Title: Testing Aptitude: Investigating Meara's (2005) LLAMA tests

Authors: Rogers, Vivienne .E, Meara, Paul. M, Aspinall, Rachel, Fallon, Louise, Goss, Thomas, Keey, Emily, & Thomas, Rosa.

Affiliation: Swansea University

Abstract

Meara (2005) developed the LLAMA tests as a free, language-neutral, user-friendly suite of aptitude tests incorporating four separate elements: vocabulary learning (LLAMA_B), phonetic (implicit) memory (LLAMA_D), sound-symbol correspondence (LLAMA_E) and grammatical inferencing (LLAMA_F) based on the standardised MLAT tests (Carroll & Sapon, 1959). Recently, they have become increasingly popular in L2 acquisition research (Grañena & Long, 2013b). However, Meara has expressed concern about the wide use of these tests without validity testing (cf. Grañena 2013a). To this end, we investigated several areas relating to the LLAMA tests, i.e. (1) the role of gender in LLAMA test performance; (2) language neutrality; (3) the role of age; (4) the role of formal education qualifications; (5) the effect of playing logic puzzles on LLAMA scores and (6) the effect of changing the test timings to scores. 229 participants from a range of language backgrounds, aged 10-75 with various education levels, typologically distinct L1s, and varying levels of multilingualism were tested. A subset of participants was also tested with varying timings for the tests. The results showed that the LLAMA tests are gender and language neutral. The younger learners (10-11s) performed significantly worse than the adults in the sound/symbol correspondence task (LLAMA_E). Formal education qualifications show a significant advantage in 3 of the LLAMA subcomponents (B, E, F) but not the implicit measure (LLAMA_D). Playing logic puzzles did not improve LLAMA test scores. The timings appear to be optimal apart from LLAMA_F, which could be shortened. We suggest that the LLAMA aptitude tests are not significantly affected by these factors although researchers using these tests should be aware of the possible impact of education level on some components of the tests.

Introduction:

In this paper we seek to establish if the LLAMA aptitude tests (Meara 2005) are influenced by factors other than aptitude, namely if the tests are unduly influenced by factors such as gender, first language, age, education level, L2 status (whether the person has already learnt a second language) etc. The rationale for this study is the increasing use of the LLAMA tests in research and Meara's concerns about the lack of validity testing of the LLAMA tests. In subsequent sections we will outline how the LLAMA tests were developed and their original purpose as teaching tools for MA students studying research methods. We make no claims in this paper to whether or not the LLAMA tests actually predict the rate that learners acquire another language (please see below for a fuller definition of language learning aptitude).

Language Aptitude has been the object of empirical study in Applied Linguistics for a considerable time. It came of age in the 1950s, when two major test batteries that claimed to measure aptitude were published. The most widely-used of these batteries was Carroll and Sapon's *Modern Language Aptitude Tests* (MLAT) published in 1959. This work identified four main factors that affected people's ability to learn a new language – the ability to learn words out of context, grammatical sensitivity, phonetic sensitivity and inductive learning ability. The final form of MLAT contained subtests that assessed each of these factors separately, and produced an overall aptitude score. It was designed for use with English native speakers to examine the rate with which they could acquire a new language. The *Pimsleur Modern Language Aptitude Battery* (PLAB) was published in 1966 by Paul Pimsleur. Like MLAT, PLAB identified a number of separate factors in Language Aptitude – vocabulary size in English is taken as a measure of overall verbal ability, language analysis measures whether test-takers can pick up grammatical patterns in a new language, sound discrimination measures auditory skills and sound-symbol association measures test-takers' ability to associate sounds to symbols other than their familiar orthographic representations. Pimsleur also included a grade point average and a measure of general interest in languages – this last component being taken as a measure of motivation for acquiring a new language. A subsequent third major aptitude test, the *Defense Language Aptitude Battery* (DLAB), an influential tool designed to assess language aptitude in military contexts, is described in Peterson and Al-Haiq (1976). The existence of these standardised tests generated a large amount of research on language aptitude, and a substantial, critical account of this work can be found in Parry and Stansfield (1990).

Moving on to the role of aptitude in language acquisition, this has been investigated in relation to several key areas, e.g. age, education levels, etc. It is clearly important to establish if the test taken to measure aptitude is unduly influenced by external factors itself (i.e. test effects) or if the results ascribed to aptitude actually pertain to aptitude. In the following section we will review some of the areas that have been investigated in relation to aptitude and which motivate the factors we will subsequently examine.

Carroll (1990: 26) defines aptitude as “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”. He included four areas for testing, namely phonemic coding ability, grammatical sensitivity, inductive language- learning ability and rote-learning ability and these form the rationale for the sub-components of the MLAT test. Since the LLAMA tests (outlined below) are based on MLAT test, we will take this definition as our working hypothesis. In terms of areas of research, aptitude is one of several individual differences (e.g. motivation, anxiety, working memory) that have been studied in second language acquisition research. Carroll's definition (given above) has been challenged on a number of grounds including whether aptitude tests should include specific tests for motivation and anxiety measures, whether aptitude is static or changes over time and whether aptitude overlaps with intelligence (Dörnyei 2010; Grañena & Long 2013a; Ortega 2009; Robinson 2001, 2013; Skehan 2002; Snow 1992; Sparks & Ganschow 2001; Wesche 1981). In the following section, we will consider three areas of aptitude research that could influence LLAMA test scores irrespective of aptitude. These areas are: the effect of education level, the role of age and the effect of the first language script.

The effect of education

Several studies have investigated the role of education level and more specifically second language learning experience on aptitude. Longitudinal studies have found similar results in test-retest participants suggesting that there is no change over time whereas research conducted on monolinguals, bilinguals and multilinguals suggested aptitude development significantly correlates to language experience and therefore would change over time (Eisenstein 1980; Kormos 2013; Sáfár & Kormos 2008; Sawyer 1992; Sparks, Ganschow, Fluharty, & Little 1995; Thompson 2013). This implies that language aptitude tests might need to consider previous language experience of the test-takers. However, Nayak, Hansen, Krueger, N., & McLaughlin (1990) found that overall there was no significant evidence that multilingual learners are more successful language learners, but that they were more able and willing to adjust their L2 learning strategy to facilitate specific language components. This suggests that aptitude, or at least performance on an aptitude test, can be altered with language training (Carroll 1981; Dornyei & Ushioda 2009; Grañena & Long 2013a; Grigorenko, Sternberg, & Ehrman 2000; McLaughlin 1990; Sáfár & Kormos 2008; Sternberg 2002). The impact of education or language learning experience may be due to the aptitude tests' focus on "analytical and analogical skills and not on the student's potential for the development of more global skills also needed for communication" (Oxford, 1990: 74). Similar effects have also been shown in intelligence (IQ) testing, e.g. Ceci's (1991) review concluded that the amount of schooling had a positive correlation with IQ.

Age

Various researchers have also questioned whether aptitude is relevant for different age groups. Several researchers have suggested that aptitude is only relevant for older learners possibly in relation to a critical period for language learning (Flege, Yeni-Komshian & Liu 1999; Johnson & Newport 1989; Bidrsong & Molis 2001). DeKeyser (2000) investigated the role of aptitude in older and younger naturalistic L2 learners in terms of morphosyntax. He found that aptitude (verbal analytical ability) was a predictor of attainment in the older learners and argued that this is due to their lack of implicit learning abilities[1] post critical period[2]. However, Grañena and Long (2013a) found that aptitude scores significantly correlated with age in relation to vocabulary and collocation measures but not for morphosyntax in their adult Chinese learners of Spanish. Abrahamsson and Hyltenstam

[1] Implicit language learning is learning without conscious awareness or intention. It can also be described as incidental learning. This is in contrast to intentional or explicit learning. Please see DeKeyser (2008) for a fuller discussion. Whether aptitude is relevant for implicit or explicit conditions has been widely debated in the field. Krashen (1981) argues that it is only relevant for explicit learning. On the other hand, Nation and MacLaughlin (1986) counter that it is only relevant for implicit learning and Robinson (2001) argues that it is relevant for both implicit and explicit learning.

[2] In this paper DeKeyser (2000) argues that apparent critical period effects in adults are actually due to a lack of access to implicit learning mechanisms.

(2008) extended DeKeyser’s study and included a wider range of language tasks (including tests of phonology, lexis and morphosyntax). They found that aptitude was a relevant factor for adult and adolescent learners (over 13s) in order to sound like a native speaker but contra DeKeyser (2000) they also found a small role for aptitude in younger L2 learners (age of onset before 11) if they were to become indistinguishable from native speakers. This result is supported by Muñóz (2014) who investigated 48 bilingual Spanish-Catalan Primary school learners of English aged 10-11 and 11-12 to assess their aptitude scores and language ability. The study compared aptitude scores with speaking, listening, reading and writing language components. The results showed significant correlations with all components. The issue of testing younger children for aptitude relies on having aptitude tests that are accessible by younger children. Muñóz used a version of the MLAT for younger children (e-MLAT: Carroll & Sapon 1967) but this is not available in all aptitude tests. In this study, we will examine how younger learners (aged 10-11) perform on two parts of the LLAMA tests.

