole or compared to each other as a single measure, particularly in high stakes situations.

5.3. Research question 3: age

The third research question considered the effect of age on LLAMA scores. We used two of the LLAMA sub-tests (vocabulary and sound recognition) with three different age groups: Group 1 aged 10–11, Group 2 aged 20–21 and Group 3 aged 30–70. We also matched for gender. The results are given in **Table 4**.

The results of a one-way between groups ANOVA for these age groups show that there were significant differences on both of the LLAMA sub-components tested: LLAMA_B F (2, 101) = 6.741, p = 0.002 and LLAMA_D F (2, 101) = 3.919, p = 0.023. Post-hoc Games-Howell (unequal variances) tests showed that for LLAMA_B Group 1 (aged 10–11, M = 28.67, s.d. = 14.920) performed worse than both Group 2 (aged 20–21, M = 45.68, s.d. = 21.529, p = 0.000) and Group 3 (aged 30–70, M = 44.33, s.d. = 24.380, p = 0.012) There were no significant differences between Group 2 and Group 3 for LLAMA_B. For LLAMA_D again Group 1 (aged 10–11, M = 18.50, s.d. = 13.528) performed significantly worse than Group 2 (aged 20–21, M = 29.32, s.d. = 17.206, $p = 0.010$). There were no significant differences between Group 3 (aged 30–70, M = 24.50, s.d. = 17.536) and either of the other groups.

Our first hypothesis was that we would not see any differences between the age groups for LLAMA_B because vocabulary is a skill thought to be relatively independent of critical or sensitive period effects (Milton, 2009).

Table 4: Results of LLAMA tests according to age.

		LLAMA_B	LLAMA_D
Group 1: 10–11	M	28.67	18.50
(n = 30)	s.d.	14.920	13.528
Group 2: 20–21	M	45.68	29.32
(n = 44)	s.d.	21.529	17.206
Group 3: 30–70	M	44.33	24.50
(n = 30)	s.d.	24.380	17.536

This hypothesis was disconfirmed; the younger groups performed significantly worse than the two older groups. Our alternate hypothesis for LLAMA_B was that the older participants would outperform the younger ones (Miralpeix, 2006, 2009) due to their superior cognitive abilities, and this hypothesis was confirmed. Our third hypothesis was that the younger group (10–11-year-olds) would outperform the older groups on LLAMA_D because this taps into implicit learning processes (Granena, 2013, 2016; Skehan, 2016), which may be subject to critical period effects. The results disconfirmed this hypothesis as well; the younger learners (10–11-year-olds) performed significantly worse than the 20–21-year-olds.

Overall the younger learners scored lower on both tests. We therefore advise caution when using the LLAMA tests with children. Separate norms may be required for younger age groups. Alternatively, we may have to conclude that the current LLAMA tests are not suitable for use with younger learners. This would be particularly relevant for researchers investigating the role of aptitude in different age groups, as purported differences between younger versus older learners in the relevance of aptitude to their learning situation may be artefacts of the test rather than any comment on aptitude itself (cf. Abrahamsson & Hyltenstam, 2008). Further investigation with larger groups across the whole LLAMA test battery would be required to fully address this. It should be noted that while the 10–11-year-olds did not report any problems in actually taking these two tests,[16] these tests are based on the original MLAT tests (Carroll & Sapon, 1959), and alternate versions of the MLAT for younger learners have since been specifically developed.

5.4. Research question 4: individual differences

For this final research question, we combined the results from this study with Rogers et al. (2016) to examine how much of the variance in LLAMA test scores can be accounted for by six individual background variables. These variables are the three we have examined so far – L1 script (language neutrality), L2 status and age plus three other variables – highest formal education qualification, gender and logic puzzles. In total we tested 404 participants, although the 10–11-year-olds did not take the whole test battery. We included these additional variables to examine whether the tests were influenced by formal education or by logic training (e.g. playing chess or sudoku) because previous research into aptitude has suggested links between IQ and MLAT scores (Sasaki, 1999; Wesche, 1981).[17]

The multiple regression results from all 404 participants show that for LLAMA_B (vocabulary), these six variables accounted for 9.1% of the variance but only L2 status (i.e. whether the participant was bilingual, monolingual or had received L2 instruction) reached significance ($\beta = -0.250$, $p < 0.05$) and contributed 6.0% to the overall variance.

