Chapter 11: LLAMA_B v2.0

Introduction

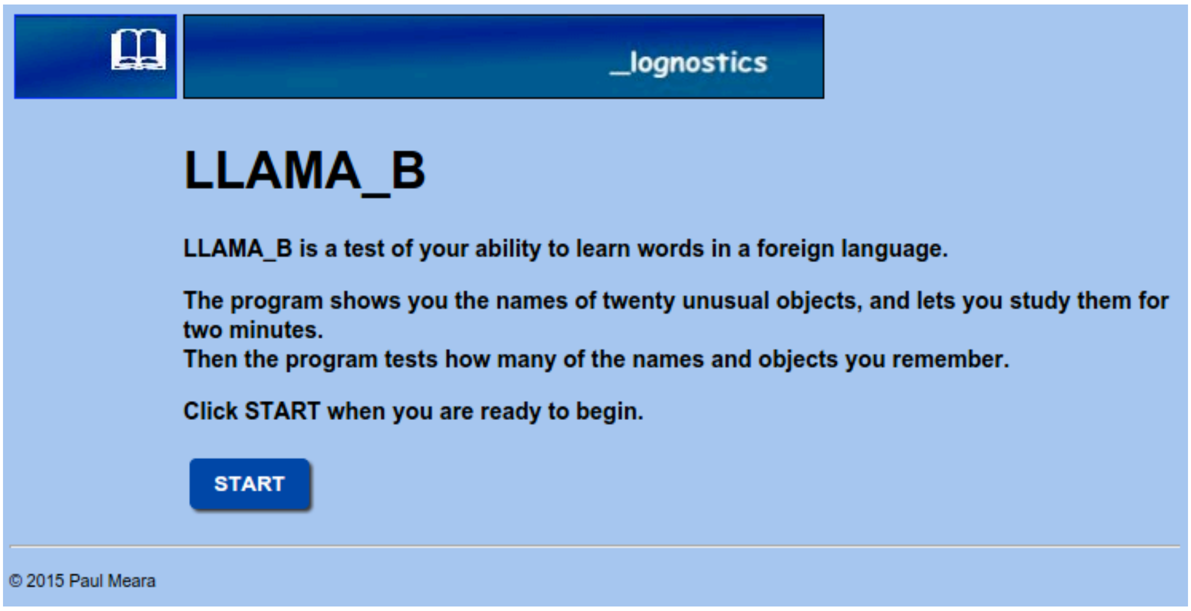
LLAMA_B is one of a suite of programs dealing with Language Aptitude. It deals specifically with vocabulary acquisition - learning names for things. The programs were developed as part of a research training program for Masters Students at Swansea University. The main motivation behind the programs was not in fact to develop a language aptitude test, but rather to provide a research environment in which students could naturally develop a critical attitude towards the data collection tools that they were using. In our experience, this critical attitude towards tools was rather rare in the students. Most of them would happily download a program from the Web, and use it uncritically in their own dissertation work, even when it was obvious to us that the tool was of dubious origins, not at all reliable, or just completely inappropriate. So we gave the students the LLAMA tests, and an instructed them to find out if they were any good. There were four tests in all, somewhat unimaginatively called LLAMA_B, LLAMA_C, LLAMA_D and LLAMA_F.

Each program was loosely based on one of the parts of Carroll and Sapon's *Modern Language Aptitude Test* (Carroll and Sapon, 1959): LLAMA_D was a sound recognition task that owed something to the work of Elizabeth Service (e.g. Service, 1992; Service & Kohonen, 1995). LLAMA_E was a sound-symbol correspondence task that tested test-takers' ability to associate familiar sounds to unfamiliar symbols. LLAMA_F was an innovative test that exposed users to an invented language, and asked them to work out what its grammatical features were. LLAMA_B covered vocabulary acquisition skills. This feature of language aptitude was considered to be very important by Carroll and Sapon, contributing a significant amount of variance to overall language learning performance. LLAMA_B was similar in content to the vocabulary learning section of MLAT, but it used a more modern interface. Carroll and Sapon had asked their testees to learn a small set of unfamiliar L2 words, but their format merely presented the test-takers with a list of English and Kurdish word pairs. We felt that this was an unattractive and demotivating format, which encouraged a rather rigid approach to vocabulary acquisition. LLAMA_B was designed with a more attractive interface, and it used pictures rather than words as stimuli. The pictures were chosen so that it was not obvious what the pictures showed, and this made the task rather more fluid and more challenging than Carroll and Sapon's list learning task. This format had the additional advantage that it allowed the test-takers a lot of freedom in how they approached the task, and did not force them to follow a predetermined learning method.