Language neutrality

A final area of discussion regarding language aptitude and the aptitude tests specifically relates to the language neutrality of the test or the role of L1 language script. 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 & Gera 1999; Hamada & Koda 2008; Green & Meara 1987). Wong and Pyun (2012) examined two separate groups of L1 English students, one learning L2 French and the other L2 Korean, to investigate the effects of sentence writing on L2 lexical retention over two tasks. One task involved the writing of new words in sentences, and the other involved repeated vocabulary and picture learning. An immediate and two delayed post-task scores revealed that the L2 Korean groups’ scores were much lower in the sentence writing activity, suggesting that sentence writing results in less retention when the L1 script differs from the L2 script. In addition, Hamada and Koda (2008) examined L1 orthographic (script) influence on decoding and meaning for L2 words. English L2 learners with typologically similar (Roman script) and typologically distinct (logographic) L1s learned the meaning of words with pictures and decoding was measured by a word naming task with phonologically regular and irregular patterns. Recall tasks found that the typologically similar (roman script) group produced greater overall retention, suggesting that L1 and L2 orthographic distance influences L2 word learning. If the language script of the L1 can influence the acquisition of the L2 as discussed above, then the question arises if the L1 script of the learner influences their aptitude scores. We will return to this point in our discussion of the LLAMA tests below.

The brief outline of some of the research questions surrounding aptitude shows that research into aptitude has had resurgence in recent years. This has led to the development of alternative aptitude tests which attempt to capitalise on the multimedia possibilities offered by new technologies. These tests are generally not as well researched as MLAT, PLAB and DLAB given their later arrival, but they appear to be more flexible, more adaptable, easier to use, and simpler to score. Above all, they offer a cheap (often free) resource and this potentially makes them attractive to researchers with limited resources. One such set of tests is the LLAMA tests (Meara 2005). These tests were not originally designed as research tools, or as formal test instruments. Rather, the LLAMA tests were an exploration of the coding possibilities provided by high level visual programming tools like DELPHI (Manning 1995), which made it very easy to produce professional looking programs with Windows-like interfaces, and could handle sound files with ease. Re-programming MLAT appeared to be a useful way of exploring these features. Once the programs were established, they were used as part of a research methods course at Swansea University. Students were provided with copies of the LLAMA programs, and asked to test whether they were any good as predictors of language aptitude. The programs had some obvious failings, such as a narrow range of scores and scores tending to cluster at one end of the rating scale. Perhaps more importantly the students found it extremely difficult to find tests of proficiency against which the LLAMA tests could be evaluated, for example there are no standardised tests of proficiency in French which could be used as a criterion variable in studies of this sort.

The LLAMA tests are loosely based on the MLAT tests, in that they attempt to measure the same factors that Carroll and Sapon (1959) identified as components of language aptitude. However, the tests were explicitly designed to overcome two problems which seemed to limit the MLAT tests. One of these limitations was the time it took to complete the MLAT tests: this was solved by making the LLAMA tests shorter than the corresponding MLAT tests, and by giving them a simpler and more appealing user interface. The tests differ from the MLAT tests in they are designed to be completed in a short space of time (less than 20 minutes as opposed to the MLAT’s one hour (Carroll & Sapon 1959)) , and to provide instant feedback to the users. The second limitation of MLAT was that these tests were designed for native speakers of English, and a separate version was required if you were working with test takers from other backgrounds. LLAMA was specifically designed to be L1-independent, in that the interfaces made little use of L1 instructions, and the tests did not require L1 responses.[3]There are four parts to the LLAMA tests: LLAMA_B is a vocabulary learning task, LLAMA_D is an implicit learning task, LLAMA_E is a sound-symbol

[3] The LLAMA tests are freely available on Paul Meara’s website (http://www.lognostics.co.uk/tools).

correspondence task and LLAMA_F is a grammatical inferencing task. Further details of each component are given below.

The LLAMA tests

There are four LLAMA tests, conventionally (though somewhat unhelpfully) referred to as LLAMA_B, LLAMA_D, LLAMA_E and LLAMA_F. The slightly awkward names arose because the tests that make up the LLAMA suite were the only survivors from a much larger set of programs that were developed around the same time. These surviving programs are described in more detail in the following paragraphs

LLAMA_B[4] is loosely based on the original vocabulary learning subtask in Carroll and Sapon (1959), but it uses a completely new interface which results in a test format that is largely independent of the test-taker's first language. This interface is shown in Figure 1 below.

"Put Figure 1 about here please"

The display presents a set of twenty unusual objects that can be loosely associated with familiar objects, but do not have obvious English language names. Clicking on one of these objects causes its name to be displayed in the central panel. These names are all real words, names for common objects in a Central American language, and they are arbitrarily assigned to the target images. Using words from a real language, as opposed to using made up words means that the target words are linguistically coherent in a way that made up nonsense names generally are not (Meara, 2013). Once the panel is displayed, test-takers have a fixed length

[4] Since this paper was written, a new web-based version of LLAMA_B has become available. This study used the previous downloadable version.

of time to work with the display. This time is usually two minutes, but the program allows the experimenter to vary this time if necessary. Test-takers can use any strategy they want to work with the data. For example, they can focus their attention on a small number of items and ignore the rest. Or they can try and learn all twenty target words. When the two-minute time interval is complete, LLAMA_B enters a test phase, where the central panel displays a target word, and the test-taker has to identify which of the twenty objects this name is associated with, by clicking on the appropriate image. Immediate feedback is provided for each answer. The entire test takes about 10 minutes. Test-takers score one point for each object that is correctly identified by its name. There is no correction for guessing.

LLAMA_D is a new task that does not appear in the work of Carroll and Sapon (1959). This test is designed to assess whether test-takers can recognise short stretches of spoken language that they were exposed to a short while previously. The program is loosely based on work by Service (e.g. Service 1992; Service and Kohonen 1995) and it also owes something to Speciale, Ellis and Bywater (2004). These writers suggest that a key skill in language ability is a person's ability to recognise patterns, particularly patterns in spoken language. If they can recognise repeated patterns, then they are more likely to be able to recognise words when you hear them for a second time. This helps them to acquire vocabulary. It also helps them to recognise the small variations in endings that many languages use to signal grammatical features. The sound sequences used in this program are computer generated. The words they are based on are the names of flowers and natural objects in a British Columbian Indian language. The sounds have been synthesised using the AT&T Natural Voices (French). This makes for a difficult set of stimuli which are unlikely to be recognised as belonging to any major language family.

The LLAMA_D interface is shown in Figure 2 below. Clicking on the ↲ button plays a series of short sound clips. In phase one of the test, test-takers simply listen to these sound clips. In phase two the original sound clips are intermingled with some new sound clips. After each clip, the test-taker has to indicate whether the sound clip has appeared before.Test-takers get one point for each repeated sound clip that they recognise, and they are penalised for guessing. The entire test takes just over five minutes. It generally gets positive comments from users, but appears to be very hard.

“Put Figure 2 about here please”

LLAMA_E is a sound-symbol correspondence task and an adaptation of the original sound-symbol correspondence test that appeared in Carroll and Sapon (1959). The approach used in LLAMA_E appears to be rather more user friendly than the original version, and it incorporates a number of features which bias the test in favour of test-takers who have some familiarity with phonetic theory. It seems to be particularly good at identifying learners who are able to dissociate sounds from the way that they are normally written in English. The interface displays a set of 24 buttons, each carrying a pair of symbols. Clicking on a button causes the program to play a sound file that consists of a single syllable. The test-taker’s task is to decipher the symbols, and work out which features of the orthography correspond to the sounds of this “language”. In phase one of the test, test-takers get two minutes to explore the interface, and are allowed to click any of the buttons as often as they wish. In phase two, the program plays a complex two syllable sound, and displays two possible spellings for this sound. Test-takers are required to decide which of the two spellings accurately represents the sound. Points are scored for correct answers; points are lost for incorrect answers.

“Put Figure 3 about here please”

LLAMA_F tries to assess test-takers’ grammatical inferencing skills. The interface is shown in Figure 4 below.

“Put Figure 4 about here please”

Each of the twenty small buttons is associated with a picture shown here on the right of the interface, and a short sentence displayed in the middle of the interface. By comparing a number of pictures, test-takers should be able to work out which parts of the sentence correspond to the different features of the picture, and from this, they should be able to establish a number of grammatical features which characterise the way the language works. These features include word-order, gender for nouns, singular, dual and plural number, and so on. Phase two of the program displays a new picture incorporating the familiar elements, and two sentences which might describe it. The test-takers' task is to identify the sentence which is grammatically and semantically correct. The entire test takes about five minutes. Test-takers get one point for each correctly identified sentence, and lose points for guessing.

Some research studies using the LLAMA tests

A number of established researchers and early career / doctoral researchers have used these tests in some major research projects. At the time of writing according to Google Scholar, the LLAMA tests appear to have been cited 45 times in published research projects since they were made available in 2005. This works out at about five citations per year. Grañena and Long (2013b) is a particularly important collection of papers that makes extensive use of the LLAMA tests. Some examples of the types of research recently carried out using the LLAMA tests are given below.

Grañena and Long (2013a) used LLAMA results to suggest aptitude develops a mitigating role on age effects within lexis and collocation. Larson-Hall and Dewey (2012) used LLAMA_B and LLAMA_F results to suggest aptitude is a more important factor than motivation to L2 acquisition. Yilmaz (2013) used LLAMA_F as a measure of cognitive ability and found that students with high aptitude scores benefited from negative correction. Grañena (2013b) used LLAMA_D as a sound sequence test and concluded sequence learning ability affected L2 attainment specifically understanding L2 agreement factors. Lundell and Sandgren (2013) expanded on Grañena and Long's (2013a) investigation using all four LLAMA subtests and concluded aptitude should be considered a personality factor. De Bot (2013) found LLAMA_E and LLAMA_F were sensitive to circadian rhythms and concluded aptitude is individualised and influenced by learner preferences. Smeds (2012) suggested LLAMA_D can identify whether blindness improves verbal or auditory functions. Xiang et al. (2012) used all four LLAMA subtests to illustrate high aptitude scores relate to strong structural connectivity of neural language pathways. This brief overview shows that the LLAMA tests are being used to address a wide range of research questions. Moreover, Larson-Hall and Dewey (2012) explicitly endorsed the LLAMA system after comparing it with the MLAT, suggesting scholars “should consider the use of the LLAMA tests in aptitude research” (2012: 74).