In total, 375 participants took the LLAMA_D test of implicit learning. The multiple regression results show that together these six factors accounted for 4.8% of the overall variance. In terms of the individual factors, L2 status and gender both reached significance. L2 status contributed 1.8% to the overall variance ($\beta = 0.136$, $p = 0.012$). Gender contributed 1.3% to the overall variance ($\beta = 0.116$, $p = 0.030$).

LLAMA_E is the measure of sound-symbol correspondence, and 370 participants took this test. The multiple regression shows that the six factors account for 3.4% of the overall variance. Only the playing of logic games reached significance ($\beta = 0.152$, $p = 0.004$) and contributed 2.3% to the overall variance.

Finally, 346 participants took LLAMA_F, the grammatical inferencing measure. Overall, the multiple regression shows that the six factors accounted for 6.6% of the overall variance with two individual factors reaching significance. These were L2 status and L1 script. L2 status contributed 2.6% to the overall variance ($\beta = -0.165$, $p = 0.002$) and L1 script accounted for 1.3% of the total variance ($\beta = 0.114$, $p = 0.036$).

Overall, the results of the multiple regression analysis suggest that the LLAMA tests can generally be used across different L1s, with male and female participants of differing education levels and with different ages, as these do not consistently affect the overall variance in LLAMA scores. The only consistent finding is that prior instruction in a second language can account for significant amounts of variance in LLAMA_B (6%) and LLAMA_F (2.6%). This suggests that the LLAMA tests are robust and not subject to significant external factors or individual variables that would influence their results although we make no claims regarding how well they measure aptitude (however defined).

6. Conclusions and Future Research

Overall using a large sample, we have shown that the LLAMA aptitude tests are robust as they are not subject to external individual differences. Our results confirm previous studies by Granena (2013) and Rogers et al. (2016). Additionally, we have identified two possible limitations of the tests in their use with younger children and in mixed L2/L3 groups. This study represents a significant step in the ongoing validation of the LLAMA tests, as we have recruited a large number of participants and provided a thorough examination of the tests in terms of targeted individual differences that could affect test performance.[18] However, the LLAMA tests still need to be validated in terms of their ability to predict language learning. Skehan (2016) highlights the changing role in aptitude validation work from the macro (large-scale) predictive studies to the more micro studies looking at language-learning processes. Within this latter framework, Skehan (2016) links aptitude to stages of acquisition and Wen (2016) considers whether working memory is the key component in language-learning aptitude. Our scrutiny of the LLAMA tests establishes a strong platform to conduct a large-scale macro validation study for the LLAMA tests to put them on a level playing field with the other tests (e.g. MLAT) and to provide crucial norming data. But as our knowledge of the interaction between different components of aptitude grows then we will also need to consider how the LLAMA tests interact with other areas of intelligence and memory (Sasaki, 1999; Wen et al., 2017).

As one of the only free aptitude tests available to researchers, the attractiveness of the LLAMA tests is ongoing and increasing based on the number of citations on Google scholar (over 700 since 2013). As the LLAMA tests are currently being developed for cross-platform online access (unlike the current Windows-only downloadable versions) with LLAMA_B already available online, we expect this interest and use of the LLAMA tests to continue. It is within this context that we hope this study provides researchers using the LLAMA tests with some useful background and helpful caveats to the use of the LLAMA tests.

Notes

[1] Unfortunately, the LLAMA tests do not currently permit an item-analysis so standard reliability testing, e.g. Cronbach's alpha, is not possible at this time. We use the terms reliable in the sense that we want to examine if the tests are subject to external factors which would change the test scores.

[2] We would like to thank an anonymous reviewer for highlighting the following quote by Carroll (1981, p. 86).

> I must also state that I am in general sympathy with writers like Neufeld (1978) who want to emphasize that foreign language aptitude, whatever it is, is not fixed or innate. They may be correct, and I would like to believe they are. I am simply neutral on this matter ... Yet, what evidence I have suggests that foreign language aptitude is relatively fixed over long periods of an individual's life span, and relatively hard to modify in any significant way.

[3] In Li (2016, p.15) LLAMA_D is grouped with LLAMA_E as an explicit measure of phonemic coding in Table 1, which outlines the different sub-components of the tests. We think this is in error. All previous work looking at the sub-components of the LLAMA tests including work by Granena (2013) and Rogers (2016) and the original conception of the tests by Paul Meara, suggests that LLAMA_D is testing something different to the other LLAMA tests. Granena (2013) has suggested this is an implicit measure and we would follow this approach.

[4] In the new online version of the test, one point is awarded for each correct answer with a maximum total of 20. As these participants took the downloadable version, we have kept those scores here.