Using LLAMA_B

You can access the program at http://www.lognostics.co.uk/tools/LLAMA_B/LLAMA_B.htm

1. Click **START** to begin the program. The LLAMA_B opening page looks like Figure 11.1.



LLAMA_B

LLAMA_B is a test of your ability to learn words in a foreign language.

The program shows you the names of twenty unusual objects, and lets you study them for two minutes.
Then the program tests how many of the names and objects you remember.

Click START when you are ready to begin.

© 2015 Paul Meara

Figure 11.1 LLAMA_B opening screen

2. When you click the START button, LLAMA_B takes you to a display page which contains twenty pictures. Each picture is an unusual object. You can find out the name of the object by moving your mouse over it. The display page looks like Figure 11.2.

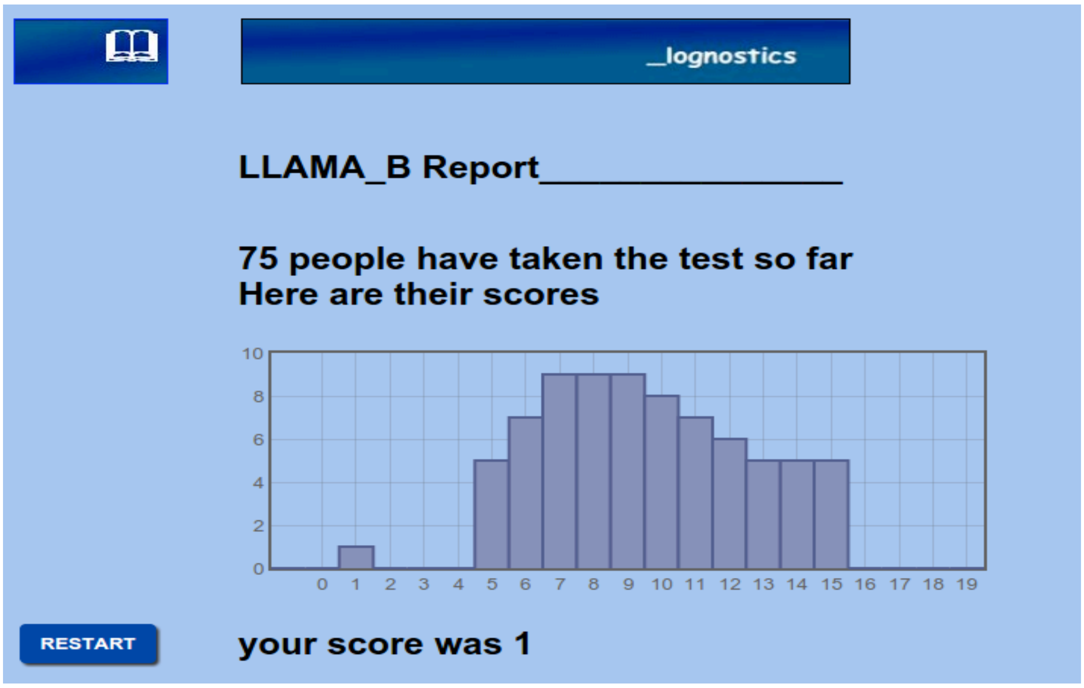
The program allows you two minutes to study this material and to learn the names of the twenty objects. You can do this in any way you like.

At the end of two minutes, the pictures will disappear. Click the CONTINUE button to go on to the LLAMA_B testing phase.

3. The testing screen looks very similar to the learning screen, but the objects have been moved around. The program presents you with each of the twenty object names in turn, and for each name, asks you to click on the picture of the relevant object.

This part of the program is not timed. When you have completed all twenty naming tasks, the program will automatically transfer you to the LLAMA_B report page.