Given the high levels of interest in the LLAMA tests as research tools, we felt that it was important to carry out some validation studies of our own, and some preliminary work of this sort is reported in this paper. Some of our validation work is based on work by Grañena (2013a), who carried out a previous analysis of the LLAMA tests. She argues that the LLAMA tests are internally consistent but that they load on two different factors. Using an exploratory Principal Component Analysis (PCA) to determine whether the different subtests are correlated and therefore may be measuring the same or similar variables, the three tests that were designed to correspond to the sub-components of the MLAT all loaded on the same factor, i.e. vocabulary (LLAMA_B), sound-symbol correspondence (LLAMA_E) and grammatical inferencing (LLAMA_F). These three tests accounted for 45.15% of the total variance. The other test, that was a new addition aimed at measuring implicit learning (LLAMA_D), loaded on another factor and accounted for 22.82% of the variance. This split between LLAMA_D on one hand and the other 3 sub-tests on the other is further supported in Grañena’s work by a series of exploratory PCAs with LLAMA_D loading on the same factor as attention control and also a probabilistic SRT task that measure implicit learning ability whereas the other LLAMA tests loaded with more explicit measure like general intelligence (GAMA test). This potentially supports the inclusion of LLAMA_D as a measure of implicit learning. Grañena highlights the fact that LLAMA_B, E and F all have an explicit learning phase, which might appeal to more explicit learning strategies whereas LLAMA_D is a receptive task. In this paper, Grañena also examined the role of language neutrality and gender in the LLAMA test results with 186 participants from 3 typologically distinct L1s (English, Spanish and Chinese). The language backgrounds were chosen because they were typologically distinct. The results showed no significant differences between gender or L1 typology. The results of this study are important as they are the first attempt at a validation of the LLAMA aptitude tests but it is not clear on the exact composition of the groups and this could be extended to other L1s.

Research Questions

Since the LLAMA tests were developed as a teaching tool, the data collected for this study was collected by a group of 5 undergraduate students for their dissertations. This allowed them to develop the research questions given here and these are typical of the kinds of questions raised by the students in the original research methods course that the LLAMA tests were designed for (as discussed above). Following Grañena (2013a) the first two research questions examine the role of gender and whether the LLAMA tests are language-neutral, i.e. is there a difference in outcome between speakers of Roman and non-Roman scripts. The remaining research questions will consider some of the other factors outlined previously that might influence LLAMA test scores, such as age and education level. As researchers are using the LLAMA tests to investigate the effects of these variables on aptitude, it is important to ensure that these variables do not affect performance on the LLAMA tests themselves. Therefore, we asked:

1. What is the role of gender in LLAMA test performance?
2. Are the LLAMA tests language neutral?
3. What is the role of age?
4. What is the role of formal education qualifications?
5. Does playing logic puzzles affect LLAMA scores?
6. What difference would changing the test timings make to scores?

In the following section, we will briefly outline the tasks and general methodology for data collection. We will then present each research question in more detail with hypotheses, further detail of participants and the results.

General Methodology

This section outlines the overall tasks and methodology for this study but the specific details of the participants and hypotheses for each of the research questions will be outlined in the relevant results section.

Tasks

The test battery consisted of the LLAMA tests, an online background questionnaire and a consent form. Copies of the instruction handout, consent form and background questionnaire are provided in the appendix. Testing took place either on an individual basis or via larger drop-in sessions in a computer lab. The background questionnaires were administered online through Limesurvey or on paper for the younger participants who required parental consent. All participants were required sign the informed consent form (or obtain signed parental consent) before taking part in the study. The LLAMA tests are programmed to give a score for each component. The scoring procedure for each test in outlined above in the discussion of the individual tests. Participants were asked to fill out a slip of paper with their scores, which were collected by the students. No changes to scoring were made. All scores were entered into SPSS as given by the LLAMA tests. Details of how the background questionnaire was coded are given below for each research question as relevant.

Participants

In total 229 participants took part in this study. The data were collected by the five students for their dissertations. They worked together to collect as much data as possible and shared it between the group. They aimed to collect from as wide a range of ages, education levels, L1s as possible but as the participants were volunteers, this is a convenience sample. There were two main groups of participants in order to address RQ 6. The first group consisted of 164 subjects who took the LLAMA tests at the standard length and 65 subjects who took the LLAMA tests at altered lengths. More details about the altered lengths will be given in the results for RQ 6. Participants were aged between 10-75.

Methodology, results and discussion for each research question:

In this section each of the research questions will be reported in turn. Details of the participants involved will also be given as well as any hypotheses before the results. In the original student dissertations, the participants were carefully matched for age, education level, L1, whether or not they spoke another language (L2 status) and education level depending on the specific research question each student addressed. However, for the purposes of this paper the whole group of participants has been analysed. We will present the results for each research question before we turn to a final conclusion.

RQ1: Gender

The first research question addressed the role of gender in LLAMA test scores. Following Grañena (2013a) we expected to find no difference between genders. This is the null hypothesis (please see the limitations section at the end of the paper for further discussion regarding conceptual issues surrounding this). Only subjects who were over 18 and who had taken all of the LLAMA tests were included. This gave a total of 135 participants; 63 male and 72 female. The mean results and standard deviations are given in Table 1. T-tests showed no significant differences between male and female subjects for any of the LLAMA tests. For LLAMA_B (vocabulary), *t*(133)=.367 *p*=.729. For LLAMA_D (sound recognition), *t*(133)=.536 *p*=.904. For LLAMA_E (sound/symbol correspondence), *t*(133)=1.005 *p* =.488 and for LLAMA_F (grammar), *t*(133)=-.404 *p*=.456.

"Put table 1 about here please"

This result confirmed Grañena's (2013a) finding that there was no difference between male and female participants on the LLAMA tests.

RQ2: Language Neutrality

The second research questions investigated whether the language background of the participant affected the scores on the LLAMA test. As outlined above, the LLAMA tests were designed to be language neutral, i.e. they can be taken by speakers of any language, but both LLAMA-B (vocabulary) and LLAMA_F (grammatical inferencing) contain written non-words in a roman script. In the vocabulary test (LLAMA_B) there are a total of 20 words whereas in the grammatical inferencing test (LLAMA_F) there are 20 phrases – typically two word pairs. To establish whether this would disadvantage people whose first language uses a non-roman alphabet, 135 participants over the age of 18 took all 4 LLAMA tests. This resulted in 18 different L1s with small numbers for some languages (e.g. there were only 2 Norwegians). Since we were keen to examine the role of the L1 script in particular whether the non-Roman script L1 participants were disadvantaged in taking the test, we grouped these individual languages according to whether they used a roman or non-roman alphabet. We also separated the English speakers as they formed a large (predominantly monolingual) group from the participants who did not have English as their first language but whose first language had a roman script. The inclusion of the non-English Roman script group allowed for a comparison between participants taking the test with the instructions in their non-native language in case any differences between the non-Roman script group and the English group were due to the instructions being given in English and not due to alphabet or script differences. This gave three groups: L1 English (n=99), non-Roman script, e.g. Chinese (n=17)[5] and non-English Roman script, e.g. French (n=18)[6]. Given that L1 script has an influence on L2 attainment (Hamada & Koda, 2008; Wong & Pyun, 2012) we hypothesised that the non-Roman script group may score lower than the other groups in LLAMA_B (vocabulary) and LLAMA_F (grammatical inferencing) as they both contain roman scripted elements (see Figures 2 & 4 above). The means and standard deviations are given in Table 2.

“Put table 2 about here please”

A one-way ANOVA showed that there was a significant effect of alphabet/ script type on LLAMA_E (sound-symbol correspondence), *F*(2,131)=3.505 *p*=.033. Post-hoc analysis for LLAMA_E showed that the non-Roman script group significantly outperformed the English group (*p*=0.036). As the non-English Roman script group also outperformed the English group on LLAMA_E, the groups were re-coded into monolingual (n=73), bilingual, i.e. spoke 1 other language (n=42) and multilingual, i.e. spoke more than one other language

[5] The non-Roman script group included Chinese (13), Arabic (2), Greek (1) and Korean (1).
[6] The non-English roman script group includes French (3), Welsh (2), Spanish (3), Norwegian (2), German (3), Swiss-German (1), Danish (1), Yoruba (1) and Italian (2).

(n=20). This allows us to examine if the differences due to alphabet/script were actually because the non-Roman participants were all bilingual and most of the English group were not. A one-way ANOVA showed no significant effect on any of the LLAMA tests although LLAMA_B (vocabulary) approached significance $F(2,131)=3.052$ $p=.051$). The majority of the non-Roman participants were L1 Chinese speakers. Given that Chinese languages are logographic it is possible that Chinese learners are more familiar with learning abstract symbols to correspond to words as opposed to the roman alphabet system. However, in a follow-up study comparing L1 Chinese and L1 Arabic speakers, Rogers, Barnett-Legh, Curry & Davey (2015) found no advantage for the L1 Chinese speakers on LLAMA_E. In an overview bringing together the results from this study and Rogers et al (2015), which totalled 404 participants taking the LLAMA tests, Rogers (2015) found that language background was not a significant factor in a multiple regression analysis for any of the LLAMA tests except in LLAMA_F (grammatical inferencing) where it accounted for 1.3% of the variance of the scores ($\beta=.114$, $p<.05$).