[5] Earlier versions of this program had a fault that meant the maximum score possible was 75 contra the LLAMA manual (Meara, 2005). This error has now been fixed but in the version reported in this study, the maximum score was 75.

[6] The LLAMA manual suggests that test-takers may make notes during this test. We have conducted two versions of the test; one in which our participants could take notes (this study, n = 211) and our previous study in which participants could not take notes (n = 135). A t-test did not show any difference ($t(344) = 0.268$, $p = 0.789$) between participants who were allowed to take notes (M = 41.42, s.d. = 26.28) and those who were not (M = 42.22, s.d. = 28.35). Anecdotally, we noticed that those who were permitted to take notes did so and also made use of the full five minutes of learning time, whereas those who could not take notes did not use the full five minutes. We also noted that quite a few of the note takers wrote out the sentences as a whole and drew pictures. They then tried to work out the rules in the testing phase rather than using the learning phase to do so. This was contrary to the instructions they were given.

[7] For a more detailed discussion of the tests see Rogers et al. (2016). The entire suite of tests and a comprehensive test manual (Meara, 2005) can be downloaded from http://www.lognostics.co.uk/tools/llama/index.htm. A newer, web-based version of LLAMA_B is available at http://www.lognostics.co.uk/tools/LLAMA_B/LLAMA_B.htm.

[8] In Rogers et al. (2016) we argued that there was a pattern recognition or detection element in LLAMA_E due to the layout of the test and that this might influence these scores.

[9] Our thanks to an anonymous reviewer for pointing this out.

[10] Unfortunately, due to the characteristics of this group, we are unable to make any specific comments on cognitive decline. This is an area we are currently investigating in a follow-up study.

[11] We would like to thank an anonymous reviewer for highlighting this alternative hypothesis.

[12] Please see Hulstijn (2005) for a detailed discussion of the differences in implicit and explicit learning, knowledge and memory.

[13] These data were collected as part of some of the authors' undergraduate dissertations.

[14] We checked that none of the sounds used in LLAMA_E were allophones in Mandarin or Cantonese in case that was the source of this differences, but they are not allophonic for Mandarin or Cantonese.

[15] As a follow-up, Rogers et al. (2017) investigate the relationship between aptitude and the number of languages known (Hyltenstam, 2016), as this study cannot tease this apart. Our results show a significant effect for learners who have learnt more than one second language.

[16] Nektaria Kourtali (p.c.) tested 147 L1 Greek, L2 English 10–13-year-olds on all four sub-components as part of her current PhD thesis and also did not find any problems with her participants taking the tests.

[17] Due to testing and time constraints we were not able to administer a standardised IQ test and used these measures as pseudo-surrogates.

[18] The next step in our on-going work with the LLAMA tests is to examine the relationship between working memory and the LLAMA tests in light of Wen's (2016) model incorporating phonological working memory and executive working memory with language-learning aptitude.

References

Abrahamsson, N., & **Hyltenstam, K.** (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*(04), 481–509. DOI: https://doi.org/10.1017/S027226310808073X

Akamatsu, N. (1999). The effects of first language orthographic features on word recognition processing in English as a second language. *Reading and Writing, 11*(4), 381–403. DOI: https://doi.org/10.1023/A:1008053520326

Bialystok, E., Luk, G., & **Kwan, E.** (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific Studies of Reading, 9*(1), 43–61. DOI: https://doi.org/10.1207/s1532799xssr0901_4

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? *Linguistic Perspectives on Second Language Acquisition, 4*, 1–68. DOI: https://doi.org/10.1017/cbo9781139524544.005

Bley-Vroman, R. (2009). The evolving context of the fundamental difference hypothesis. *Studies in Second Language Acquisition, 31*(02), 175–198. DOI: https://doi.org/10.1017/S0272263109090275

Carroll, J. B. (1973). Implications of aptitude test research and psycholinguistic theory for foreign-language teaching. *Linguistics, 11*(112), 5–14. DOI: https://doi.org/10.1515/ling.1973.11.112.5

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In: Diller, K. C. (Ed.), *Individual Differences and Universals in Language Learning Aptitude*, 83–118. Newbury House.

Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In: Parry, T. S., & Stansfield, C. W. (Eds.), *Language Aptitude Reconsidered*, 11–29. Washington, D.C.: ERIC Clearinghouse of Languages and Linguistics.