Figure 11.2 The LLAMA_B display page
4. The Report page is shown in Figure 11.3



LLAMA_B Report

75 people have taken the test so far
Here are their scores

your score was 1

Figu
re 11.3 The LLAMA_B report page

LLAMA B keeps a record of all the scores people get on the test, and prints out a graph showing how these scores are distributed. This means that the program will build up over a time a fairly comprehensive picture of what "normal" performance on the test is like. This is a significant improvement on the impressionistic reporting that was incorporated into the early versions of the test.

LLAMA_B also reports your own score on the test. This figure ranges from 0-20. Random guessing should produce a score of 1.

The technical bits
The illustrations used in the LLAMA_B test are part of the Microsoft clip-art collection.
The names of the objects are names for naturally occurring plants and animals in Mixtec, an Oto-Manguean language spoken in Central America.

Background Reading

Rogers, V. (2015) A brief evaluation of the LLAMA_B Test.

The LLAMA tests (Meara, 2005) were developed as a suite of research training tools aimed at students doing small-scale reserch projects. The LLAMA suite is loosely based on the work of Carroll and Sapon (1959), but it differs from that early work in taking advantage of the possibilities offered by computers to collect and analyse data. The LLAMA tests were not originally intended as standard Language Aptitude tests, although a surprisingly large number of people have used them in this way. The intention behind developing these tests was to provide a space where final year undergraduate students and Masters level students doing small-scale independent research could develop their research skills by thinking critically about the strengths and weaknesses of the research tools that they adopt for this work. Surprisingly, perhaps, the tests have been used by a number of researchers in high stakes contexts that are not training exercises. Given that the LLAMA tests have never been properly validated, this is something of a problem, and particularly so as some of these users appear not to have heeded the warnings in the LLAMA manual about the tentative, exploratory nature of the tests, and the need for the results they generate to be treated with appropriate caution.

One unanticipated side-effect of people using the LLAMA tests outside the context they were intended for is that the student projects evaluating the LLAMA tests have become considerably more significant than is normally the case for projects of this sort. For this reason, we are summarising in this paper a set of undergraduate projects carried out in 2013 and 2014. The work reported was developed by Rachel Aspinall, Thomas Barnett-Leigh, Clare Curry, Emma Davie, Louise Fallon, Tom Goss, Emily Keey and Rosa Thomas in their final year projects. The work reported here is concerned with the LLAMA_B test. We hope that this work will be of interest to users of the LLAMA_B test.

LLAMA_B is the LLAMA equivalent of MLAT's Paired Associate Learning Task, and deals with test-takers' ability to learn new L2 words; that is, LLAMA_B attempts to assess how good a testee is at learning new vocabulary. It does this by presenting test-taker with pictures of twenty unusual objects and the names for these objects in an unfamiliar language. The test-takers have two minutes to work with this material, and to study the new vocabulary in any way that they want to. After two minutes, the program enters a test phase, where each of the new words are presented one at a time, and the test-taker has to identify the object that it refers to.

The students were asked to work with this test and to rigorously evaluate it. Between the two year groups, data from over 400 participants has been gathered for this test. Each student group project may collect data from only twenty or thirty subjects. However, because we teach the student how to properly archive their data, it is possible for us to interrogate the data in ways which the students did not envisage at the time it was collected. And this means that the LLAMA evaluation projects become increasingly interesting as the database increases in size.

The types of questions the students typically ask when they are left to themselves include:
1: Are the LLAMA tests language neutral?
2: Does previous instructed language learning experience influence LLAMA scores?3: Is there an effect of age or gender on LLAMA test scores?4: Does a person's general education level influence their LLAMA scores?
5: Do LLAMA scores change if we vary the time allowed for learning the names?

These questions will be addressed in this report. The first four questions are related to individual variables and their possible effect on the test scores. As you will see below, some of the 404 test takers did not provide all the necessary personal information for them to be included in several analyses we wanted to perform (e.g. on the role of the L1, age, etc.) and therefore had to be excluded from particular analyses. The exact numbers of participants involved in the analyses for each of our research questions are specified below. The last research question examines a test condition (the study time allowed to learn the words). We investigated this point with some additional participants who were given more or less time in the first phase of the test to learn the vocabulary that would be tested.