This result did not completely agree with Grañena's (2013a) finding regarding language neutrality as we found that the non-Roman script group significantly outperformed the English group on LLAMA_E. However, the participant numbers in the non-Roman script group and non-English Roman script group were comparatively small. Contrary to our hypothesis that the non-Roman script group might perform lower on LLAMA_B and F, they performed the same as the other groups on LLAMA_B (vocabulary) and outperformed the others on LLAMA_F (see Table 2).

When the results were re-coded into monolingual, bilingual and multilingual, there were no significant differences between any of the groups (LLAMA_B: $F(2,131)=3.052$ $p=.051$, LLAMA_D: $F(2,131)=.088$ $p=.916$, LLAMA_E: $F(2,131)=1.970$ $p=.144$, LLAMA_F: *F*(2,131)=.311 *p*=.733). This is counter to the findings that aptitude can be altered with language training (e.g. Grigorenko et al. 2000; McLaughlin 1990; Sternberg 2002), or language experience (e.g. Eisenstein 1980; Kormos 2013; Sáfár & Kormos 2008; Sawyer 1992; Sparks et al. 1995; Thompson 2013). It should also be noted that throughout the rest of the results, participants score highest on LLAMA_E (sound/symbol correspondence) regardless of how the groups are calculated and it may be that this test needs to be re-examined. Carroll (1990 p. 13) found that scores in the phonetic discrimination task of the MLAT also was negatively skewed, suggesting that the test was too easy. As LLAMA_E is based on that section of the MLAT, this result should not be surprising.

RQ3: Age

The third research question was looking at the effect of age on LLAMA test scores. 157 subjects took LLAMA_B (vocabulary) and LLAMA_E (sound-symbol correspondence). Participants ranged in age from 10 to 75 and were divided into 5 groups (10-11 year olds (n=14), 18-21 year olds (n=66), 22-25 year olds (n=32), 26-35 year olds (n=18) and 36-75 year olds (n=27). This range of ages was included to be able to identify any effects of increasing age or cognitive decline on LLAMA test scores. Salthouse (2009) argues that cognitive decline begins in the 20s or 30s. Only LLAMA_B and E were chosen due to the young age of some of the subjects. We hypothesized that vocabulary acquisition (LLAMA_B) should not change according to age but that if there is a critical period for language learning (Lenneberg 1967) then we might expect the 10-11 year olds to outperform the adults on the sound/symbol correspondence (LLAMA_E) (e.g. see Munoz & Singleton 2011 and Piske, MacKay & Flege 2001 for reviews). The means and standard deviations are given in Table 3.

“Put table 3 about here please”

A one-way ANOVA showed no overall significant effects for age with vocabulary (LLAMA_B: *F*(4,152)=.282 *p*=.889) or with sound-symbol correspondence (LLAMA_E: *F*(4,152)=1.737 *p*=.145).However, the results in Table 3 show that the 10-11 year olds profile differently in comparison with the older groups. This can be seen in Figure 1. A post-hoc Games-Howell showed a significant difference on LLAMA_E between the 10-11 year olds and the 18-21s (*p*=.007) and the 36-75s (*p*=.014) but no differences with LLAMA_B (vocabulary).

“Put Figure 5 about here please”

This result suggests that there is no difference between adults and children (aged 10-11) in terms of their ability to acquire vocabulary as measured by LLAMA_B. This confirms our hypothesis for vocabulary and it possibly not surprising given that we acquire new words throughout our lives. However, contrary to our prediction that the younger learners would outperform the adults on sound/symbol correspondence, the younger learners actually performed significantly worse than the 18-21s and 36-75s. This may be due to the small number of participants in the 10-11s group (n=14) but as the task was not designed for children, it may be that this age group found the task too difficult and therefore these tests should be used with caution with younger learners. However, these results may support the idea that aptitude may be a factor for younger learners (Abrahamson & Hyltenstam 2008; Muñóz’s 2014 contra DeKeyser 2000).

RQ4: Formal education qualifications

The fourth research question investigated the role of formal education in LLAMA test scores. 135 participants over the age of 18 took all 4 LLAMA tests. In the background questionnaire, participants were asked for their highest formal qualification. These were then grouped into four categories. Group 1 had the lowest qualifications. These were the equivalent of leaving school aged 16 at the end of compulsory schooling (n=10). Group 2 had the equivalent of A-levels or qualifications that marked the end of secondary schooling (n=64). Group 3 had an undergraduate degree (n=40) and Group 4 had obtained a postgraduate qualification (n=21).We hypothesise that as, according to Grañena (2013a), three of the LLAMA tests seem to relate to more explicit learning/ analytical skills measures (LLAMA_B, LLAMA_E and LLAMA_F) that these scores may be affected by education level. Namely, that participants with higher formal education levels would perform better on the analytical skills required in the tests (following Oxford’s (1990) argument outlined above. The means and standard deviations are given in Table 4.

“Put table 4 about here please”

A one-way ANOVA showed a significant effect for highest formal qualification for Vocabulary (LLAMA_B) *F*(3,131)=3.413 *p*=.019, Sound/symbol correspondence (LLAMA_E) *F*(3,131)=7.684 *p*<.001 and grammar inferencing (LLAMA_F) *F*(3,131)=4.724 *p*=.004. Sound recognition (LLAMA_D) did not reach statistical significance, *F*(3,131)=2.439 *p*=.067. The postgraduate group (group 4) consistently outperformed the other groups (see Table 4). Group 2 (end of secondary school group) outperformed both

Group 1 (end of compulsory schooling) and Group 3 (undergraduate degree). Groups 1 & 3 did not perform differently from each other. However, it should be noted that Group 2 (end of secondary school group) was mainly comprised of students studying for their undergraduate degree. It is possible that some of the people who said they had an undergraduate degree (Group 3) misread the question on the background questionnaire and were actually studying for their degree. These subjects should have therefore been in group 2. This is merely speculative given our impressions of the participants at the larger group sessions.

The results suggest that having studied formally to postgraduate level significantly improves your aptitude test results as measured by the three components of LLAMA that correspond to the MLAT tests, i.e. LLAMA_B (vocabulary), LLAMA_E (sound/symbol) and LLAMA_F (grammatical inferencing). Formal education level does not appear to have an effect on LLAMA_D (sound recognition). Meara (2005) added LLAMA_D to the other components as a measure of implicit learning. Grañena (2013a) found that LLAMA_D measured something different to the other LLAMA tests, as discussed previously. The finding that formal education can affect your aptitude for the sub-tests that measure more explicit learning is perhaps not surprising given Ceci’s (1991) review findings that education level can influence IQ scores. This suggests that using the LLAMA tests in groups with mixed educational levels may need to be interpreted with caution.

RQ5: Logic Puzzles

The fifth research question addressed whether playing logic puzzles/games helped LLAMA test scores. Following on from research question 4 on education levels, we wanted to investigate whether the education effect shown may actually be due to developing analytical skills through playing of logic puzzles rather than through formal education. We hypothesized that playing logic puzzles may help with the more analytical measures of LLAMA_B (vocabulary), LLAMA_E (sound-symbol correspondence) and LLAMA_F (grammatical inferencing test). Participants were asked if they did logic puzzles, e.g. Sudoku, Brain Training games and if so, how often. Few participants played on a regular basis so 135 participants over the age of 18, who had completed all the LLAMA tests at the standard times, were divided into two groups; no games played (n=60) and played games (n=75). The means and standard deviation results are given in Table 5. A t-test for each LLAMA test showed a significant effect for playing logic puzzles with LLAMA_E (sound/symbol correspondence) only: *t*(133)=-2.781 *p*=.006.

“Put table 5 about here please”

This result contradicted our hypothesis that playing logic puzzles would help with LLAMA_B (vocabulary) or LLAMA_F (grammatical inferencing). This may be due, in part, to the methodology. Participants were asked not to write anything down in LLAMA_F and some reported that this made the task more difficult. Participants were not allowed to write anything down to ensure that all participants did the task in the same way and avoid the confounding variable of whether some participants had written things down and others had not.[7] The result that playing logic puzzles improved LLAMA_E (sound/symbol correspondence) scores initially seems surprising as LLAMA_E is supposed to be a measure

[7] In a follow up study reported in Rogers et al (2015), we allowed participants to take notes for LLAMA_F (grammatical inferencing). An independent t-test (unequal variances assumed) between the data reported here (n=135, M=42.22, s.d = 28.355) and in the 2015 study (n=211, M=41.42, s.d. 26.284) showed no significant difference for whether participants were allowed to take notes or not (*t*(344) = .263, *p*>0,5)

of phonetic discrimination. However, the actual task does have a puzzle element to it in that the participants have to match sounds to symbols laid out in a grid. There is a pattern to the groupings on the grid and participants may have deduced this. This suggests that the finding for the effect of education in RQ4 is not solely due to the development of greater analytical skills through formal education but may be also be due to increased analytical skills as measured through the playing of logic related puzzles.