Carroll, J. B., & **Sapon, S. M.** (1959). *Modern Language Aptitude Test*. New York: Psychological Corporation.

Crystal, D. (1987). *The Encyclopedia of Language*. Cambridge: Cambridge University Press.

De Graaff, R. (1997). The eXperanto experiment. *Studies in Second Language Acquisition, 19*(02), 249–276. DOI: https://doi.org/10.1017/S0272263197002064

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(04), 499–533.

Dörnyei, Z., & **Skehan, P.** (2003). Individual differences in second language learning. In: Doughty, C. J., & Long, M. H. (Eds.), *The Handbook of Second Language Acquisition*, 589–630. Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470756492.ch18

Gholamain, M., & **Geva, E.** (1999). Orthographic and cognitive factors in the concurrent development of basic reading skills in English and Persian. *Language Learning, 49*(2), 183–217. DOI: https://doi.org/10.1111/0023-8333.00087

Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In: Granena, G., & Long, M. H. (Eds.), *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment, 35*, 105–130 Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/lllt.35.04gra

Granena, G. (2016). Explicit and implicit cognitive aptitudes and information-processing styles: An individual differences study. *Applied Psycholinguistics, 37*(3), 577–600. DOI: https://doi.org/10.1017/S0142716415000120

Green, D., & **Meara, P.** (1987). The effects of script on visual search. *Second Language Research, 3*(2), 102–113. DOI: https://doi.org/10.1177/026765838700300202

Grigorenko, E. L., Sternberg, R. J., & **Ehrman, M. E.** (2000). A theory-based approach to the measurement of foreign language learning ability: The canal-F theory and test. *The Modern Language Journal, 84*(3), 390–405. DOI: https://doi.org/10.1111/0026-7902.00076

Hamada, M., & **Koda, K.** (2008). Influence of first language orthographic experience on second language decoding and word learning. *Language Learning, 58*(1), 1–31. DOI: https://doi.org/10.1111/j.1467-9922.2007.00433.x

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition, 27*(02), 129–140. DOI: https://doi.org/10.1017/S0272263105050084

Hyltenstam, K. (2016). *Advanced Proficiency and Exceptional Ability in Second Languages*. Walter de Gruyter GmbH & Co KG. DOI: https://doi.org/10.1515/9781614515173

Kaushanskaya, M., & **Marian, V.** (2009). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review, 16*(4), 705–710. DOI: https://doi.org/10.3758/PBR.16.4.705

Kormos, J. (2013). New conceptualizations of language aptitude in second language attainment. In: Granena, G., & Long, M. H. (Eds.), *Sensitive Periods, Language Aptitude and Ultimate Attainment*, 131–152. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/lllt.35.05kor

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition, 38*(4), 801–842. DOI: https://doi.org/10.1017/S027226311500042X

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Doughty, C. J., et al. (2013). Hi-lab: A new measure of aptitude for high-level language proficiency. *Language Learning, 63*(3), 530–566. DOI: https://doi.org/10.1111/lang.12011

McLaughlin, B. (1990). The relationship between first and second languages: Language proficiency and language aptitude. In: Harley, B. (Ed.), *The Development of Second Language Proficiency*, 158–178. Cambridge University Press. DOI: https://doi.org/10.1017/cbo9781139524568.014

Meara, P. (2005). *LLAMA Language Aptitude Tests: The Manual* (Tech. Rep.). Swansea: Lognostics.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Miralpeix, I. (2006). Age and vocabulary acquisition in English as a foreign language (EFL). In: Muñoz, C. (Ed.), *Age and the Rate of Foreign Language Learning*, 89–106). Bristol: Multilingual Matters.

Miralpeix, I. (2009). Lexical knowledge in instructed language learning: The effects of age and exposure. *International Journal of English Studies, 7*(2), 61–84.

Muñoz, C. (2014). The association between aptitude components and language skills in young learners. In: Pawlak, M., & Aronin, L. (Eds.), *Essential Topics in Applied Linguistics and Multilingualism*, 51–68. Springer.

Nayak, N., Hansen, N., Krueger, N., & **McLaughlin, B.** (1990). Language-learning strategies in monolingual and multilingual adults. *Language Learning, 40*(2), 221–244. DOI: https://doi.org/10.1111/j.1467-1770.1990.tb01334.x

Neufeld, G. G. (1978). A theoretical perspective on the nature of linguistic aptitude. *IRAL-International Review of Applied Linguistics in Language Teaching, 16*, 15–26. DOI: https://doi.org/10.1515/iral.1978.16.1-4.15

Patkowski, M. S. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning, 30*(2), 449–468. DOI: https://doi.org/10.1111/j.1467-1770.1980.tb00328.x

Petersen, C. R., & **Al-Haik, A. R.** (1976). The development of the defense language aptitude battery (DLAB). *Educational and Psychological Measurement, 36*(2), 369–380. DOI: https://doi.org/10.1177/001316447603600216

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (Forms)*. Harcourt, Brace and World Incorporated.