Exploring the Data

The effects of L1 background

One of the design features of the LLAMA_B test is that it uses picture stimuli and non-English words. This feature was a deliberate improvement on the vocabulary learning task in MLAT which used pairs of words in English and Kurdish. Obviously, the original MLAT tests assumed that the test takers would be L1 English speakers. Later on, a large number of MLAT variants were developed in order to cope with test takers whose L1 was not English (e.g. Stansfield et al., 2005; Suárez & Muñoz, 2011). For LLAMA_B translations should not be necessary, as the material is essentially language neutral. However, students naturally tend to ask whether the LLAMA tests really are language-neutral in the way they are intended to be. This question was first examined by the 2013-14 student group, and then re-examined in more detail by the 2014-15 group. Collapsing all the data together, a large number of participants have taken this test, and this has allowed us to look at the performance of several different L1 testee groups. This data is summarised in Table 1.

Table 1. Mean scores on the LLAMA_B test by L1 background.

L1 group	N	Mean	*sd*
Arabic	34	10.91	4.73

A one-way ANOVA showed an overall effect of language background [F(3, 387)=6.669, p<.001]. A post-hoc Games-Howell analysis assuming unequal variances found a statistically significant difference between the L1 English speakers and the Chinese and Arabic speakers, with the latter two groups outperforming the L1 English group (p<.05).

Clearly, there **is** an effect of L1 background on the way test-takers perform on the LLAMA_B test. However, we think it probable that the efffect found here is actually something rather different. We suspect that the most likely explanation of this unexpected result is that the L1 English speakers were mostly monolingual subjects with little or no foreign language learning experience, whereas the L1 Chinese and Arabic speakers were all following English language courses outside of their countries, and were to some extent pre-selected as good language learners. An obvious follow-up study here would be to identify a gtoup of L1 English speakers with extensive language learning experience, and compare this group with other L1 speakers with limited experience of learning foreign languages.

Generally speaking, there is considerable variation within the groups on this test – the standard deviations are about 50% of the mean score for all the groups. Another obvious follow up study would be to examine whether test-takers who score high on the LLAMA_B test tend to have have larger L1 vocabularies than test-takers with lower scores on the LLAMA_B test.

The effects of previous language learning experience

The results reported in the previous section suggest that language experience might be a factor which impacts on the LLAMA_B scores. We can address this question by interrogating our data in terms of the test-takers' language learning background. Specifically, the data allowed us to identify three groups of test-takers who differed in this respect: 222 of the test-takers were learning a second language; 51 considered themselves to be bilingual from childhood; and 130 were monolingual English speakers. We therefore re-analysed the LLAMA_B data in these terms, and the results are shown in Table 2.

Table 2. Mean scores on the LLAMA_B test by language learning experience.

Language experience	**N**	**Mean**	***sd***
L2 learners	222	10.27	4.76
Monolinguals	130	7.84	4.09
Bilinguals	51	7.27	3.97
Total	403		

A one-way ANOVA again showed an overall significant group effect [F(2,400)=17.209, p<.001]. A post-hoc Games-Howell test assuming unequal variances found statistically significant differences between the L2 learner group and the other two groups (p<.05), with the L2 learner group outperforming both of the other groups. This finding is not entirely surprising. The obvious explanation of the result is that second language learners will generally have had to learn new vocabulary explicitly, and will have developed their own strategies for doing this. Monolinguals, and childhood bilinguals, on the other hand, will have developed their vocabularies through implicit learning, rather than explicit learning, and may not have developed more formal vocabulary acquisition strategies.

The effects of age

The LLAMA tests were not originally intended for use with young children, but some of the work on language aptitude has been using the tests with adolescent language learners. This raises the question of whether the tests are age sensitive.

Alongside our older testees, we tested fifty 10-11 year olds, and the scores of these learners were compared with the scores of several groups of older learners. Results are reported in Table 3.

Table 3. Mean scores on the LLAMA_B test by age.