RQ6: Timings

The final research question investigated the timings of the ‘learning’ part of the LLAMA tests. The LLAMA tests comprise of an initial learning phase and then a testing/answering phase in which participants are tested on what they have just learnt. The default timings for the ‘learning’ part for LLAMA_B (vocabulary), LLAMA_D (sound recognition) and LLAMA_E (sound/symbol correspondence) are each 2 minutes in length. For LLAMA_F (grammatical inferencing) it is 5 minutes. Learners are not time constrained in the ‘answering’ part of the LLAMA tests. Timings were modified in two conditions. In the first, the timings were shortened by 1 minute each and in the second condition, the timings were lengthened by one minute each. LLAMA_D was not included as it involves a recording and this could not be modified. We hypothesized that participants would perform worse when the timings were shortened and better when the timings were increased. 98 participants in total took the 3 LLAMA tests. 32 participants took the shorter times, 33 the default times and 33 the longer times. The participants were all matched for gender, age, education level and whether they spoke another language (L2 status). As the results were not normally distributed, non-parametric statistics were performed. Kruskal-Wallis tests found no statistical significant difference between groups (Age: $\chi^2(2) = .688$, $p = .709$, Gender: χ^2 (2) =

.192, *p* = .908, Education: χ^2 (2) = .809, *p* = .667 and L2 Status: χ^2 (2) = .161, *p* = .922). Table 6 gives the mean, median and range for each group and Figure 6 shows the mean results in a graph.

“Put table 6 and figure 6 about here please”

In order to establish if changing the times affected the LLAMA scores, a Kruskal-Wallis test identified altering time-limits affected participant score in LLAMA_B (vocabulary) and LLAMA_E (sound/symbol) but not for LLAMA_F(grammar) (LLAMA_B χ2(2)=8.9, *p*= 0.11; LLAMA_E χ2(2)=11.1, *p*=.004; LLAMA_F χ2(2)=1.66, *p*= 0.44). For LLAMA_B (vocabulary) a Mann Whitney U test (p=.003) showed that there was a significant difference between the shorter time of 1 minute (Median=35) and the default time of 2 minutes (median=45). For LLAMA_E (sound/symbol), another Mann Whitney U test (p=.01) also showed that there was a significant difference between the shorter time of 1 minute (Median=50) and the default time of 2 minutes (median=70). However, for LLAMA_E (sound/symbol), a Mann Whitney U test (p=.002) also showed that there was a significant difference between the shorter time of 1 minute (Median=50) and the longer time of 3 minutes (median=80). Therefore, time-limit reductions had the strongest effect on LLAMA_B and LLAMA_E scores with the shorter times producing lower scores. Only on LLAMA_E did increasing the time give a statistically significant improved score. The lack of change in the LLAMA_F scores may, in part, be due to many participants finishing early (i.e. before the default time of 5 minutes was over) so altering the times did not impact on their performance. As previously mentioned under RQ5, this may have been because we did not allow the participants to write anything down for this task.

As the participants were matched according to gender, age, education level and L2 status (see previous), we conducted a Quade's rank ANCOVA (1967) to test if participant attributes for gender, education level and L2 status acted as covariates to time-limit effects on LLAMA scores (Carlsson et al., 2014). Rank ANCOVA results suggested L2 Status influenced time-limit effects on LLAMA_B scores (*F*(2,95)= 4.751 *p*=.011) and in LLAMA_E scores (*F*(2,95)= 6.196 *p*=.003). Education also influenced time-limit effects on LLAMA_E scores (*F*(2,95)=7.825 *p*=.001). Results identified no attribute influence on time-limit effects in LLAMA_F (Gender: *F*(2,95)=.756 *p*=.472; Education: *F*(2,95)=.925 *p*=.4; L2 Status *F*(2,95)=.76 *p*=.471). Therefore, participant attributes influenced the effect of changed time-limits in LLAMA_B and LLAMA_E but could not influence time effects in LLAMA_F.

A Kruskal-Wallis test examined if attributes influenced the reduced or increased time-limit scores. Results showed that monolingual scores were more affected by the shorter time-limits (LLAMA_B: *p*=.018) and (LLAMA_E: *p*=.003) than bilingual & multilingual scores (LLAMA_B: *p*=.645; LLAMA_E: *p*=.730). In terms of education level, Undergraduate & Graduates were most affected by the shorter time-limits for LLAMA_E (*p*=.005), whereas participants with GCSE's & A-levels were most affected by shortened time-limits in LLAMA_B (p=.057). In relation to gender, males were more affected by the shorter time-limits (LLAMA_B: p=.022; LLAMA_E: p=.004) than females (LLAMA_B: p=.355; LLAMA_E p=.478).

We initially hypothesised that participants would perform worse when the timings were shortened by one minute. This hypothesis was confirmed for LLAMA_B (vocabulary) and LLAMA_E (sound/symbol). There was no statistically significant difference for LLAMA_F (grammar) although the results are suggestive (median=45 for the shorter time and median=60 for the default time). This supports similar findings by Partchev et al (2012) of lower scores with shorter times on the GRE. We also hypothesised that participants would perform better when allowed extra time. This hypothesis was only partially confirmed. Students performed statistically significantly better on LLAMA_E (sound/symbol correspondence) only. In fact participants scored lower on the increased times in LLAMA_F (median = 60 on default time and median = 50 on increased time). We acknowledged previously that some participants finished early with the LLAMA_F learning sequence and were explicitly encouraged to continue learning until time had elapsed. This positions LLAMA_F as a power-test and duplicates Wild et al.'s (1982) finding that extra time did not benefit participants as test difficulty prevented score increase. LLAMA_F default time runs the risk of participant fatigue and demotivation, especially as it appears last in the LLAMA test sequence. A reduced time could benefit both individual score and overall LLAMA test efficiency.

In conclusion for this research question, it appears that the default times of 2 minutes for LLAMA_B (vocabulary) and LLAMA_E (sound/symbol correspondence) are optimal. We suggest that LLAMA_F could be shortened without a significant effect on results although we acknowledge that our findings may have been influenced by the prohibition on participants writing notes during this test (however, see comments in footnote 8).

Overall conclusion

We set out to examine in a series of research training projects if the LLAMA tests (Meara 2005) were affected by a number of independent variables, including gender, L1 language alphabet/script, age, formal education, playing of logic games and altering the timings of the tests. The first two research questions mirrored those by Grañena (2013a) and we confirmed that there was no difference between the overall performances according to gender. The language neutrality question was only partially confirmed as our non-Roman script group outperformed the other groups on LLAMA_E (sound/symbol). We found significant effects for formal education (educated to postgraduate level) on all tests and of playing logic puzzles on LLAMA_E. The default timings appear to be optimal for LLAMA_B & E but could be shortened for LLAMA_F. We also found certain trends in the data that did not quite reach statistical significance: Participants tended to perform highly on LLAMA_E throughout and it may be that it does not discriminate well (see Carroll, 1990), and younger learners profiled differently than adults on LLAMA_B & E as they scored higher on LLAMA_B (vocabulary) than LLAMA_E (sound/symbol) whereas the adults had the opposite findings.

Limitations

There are several important limitations to this study. As always with research training projects, there are a number of the limitations in these studies which only become apparent after the event and in these cases were largely due to the short time-frame for data collection by the students. There is an over-dominance of monolingual undergraduate participants in comparison to other groups. Some of the sub-groups were quite small, for example the 10-11 year old group consisted of 14 participants and the non-Roman script group consisted of 17. No comparison measure were conducted, e.g. MLAT, IQ or working memory. These are all areas that we hope to address in the future. There are two other more general caveats to these results. Firstly, statistically-speaking the large standard deviations in some of the groups make finding a significant difference less likely. These large standard deviations are partly due to the small sample sizes for some groups and future replications should address this with larger samples. However, this leads to the more difficult conceptual point that in order to show that the LLAMA tests work in the same way for different groups, we do not usually want to find significant differences. In effect we are trying to ‘prove’ the null hypothesis. This is a difficult issue to address and is a problem for many validation studies of this kind. The underlying philosophy of our research training is that students need to learn to be critical of the tools they use, and do not accept things at face value. At the outset of these problems, they were told that the object of the exercise was to examine the performance of the LLAMA tests in detail and “test them to destruction” in ways that we had not examined them in the past. The fact that the LLAMA tests appear to have survived this critical probing is pleasing, but rather surprising.

References:

Abrahamsson, N., & Hyltenstam, K. 2008. The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*(4), 481-509.

Anufryk, V. 2011. *Intonational variation and pronunciation aptitude*. Doctoral Thesis. Stuttgart.

Birdsong, D., & Molis, M. 2001. On the evidence for maturational constraints in second-language acquisition. *Journal of memory and language*, *44*(2), 235-249.

Carlsson, M. O., Zou, K. H., Yu, C. R., Liu, K., & Sun, F. W. 2014. A comparison of nonparametric and parametric methods to adjust for baseline measures. *Contemporary clinical trials*, *37*(2), 225-233.

Carroll, J. B. 1981. Twenty-five years of research on foreign language aptitude. *Individual differences and universals in language learning aptitude*, 83-118.

Carroll, J. B. 1990. Cognitive abilities in foreign language aptitude: Then and now. *Language aptitude reconsidered*, 11-29.

Carroll, J. B., & Sapon, S. M. 1959. *Modern language aptitude test.* San Antonio, Texas: The Psychological Corporation.

Carroll, J. B., & Sapon, S. M. 1967. *Modern Language Aptitude Test-Elementary Manual*. San Antonio, Texas: The Psychological Corporation.

Ceci, S. J. 1991. How Much Does Schooling Influence General Intelligence and Its Cognitive Components? A Reassessment of the Evidence. *Developmental Psychology, 27*(5), 703-722.