Rogers, V. E. (2016). Measuring Individual Differences with the LLAMA Aptitude Tests. *Paper presented at Psychology of Language Learning, 2*. University of Jyväsklä, Finland.

Rogers, V. E., Chisholm, M., Clothier, J., Cobner, A., Galvin, T., & **Greenfield, I.** (2017). *How Does Aptitude Relate to Working Memory?* (Accepted for poster presentation at EUROSLA 2017, University of Reading).

Rogers, V. E., Meara, P., Aspinall, R., Fallon, L., Goss, T., Keey, E., & **Thomas, R.** (2016). Testing aptitude. *EUROSLA Yearbook, 16*(1), 179–210. DOI: https://doi.org/10.1075/eurosla.16.07rog

Safar, A., & **Kormos, J.** (2008). Revisiting problems with foreign language aptitude. *IRAL-International Review of Applied Linguistics in Language Teaching, 46*(2), 113–136. DOI: https://doi.org/10.1515/IRAL.2008.005

Sampson, G. (1985). *Writing Systems: A Linguistic Introduction*. Stanford University Press.

Sasaki, M. (1999). *Second Language Proficiency, Foreign Language Aptitude, and Intelligence*. Peter Lang.

Sawyer, M. (1992). Language aptitude and language experience: Are they related? *Working Papers, 3*, 27–45.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In: Granena, G., Jackson, D. O., & Yilmaz, Y. (Eds.), *Cognitive Individual Differences in L2 Processing and Acquisition*. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/bpa.3.02ske

Sparks, R. L., Ganschow, L., Fluharty, K., & **Little, S.** (1995). An exploratory study on the effects of Latin on the native language skills and foreign language aptitude of students with and without learning disabilities. *The Classical Journal, 91*(2), 165–184.

Sparks, R. L., & **Patton, J.** (2013). Relationship of L1 skills and L2 aptitude to L2 anxiety on the foreign language classroom anxiety scale. *Language Learning, 63*(4), 870–895. DOI: https://doi.org/10.1111/lang.12025

Sternberg, R. J., & **Grigorenko, E. L.** (2002). *Dynamic Testing: The Nature and Measurement of Learning Potential*. Cambridge University Press.

Tan, L. H., Spinks, J. A., Feng, C.-M., Siok, W. T., Perfetti, C. A., Xiong, J., Gao, J.-H., et al. (2003). Neural systems of second language reading are shaped by native language. *Human Brain Mapping, 18*(3), 158–166. DOI: https://doi.org/10.1002/hbm.10089

Thompson, A. S. (2013). The interface of language aptitude and multilingualism: Reconsidering the bilingual/multilingual dichotomy. *The Modern Language Journal, 97*(3), 685–701. DOI: https://doi.org/10.1111/j.1540-4781.2013.12034.x

Tolchinsky, L., Levin, I., Aram, D., & **McBride-Chang, C.** (2012). Building literacy in alphabetic, abjad and morphosyllabic systems. *Reading and Writing, 25*(7), 1573–1598. DOI: https://doi.org/10.1007/s11145-011-9334-7

Wen, Z. E. (2016). *Working Memory and Second Language Learning: Towards an Integrated Approach*. Bristol: Multilingual Matters.

Wen, Z. E., Biedroń, A., & **Skehan, P.** (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching, 50*(1), 1–31. DOI: https://doi.org/10.1017/S0261444816000276

Wesche, M. (1981). Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In: Diller, K. C. (Ed.), *Individual Differences and Universals in Language Learning Aptitude*, 119–154. Rowley, MA: Newbury House.

How to cite this article: Rogers, V., Meara, P., Barnett-Legh, T., Curry, C. and Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, 1(1), 49–60, DOI: https://doi.org/10.22599/jesla.24

Submitted: 17 January 2017 **Accepted:** 28 June 2017 **Published:** 01 August 2017

JESLA *Journal of the European Second Language Association* is a peer-reviewed open access journal published by White Rose University Press.

OPEN ACCESS