22-29	124	10.03	4.92
Total	395		

The results show that the young learner group has the lowest mean score of all the age groups. When a one-way ANOVA was conducted, significant differences were found between groups [F(4,390)=4.832, p=.001]. More specifically, the Games-Howell post-hoc test revealed that the young learners performed significantly worse than all the other groups, except for the 'above 45' group. That is, this young group performs significantly more poorly than adolescents and young adults, but the difference in mean scores with the oldest groups does not reach significance. The performance of the three groups comprising test takers aged fifteen plus is consistent and there are no significant differences between these groups. It should be noted, though, that all the standard deviations are again relatively large (about 50% of the mean group score), and this indicates that there are substantial within-group differences in this test.

The results confirm our view that the LLAMA_B tests should be treated with caution when it is used with children. The results may also suggest that aptitude changes with age, with general cognitive ability being a defining factor in young children, but less so in older subjects. In addition, although the oldest group ('above 45') is the least numerous (N=31), the findings could also indicate a variation in aptitude beyond 45, as the mean for this group does not significantly differ from that of the youngest group (10-11). However, there is not a significant difference between the mean of the oldest group and that of adolescents and younger adults, so further research on these age ranges is required to throw light on this issue.

The effects of gender

Grañena (2013) investigated whether gender had an effect on LLAMA test scores. She failed to find any gender effect in a study with 186 participants. This finding was replicated in our own data with an even larger participant group (N=404). These results are reported in Table 4. An independent-samples t-test did not show any significant differences in the scores due to gender [*t*(402)=.328, p=.743].

Table 4. Mean scores on the LLAMA_B test by gender

Gender of participants	N	Mean	*sd*
male	186	9.19	4.42
female	218	9.04	4.81
total	404		

The effects of educational level

A number of previous studies have suggested that social class or educational levels (e.g. Skehan, 1989) might affect language aptitude and our data also allows us to investigate this question. The descriptive results are presented in Table 5.

A one-way ANOVA found an overall significant effect of educational level [F(2,401)=11.668, p<.001]. A post hoc Games-Howell test indicated that there were statistically significant differences between the *No qualifications or GCSE* group and the other two groups (*A-level* and *graduate/postgraduate* groups), with the lowest qualification group performing significantly worse than the other two. There were no other significant differences in the data.

It should be noted that the wording of the question asked to the test-takers was "What is the highest qualification you have already achieved?". Therefore, current undergraduate students should have indicated that their highest qualification is an A-level or equivalent. We suspect that some of the participants did not read the question properly, and answered with the qualification that they were currently studying for. This means that the data in the graduate group may be unreliable. However, significant differences were found between 'below A-level' participants and 'A-level equivalent and beyond' participants.

Table 5. Mean scores on the LLAMA_B test by educational level

Highest qualification obtained	N	Mean	*sd*
No qualifications or GCSE equivalent	74	6.88	3.38

A-level equivalent	178	9.36	4.68
University grad. and/or postgrad.	152	9.90	4.77
Total	404		

Note 1: In the UK, GCSE examinations are usually taken at age 16, A-level examinations are usually taken at age 18. The legal school leaving age is 16. The first test taker group therefore consists of people who have completed compulsory education in the UK. About 50% of the population go on to take university degrees.

Individual variables as predictors of aptitude

The variables explored in the previous sections may probably affectt the aptitude scores to different extents. We carried out a multiple linear regression analysis (N=381) to assess the possible contribution of the test-takers' individual variables (predictor variables) to the LLAMA B scores (dependent variable). Our data did not violate the assumptions of normality, linearity and multicollinearity, but in some cases homogeneity of variance was not met. For this reason, we used bootstrap regression.

The analysis showed that our five variables (i.e. educational level, previous experience with learning languages, L1, age range and sex) accounted for 10.3% of the variance in the scores. Only two factors made a significant contribution to the model (p<.05): previous experience in learning foreign languages (B coefficient= 1.590, std.error= .310, p=.001) and educational level (B coefficient=.826, std.error=.340, p=.016).

These results suggest that the LLAMA_B test is measuring something independent of these factors. Future research will need to examine other factors (e.g. IQ, working memory or other cognitive abilities), which might be able to account for more of the variance in the scores.

The mechanics of the LLAMA_B test

Finally, we also examined the length of time allowed to learn the 20 words in the LLAMA_B test. We collected data from 99 participants, equally divided into three groups, matched for age, gender and educational level. One group took the test with the default 2 minutes learning time. A second group took the tests with a reduced learning time of one minute. The third group took the test with an increased learning time of three minutes. The results of this study are shown in Table 6.