De Bot, K. 2013. Circadian rhythms and second language development. *International Journal of Bilingualism*, online access. DOI: 1367006913489201.

DeKeyser, R. 2000. The robustness of critical period effects in second language acquisition. *Studies in second language acquisition*, *22*(04), 499-533.

DeKeyser, R. 2008. 11 Implicit and Explicit Learning. In .Doughty, C., & Long, M. (eds). *The handbook of second language acquisition.* Blackwell handbooks in linguistics). Malden, MA: Blackwell Pub.

Dörnyei, Z. 2010. “The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective”. In E. Macaro (Eds.), *Continuum companion to second language acquisition*, 247-267. London: Continuum.

Dörnyei, Z., & Ushioda, E. (eds.). 2009. *Motivation, language identity and the L2 self*. Multilingual Matters.

Eisenstein, M. 1980. Childhood bilingualism and adult language learning aptitude. *Revue Internationale De Psychologie Appliquee/International Review of Applied Psychology, 29*(1-2), 159-172.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. 1999. Age constraints on second-language acquisition. *Journal of memory and language*, *41*(1), 78-104.

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.

Grañena, G. 2012. *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment.* Doctoral thesis: University of Maryland.

Grañena, G. 2013a. "Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test" In G. Grañena, & M. Long (eds.) *Sensitive periods, language aptitude, and ultimate L2 attainment*, 105-129 Amsterdam: John Benjamins Publishing.

Grañena, G. 2013b. Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*(4), 665-703

Grañena, G., & Long, M. 2013a. Age of onset, length of residence, language aptitude, and ultimate l2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311-343.

Grañena, G., & Long, M. (eds.). 2013b. *Sensitive periods, language aptitude, and ultimate L2 attainment* (Vol. 35). Amsterdam: John Benjamins Publishing.

Green, D., & Meara, P. 1987. The effects of script on visual search. *Second Language Research, 3*(2), 102-113.

Grigornko, 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.

Hamada, M., & Koda, K. 2008. Influence of first language orthographic experience on second language decoding and word learning. *Language Learning*, *58*(1), 1-31.

Johnson, J. S., & Newport, E. L. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology*, *21*(1), 60-99.

Kormos, J. 2013. New conceptualizations of language aptitude in second language attainment. *Sensitive periods, language aptitude, and ultimate L2 attainment*, *35*, 131.

Krashen, S. D. 1981. Aptitude and attitude in relation to second language acquisition and learning. *Individual differences and universals in language learning aptitude*, 155-175.

Larson-Hall, J., & Dewey, D. 2012. An examination of the effects of input, aptitude, and motivation on the language proficiency of missionaries learning Japanese as a second language. *Second language acquisition abroad: The LDS missionary experience*, 45, 51.

Lenneberg, E. H. 1967. *Biological foundations of language* (Vol. 68). New York: Wiley.

Lundell, F. F., & Sandgren, M. 2013. High-level proficiency in late L2 acquisition Relationships between collocational production. In Grañena, G., & Long, M. (eds.). *Sensitive periods, language aptitude, and ultimate L2 attainment*. Amsterdam: John Benjamins Publishing.

Manning, M. 1995. *Borland's official no-nonsense guide to Delphi 2*. Sams.

McLaughlin, B. 1990. The relationship between first and second languages: Language proficiency and language aptitude. *The development of second language proficiency*, 158-178.

Meara, P. 2005. LLAMA language aptitude tests: The manual. *Swansea: Lognostics*.

Meara, P. 2013 “Imaginary Words”. In Chapelle, C. (ed.) *The encyclopedia of Applied Linguistics*. Chichester: Wiley

Muñóz, 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.Studies in Honor of David Singleton*, 51-68. London: Springer International Publishing.

Muñoz, C., & Singleton, D. 2011. A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*(01), 1-35.

Nation, R., & McLaughlin, B. 1986. Novices and experts: An information processing approach to the “good language learner” problem. *Applied Psycholinguistics*, *7*(01), 41-55.

Nayak, N., Hansen, N., Krueger, N., & McLaughlin, B. 1990. Language-Learning Strategies in Monolingual and Multilingual Adults. *Language Learning*, *40*(2), 221-244.

Opitz, C. 2011. *First language attrition and second language acquisition in a second language environment* (Doctoral dissertation, Trinity College Dublin.).

Ortega, L. 2009. Foreign Language Aptitude. In *Understanding second language acquisition*. London: Hodder Education, 145-166.

Oxford, R. 1990. *Language Learning Strategies. What every teacher should know.* Boston: Heinle & Heinle

Parry, T. S., & Stansfield, C. W. 1990. *Language aptitude reconsidered* (Vol. 74). New Jersey: Prentice Hall.

Partchev, I., De Boeck, P., & Steyer, R. 2013. How much power and speed is measured in this test? *Assessment*, *20*(2), 242-252.

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.

Pimsleur, P. 1966. *Pimsleur language aptitude battery*. New York: Harcourt, Brace & World.

Piske, T., MacKay, I. R., & Flege, J. E. 2001. Factors affecting degree of foreign accent in an L2: A review. *Journal of phonetics*, *29*(2), 191-215.

Quade, D. 1967. Rank analysis of covariance. *Journal of the American Statistical Association*, *62*(320), 1187-1200.

Robinson, P. 2001. Individual differences, cognitive abilities, aptitude complexes and learning conditions in second language acquisition. *Second language research*, *17*(4), 368-392.

Robinson, P. 2013. “Aptitude in Second Language Acquisition”. In Chapelle, C. (ed.) *The Encyclopedia of Applied Linguistics*. Chichester: Wiley

Rogers, V. 2015 “Testing the LLAMA aptitude tests: an overview”. *Presented at Language Research Centre seminar series, Swansea University.* Available at: http://viviennerogers.info/wp-uploads/2011/06/LRC2015_Rogers.pdf

Rogers, V., Barnett-Legh, T., Curry, C. & Davey, E. 2015 “Validating the LLAMA aptitude tests”. *Presented at EUROSLA 2015, York.* Available at: http://viviennerogers.info/wp-uploads/2011/06/EUROSLA2015_Rogers.pdf

Sáfár, A., & Kormos, J. 2008. “Revisiting problems with foreign language aptitude”. In P., Jordans & L., Roberts (eds.), *International Review of Applied Linguistics in Language Teaching*, *46*(2), 113-136.

Salthouse, T. A. 2009. When does age-related cognitive decline begin? *Neurobiology of aging, 30*(4), 507-514.

Santana Rollán, M. E. 2013. *La aptitud lingüística en estudiantes ciegos*. Doctoral dissertation, Universidad Complutense de Madrid.

Sawyer, M. 1992. Language Aptitude and Language Experience: Are They Related? *The language programs of the International University of Japan: working papers*, *3,* 27-45.

Service, E. & Kohonen, V. 1995. Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*(02), 155-172.

Service, E. 1992. Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology*, *45*(1), 21-50.

Skehan, P. 2002. Theorising and updating aptitude. *Individual differences and instructed language learning*, *2*, 69-94.

Smeds, H. 2012. Perceptual compensation in blind second language (L2) learners Paper presented at the Nordic conference for bilingualism, University of Copenhagen, Denmark. Retrieved from: http://www.biling.su.se/om-oss/kontakt/medarbetare/helena-smeds-1.91491

Smeds, H. 2013. *Perceptual compensation in blind second language (L2) learners*. PhD thesis: Stockholm University

Snow, R. E. 1992. Aptitude theory: Yesterday, today, and tomorrow. *Educational psychologist*, *27*(1), 5-32.

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. *Classical Journal*, 165-184.

Sparks, R., & Ganschow, L. 2001. Aptitude for learning a foreign language. *Annual Review of Applied Linguistics, 21*, 90-111.

Speciale, G., Ellis, N. C., & Bywater, T. 2004. Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied psycholinguistics*, *25*(02), 293-321.

Sternberg, R. J. 2002. "The theory of successful intelligence and its implications of language aptitude testing". In P. Robinson (ed.), *Individual differences and instructed language learning.* 13-44. Amsterdam: John Benjamins Publishing.

Thompson, A. S. 2013. The interface of language aptitude and multilingualism: Reconsidering the bilingual/multilingual dichotomy. *The Modern Language Journal*, 97, 685–701.

Wesche, M. 1981. Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In. K. Diller, (ed.) *Individual differences and universals in language learning aptitude*, 119-139.

Wild, C. L., Durso, R., & Rubin, D. B. 1982. Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement, 19*(1), 19-28.

Wong, W., & Pyun, D. O. 2012. The Effects of Sentence Writing on Second Language French and Korean Lexical Retention. *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *68*(2), 164-189.

Xiang, H., Dediu, D., Roberts, L., Oort, E. V., Norris, D. G., & Hagoort, P. 2012. The structural connectivity underpinning language aptitude, working memory, and IQ in the perisylvian language network. *Language learning*,*62*(s2), 110-130.

Yalcin, S. 2012. *Individual Differences and the Learning of Two Grammatical Features with Turkish Learners of English* (Doctoral dissertation, University of Toronto).

Yilmaz, Y. 2013. Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics*, *34*(3), 344-368.

Tables:

Table 1: RQ1: LLAMA results grouped by gender

	LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F
Male	36.80	21.71	52.18	36.11

(n=63)	(24.708)	(18.217)	(35.756)	(29.391)
Female (n=72)	38.40 (25.859)	23.44 (19.182)	58.27 (34.438)	34.10 (28.285)

Table 2: RQ2 LLAMA results grouped by L1 alphabet/ script

	LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F
English (n=99)	38.28 (25.013)	21.66 (18.133)	50.51 (35.623)	33.71 (28.962)
Non-Roman script (n=17)	44.71 (25.488)	28.53 (20.673)	67.06 (21.727)	35.88 (26.706)
Roman script (n=18)	38.70 (27.347)	21.19 (19.577)	69.61 (37.017)	39.58 (29.802)

Table 3: RQ3 LLAMA results grouped by age

	10-11 (n=14)	18.21 (n=66)	22-25 (n=32)	26-35 (n=18)	36-75 (n=27)
LLAMA_B	42.50	39.16	35.27	38.91	40.56

	(17.623)	(26.685)	(28.158)	(23.769)	(30.551)
LLAMA_E	31.43 (19.158)	56.01 (35.443)	52.67 (36.556)	56.72 (35.562)	57.78 (30.551)

Table 4: LLAMA results grouped by Highest formal qualifications achieved

	LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F
Aged 16/ end	34.50	16.00	45.00	40.00

of compulsory schooling (n=10)	(18.174)	(13.499)	(38.658)	(18.856)
Aged 18/ secondary school (n=64)	41.66 (25.120)	23.84 (18.345)	59.56 (31.237)	39.08 (29.934)
Undergraduate degree (n=40)	27.93 (25.819)	18.37 (19.682)	39.03 (38.220)	21.71 (25.892)
Postgraduate degree (n=21)	45.48 (37.66)	30.24 (17.852)	79.05 (20.225)	45.71 (26.376)

Table 5: LLAMA results grouped by playing logic games

	LLAMA_B	LLAMA_D	LLAMA_E	LLAMA_F
No games (n=60)	35.75 (23.780)	21.88 (18.494)	46.28 (32.939)	30.58 (21.177)
Played (n=75)	39.98 (26.287)	23.24 (18.944)	62.75 (35.194)	38.60 (29.587)

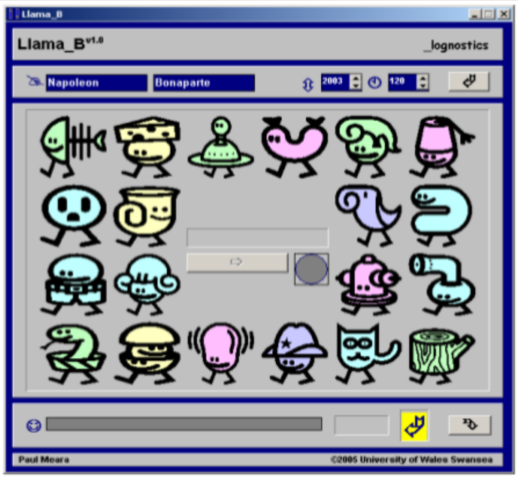
Table 6: LLAMA results grouped by altered timings (1 min shorter, default, 1 min longer)

	LLAMA_B	LLAMA_E	LLAMA_F

	Shorter	Default	Longer	Shorter	Default	Longer	Shorter	Default	Longer
Mean	38.13	50.30	53.94	46.09	65.76	70.30	42.97	51.82	48.33
Median	35	45	60	50	70	80	45	60	50
Range	10-90	10-100	10-100	0-100	10-100	10-100	0-90	0-100	0-100

Figures:

Figure 1: LLAMA_B interface



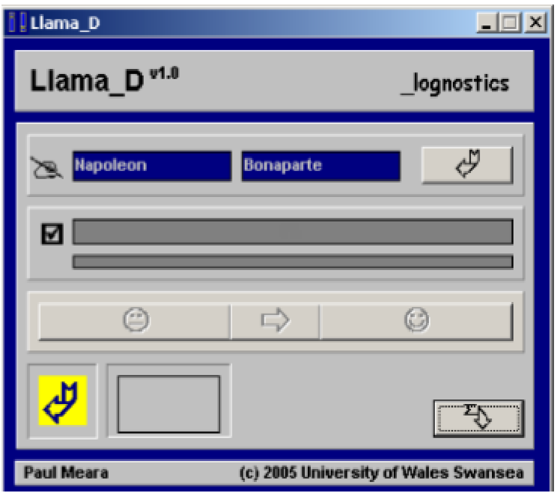
Napoleon

Bonaparte

2003

120

Figure 2: LLAMA_D interface

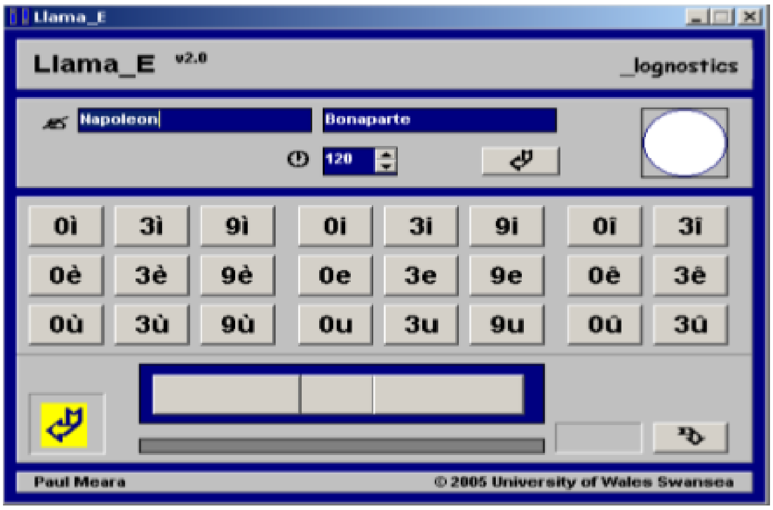


Napoleon

Bonaparte

☑

Figure 3: LLAMA_E interface

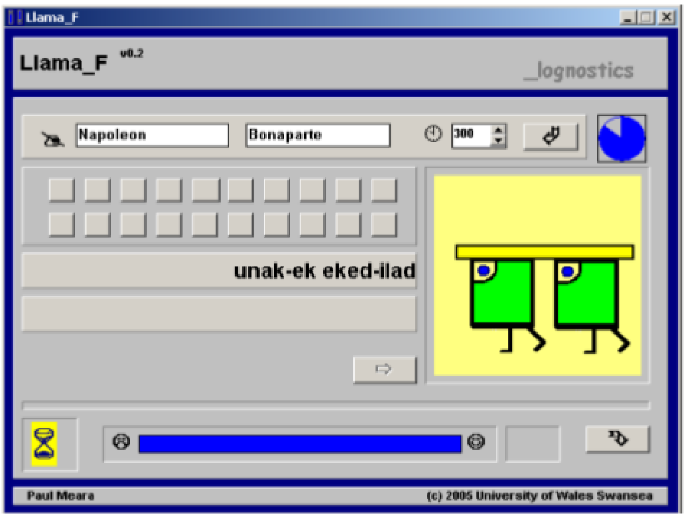


Napoleon

Bonaparte

120

Figure 4: LLAMA_F interface



Napoleon

Bonaparte

300

unak-ek eked-ilad

Figure 5: LLAMA scores by age groups

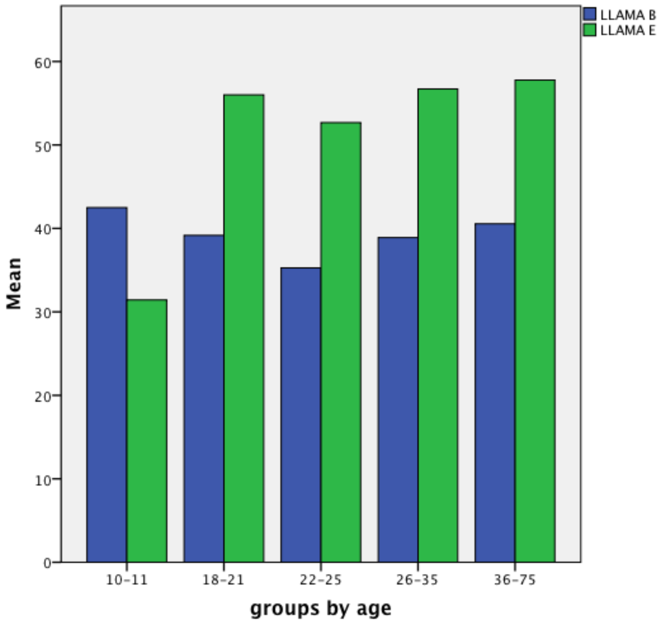
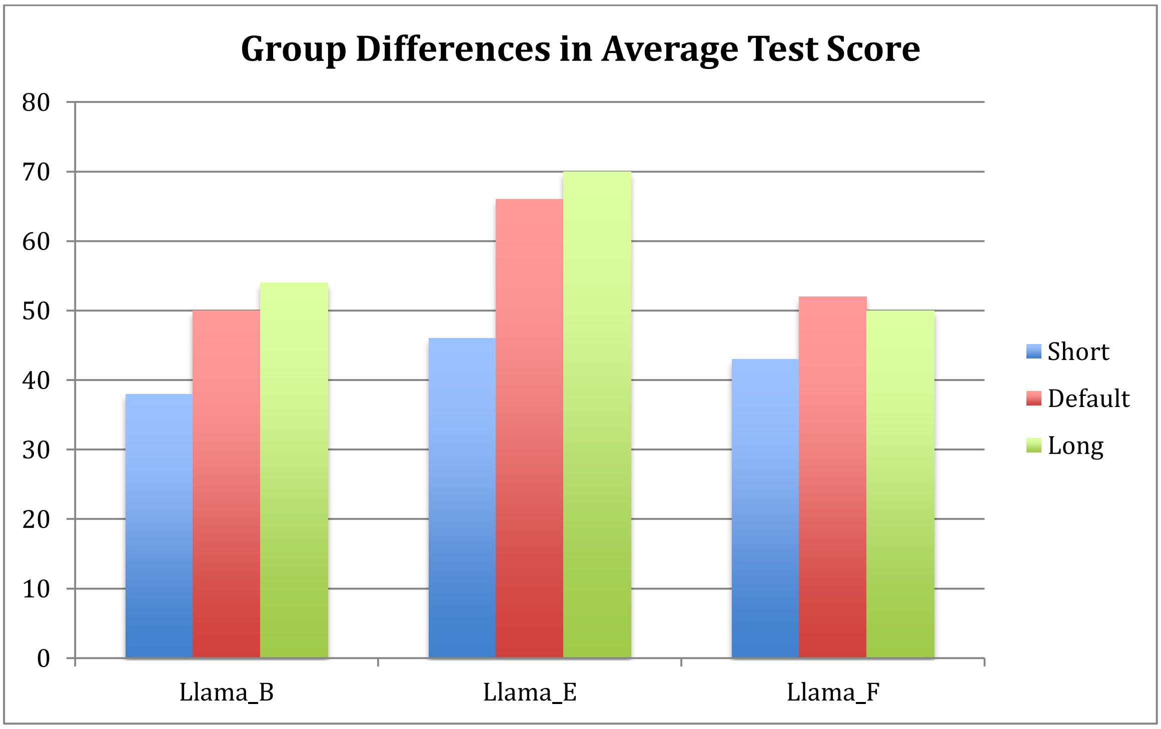