Table 6. The effect of study time on the LLAMA_B test scores.

Study time available	**N**	**Mean**	***sd***
Short (1 minute)	30	6.70	2.64
Default (2 minutes)	33	10.06	4.90
Long (3 minutes)	33	10.79	4.56
Total	96		

The results indicate that test-takers perform worse when they are allowed less time to complete the task, and slightly better when they are allowed more time. An ANOVA was performed to compare the results in the three conditions and it showed a significant study time effect [F(2,93)=8.370, p<.001]. Post hoc Tukey tests indicated that this effect was due to the lower scores in the short study time condition, as this group was significantly different from the other two. The difference between the two longer times was not significant. We interpret this to mean that the default of two minutes study time is optimal, since shortening the time available for study produces significantly worse results, while increasing the time does not provide any

significant gains
.

References

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

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

Skehan, P. (1989, 2014) *Individual differences in second language learning* (2nd edn). New York: Routledge

Stansfield, C.W., Reed, D.J. and Velasco, A.M. (2005) *Prueba de aptitud para lenguas extranjeras - version de primaria (MLAT-ES).* Rockville, MD: Second Language Testing, Inc.

Suárez, M. and Muñoz, C. (2011) Aptitude, age and cognitive development: The MLAT-E in Spanish and Catalan. *EuroSLA Yearbook* 11, 5-29.

Reflections on Rogers and Meara 2015

For some years now, we have used the LLAMA language aptitude tests (Meara, 2005) with our final year undergraduates at Swansea University, and with our Masters level students. These students are required to undertake a small piece of empirical work and submit it in the form of a dissertation which counts towards their final degree classification. The undergraduate students ind dissertation work extremely hard, and for this reason, we have been developing group projects where the students are put into small teams and given specific projects to work on. This seems to us to be a realistic way of working with inexperienced students, in that it more closely resembles the sort of problem that they might have to deal with in future employment. Not everyone in real life has the luxury of being able to pursue a topic of their own choice, and there are obvious benefits to be had from working as apart of a team rather than struggling alone. The approach also has the significant advantage that it makes the students spend time and effort on evaluating the tools they are asked to use rather than encouraging them to develop new tools from scratch.

The LLAMA tests were devised as an umbrella project which could provide a useful framework for small projects of this sort. Our overall philosophy here is that it makes a lot of sense for groups of final year students to work together on projects which are all linked together by a single over-arching theme. This way of working naturally provides spaces where students can evaluate their different approaches to the questions they are researching, and to the data they collect. Over a number of years, this approach leads to a number of linked studies, which can be exploited by the students in later work. This set of reports make up a small literature set which is generally more accessible to students in a way that standard journal articles usually are not.

Despite its simplicity, LLAMA_B has turned out to be one of the most robust of the LLAMA tests. The test produces scores ranging from 0-20, and the task we gave the students was to establish whether these scores were measuring useful traits in language learners. This task is a lot trickier than it looks at first sight. The test generates a reasonable range of scores in the target populatio i.e. it produces a good range of scores, with little evidence of a ceiling effect, but it is not easy to establish whether these scores are related to actual language learning performance. In practice, the main difficulty that the students ran up against was finding a good measure of language learning performance that could be used as a criterion to evaluate the LLAMA_B test. This problem caused the students to reflect seriously on the nature of language testing, and by extension to question the way the University evaluated their own competence in the languages they were studying as part of their degree courses. The problem also raised a whole series of discussions about the nature of language aptitude, what it meant to say that vocabulary acquisition skills contributed X% to overall language learning performance, how you might set about quantifying this contribution or what information you needed to classify someone as "a good language learner". As a training instrument, LLAMA_B was very effective.

However, there remain some internal test issues that need to be addressed. LLAMA_B is basically about learning nouns, and it is not obvious that results for this part of speech also apply to verbs and adjectives. The physical properties of the words to be learned also seem to be important. The words in LLAMA_B are all short and relatively easy to remember, and we might expect very different results if we used words which were structurally or semantically more complex (cf. Laufer, 1997). More important, perhaps, is the fact that LLAMA_B asks test-takers to learn only 20 words. It is not at all obvious that a twenty word test provides very much information about learners' ability to learn the much larger numbers of words that are required for fluent L2 performance. Nor is it obvious that a twenty-word test is really capable of discriminating between test-takers in a meaningful way. At the moment, all we can say is that most test-takers appear to score about 40% on the test - i.e. they manage to associate about eight of the words to the correct picture, but there is considerable variation around this figure. A significant number of test-takers score very badly on the test, but on further examination these low scores can mostly be ascribed to learning strategies that do not match the test format very well, rather than poor vocabulary learning ability. A very few test-takers score very highly on the test. This suggests that the LLAMA_B test might be good at identifying very poor learners and exceptionally good ones, but might perform less well with learners whose scores are only middling. Some of these technical issues are highlighted in the More research is needed section that follows.

Much to our surprise a lot of people have used the LLAMA tests in serious research on Language Aptitude. At the time of writing, Google Scholar lists several hundred papers which have cited work based on the LLAMA tests – though only a few of these studies include a reference to the LLAMA Manual, where our reservations about the robustness of the tests and the limited nature of the data they generate are explicitly documented. It is obvious to us that some of this work uses the LLAMA programs in inappropriate contexts, and some of the work makes claims about language aptitude which are very difficult to justify given the tentative nature of the testing tools. This high level of interest reinforced our view that there is a real need for properly documented research tools for vocabulary researchers, and convinced us to include this revised version of LLAMA_B in this volume. We hope to revise the remaining LLAMA tests in the near future.

More research is needed...

This section contains some ideas that could easily be worked up into small-scale research projects.

- Are scores on the LLAMA_B test affected by the instructions given to test-takers?

- Does LLAMA_B discriminate between good language learners and weaker learners?

- What strategies do test-takers use when they are tested on LLAMA_B? Are some strategies more effective than others?

- This version of LLAMA_B presents you with a word and asks you to identify the associated picture. Would you get the same results if you gave test-takers a picture and asked them to identify its name from a list of words?

- How well would LLAMA_B work if it used spoken input rather than written words to name the pictures? Do people differ in their ability to learn spoken and written words?

- Are scores on LLAMA_B affected by test-takers' short term memory span?

- Apart from the obvious learner variables discussed in Rogers' report, what other personality variables and individual differences might you expect to influence the LLAMA_B scores?

- Can you train students to perform well on a LLAMA_B type of test?

- LLAMA_B uses real words from an unfamiliar language. Does it matter that the words to be learned are very different from English nonsense words? (There is a very large literature on how people learn sets of nonsense words paired with real English words. How much of this literature is relevant to L2 vocabulary learning?)

- How do LLAMA_B scores relate to learners' self-assessment of their ability to learn vocabulary?

References

Carroll, J.B. and Sapon, S.M. (1959) *Modern language aptitude test (MLAT).* San Antonio: Psychological Corporation.

Laufer, B. (1997) What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt and M. McCarthy (eds.) *Vocabulary: Description, Acquisition and Pedagogy* (pp.140-155). Cambridge: Cambridge University Press.

Meara, P.M. (2005) *LLAMA language aptitude tests: The Manual. Swansea: Lognostics.*

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

Service, E. and Kohonen, V. (1995) Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics* 16 (2), 155-172.

Further reading

Dahlen, K. and Caldwell-Harris, C. (2013) Rehearsal and aptitude in foreign language vocabulary learning. *Modern Language Journal* 97(4), 902-916.

Grañena, G. (2013) Cognitive aptitudes for second language learning and the LLAMA Language Aptitude test. In G. Grañena and M. Long (eds.) *Sensitive Periods, Language Aptitude, and L2 attainment* (pp. 105-130). Amsterdam: John Benjamins.

Parry, T. and Stansfield, C.W. (eds.) (1990) *Language Aptitude Reconsidered*. Englewood Cliffs, NJ: Prentice Hall.

Qais, A. How to pass the MLAT. Retrieved from: http://www.suitqaisdiaries.com/how-to-pass-the-mlat/

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

Stansfield, C.W. and Reed, D.J. (2004) The story behind the Modern Language Aptitude Test: An interview with John B. Carroll (1916-2003). *Language Assessment Quarterly* 1(1), 43-56.