Figure 6: LLAMA_B, E & F results with altered timings.



Appendices:

Appendix 1: Consent form

Appendix 2: Instructions for participants

Appendix 3: Background questionnaire (this was administered online).

Appendix 1: Research Project Consent Form

Area of Research

Investigating language learning aptitude.

Study's Purpose

The purpose of our study is to examine Paul Meara's LLAMA tests along the parameters of:

- The effect of language background
- The effect of educational background
- The effect of differing test times

Procedure

This study takes approximately half hour to complete. It will require you to complete this consent form, four LLAMA tests, and to fill in a short questionnaire.

Data Collection

Your results will be used confidentially throughout the study, and may be stored anonymously by Swansea University for further research. Participant's names will not be used in the report.

The research supervisor is Dr. Vivienne Rogers, along with undergraduate researchers:
Rachel Aspinall, Louise Fallon, Thomas Goss, Emily Keey and Rosa Thomas

Results

If any participants wish to learn the results of this study, they can receive this information by contacting Dr. Vivienne Rogers by email: V.E.Rogers@swansea.ac.uk

Agreement

By signing this form you are stating that you have read through and understood the information provided. It also shows you consent to taking part in the study and to your results being used in the study. You are free to not answer any questions or withdraw from the study at any time. If you have any questions please ask.

_________________________ __________________

Participant Date

_________________________ __________________

Investigator/Witness Date

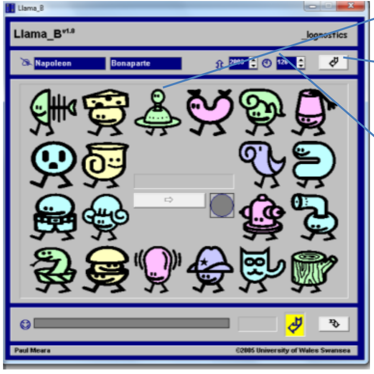
A copy of this consent form has been given to you to keep for your records and reference

Appendix 2: Instructions for participants

These tests are designed to assess your aptitude for learning a foreign language.

LLAMA B: A vocabulary learning task.

Open the program and this screen will appear.



Napoleon Bonaparte

Enter your name in these boxes.

Begin the memorising here.

Adjust the time here.

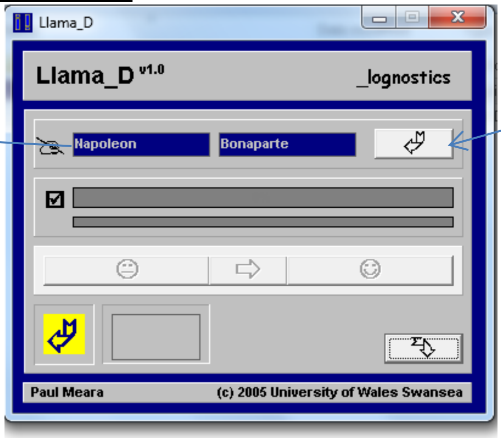
Your task is memorise the twenty objects in 2 minutes, you can click on the objects as many times as you like.

The clock in the centre displays your time. Once your time is complete the buttons will be deactivated and a beeping sound will be triggered.

Once this stage is complete you must begin the testing stage by pressing the ⇨ button.

The name of the object will be displayed in the central panel and you must find the matching object. A ding sound is activated when a correct answer is given and a bleep for a wrong answer.

LLAMA D: A sound recognition task

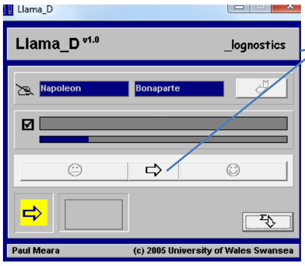


Napoleon Bonaparte

Enter your name in these boxes.

Click this symbol for the recordings

You must listen to the sound recording and it will play with a number of made up words. Your task is to learn and memorise as many of these words as possible.



Napoleon Bonaparte

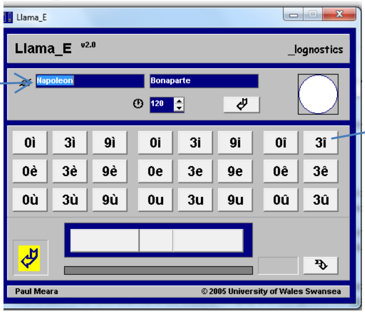
Click the arrow in the middle of the screen to begin the test.

The test will play one word and then give you a pause.

If the word you hear was in the recordings, you must click the smiley face (on the right)
You must click the sad face (on the left) if the word did not feature.
You must click the centre arrow to continue to the next word.

LLAMA E: A sound~symbol correspondence task

Enter your name in these boxes.

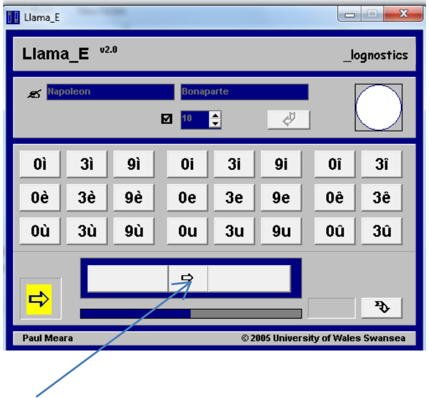


Napoleon Bonaparte 120

0ì	3ì	9ì	0i	3i	9i	0î	3î
0è	3è	9è	0e	3e	9e	0ê	3ê
0ù	3ù	9ù	0u	3u	9u	0û	3û

Click on a symbol as many times as you like.

You have 2 minutes to click the symbols and memorise the corresponding sounds.

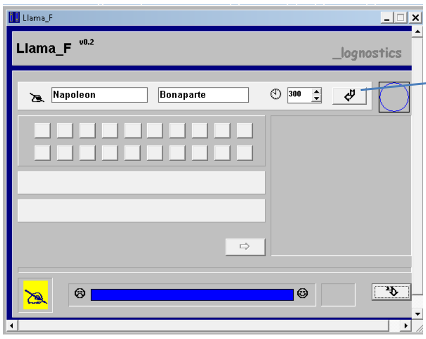


Napoleon Bonaparte ☑ 10

0ì	3ì	9ì	0i	3i	9i	0î	3î
0è	3è	9è	0e	3e	9e	0ê	3ê
0ù	3ù	9ù	0u	3u	9u	0û	3û

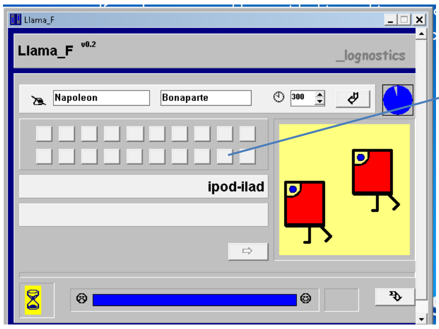
Next click the small arrow in the centre.
When you click the arrow a sound will play and two symbols will appear.
Click the symbol that corresponds with the sound.

LLAMA F



Napoleon Bonaparte 300

Click the arrow to start memorising



Napoleon

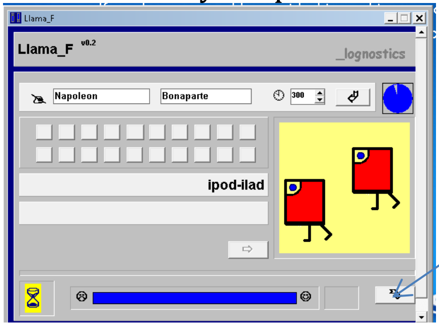
Bonaparte

300

ipod-ilad

Click each box to reveal a phrase and a corresponding picture.

Your task is to recognise patterns shared by the phrase and image.



Napoleon

Bonaparte

300

ipod-ilad

Click this arrow in order to match the image to the appropriate phrase.

Near the end of the test you will be given pictures you have not seen before, use any rules you may have noticed to help you answer. This part of the test is not timed.