Chapter 3

Testing Language Aptitude: LLAMA evolution and refinement

Vivienne Rogers, Paul Meara and Brian Rogers, Swansea University.

Introduction

Language learning aptitude is an area that has attracted increasing attention in recent years (Singleton, 2017; Wen et al., 2019). However, much research into aptitude has been driven by the availability of aptitude tests without a clear theoretical framework (cf. Skehan, 2016; Wen et al., 2017). Creating an aptitude test poses various different challenges: identifying which aspects of learning should be targeted, providing a user-friendly interface, engaging with ever-evolving technological requirements, and developing a reliable test, to name but a few. One test battery that has tackled these areas, in particular engaging with users, is the LLAMA aptitude tests originally developed by Paul Meara (Meara, 2005). This has resulted in a series of smaller, iterative changes and developments in response to user feedback. This chapter presents the story behind the creation of the original LLAMA aptitude tests, the rationale behind creating an online version in response to user feedback before introducing the new, revised online release of the LLAMA tests, created to improve their reliability (Meara & Rogers, 2021; v.3.2).[1]

The LLAMA aptitude tests (Meara, 2005) have been extensively used by researchers studying the impact of language learning aptitude on various areas of language development. Studies have ranged from the effects of different types of feedback (Kourtali & Révész, 2020; Yilmaz, 2013; Yilmaz & Grañena, 2019); the role of aptitude in language attrition (Bylund et al., 2010; Bylund & Ramírez-Galan, 2016); its relationship with explicit and implicit knowledge (Suzuki & DeKeyser, 2017b); the development of proficiency (Artieda & Muñoz, 2016; Saito, 2017; Saito et al., 2019; Suzuki & DeKeyser, 2017a); age effects (Saito, 2015); study abroad (Serrano & Llanes, 2015); and near-native language attainment (Abrahamsson & Hyltenstam, 2008) to working memory (Wen, 2016). This list barely scratches the surface, representing just a few of the areas that have been investigated. The tests have been used as an entire suite, or individual sub-tests have been used for specific research questions. The current situation highlighting the wide-spread use of the LLAMA tests is aptly summarised by Ameringer in a recent edited volume on aptitude (Reiterer, 2018):

> "A rather recent and very useful language aptitude test is the LLAMA (Meara, 2005). … It has certainly gained popularity and, as Grañena (2013) points out, only the LLAMA test does not suffer from any limitation or restriction, e.g., being difficult to get, being available only in pencil-and-paper format or only being used for military purposes. … The LLAMA test is also the test that will appear most frequently in this volume." (Ameringer, 2018, p. 27)

[1] We would like to thank all the people who sent us comments about the tests and the undergraduate students who administered the tests in a series of dissertation studies that have formed part of our ongoing investigations into the LLAMA tests.

Background to the LLAMA tests

The LLAMA tests first appeared in 2005 as a suite of programs written in the Delphi dialect of the Pascal programming language. They were designed to run on stand-alone computer workstations running the Windows OS. The programs were originally created as support materials for students following a research methods module on an MA course at Swansea University (hence the acronym LLA_MA). The programs were largely based on ideas developed in JB Carroll's *Modern Language Aptitude Test (MLAT)* (Carroll & Sapon, 1959). By 2005, MLAT was looking somewhat old-fashioned, and our students reported that they found it increasingly demotivating to work with. The LLAMA tests were an attempt to produce a more appealing interface that made use of up-to-date technology. They also included some more recent ideas in aptitude research, particularly work by Service & Kohonen (1995) and Speciale et al. (2004).

The original suite consisted of five programs, rather unimaginatively called LLAMA_A (aural memory), LLAMA_B (vocabulary learning), LLAMA_C (grammatical patterns), LLAMA_D (sound recognition), and LLAMA_E (sound-symbol correspondence). LLAMA_A and LLAMA_C turned out not to work very well: LLAMA_A required test-takers to assess their own output while LLAMA_C did not have any images making the task fairly user-unfriendly. They were both quickly abandoned and did not appear in the publicly available version of the tests, although both were part of the original battery used in the influential paper by Abrahamsson & Hyltenstam (2008). LLAMA_C was subsequently re-worked into LLAMA_F (grammatical inferencing). Downloadable versions of the final four tests were posted on the lognostics web site: http://www.lognostics.co.uk/tools/.

We had intended that the programs would mainly be used by our own MA students as part of their course work and projects, but surprisingly the website attracted a lot of traffic from other users outside of Swansea University. By 2010, the LLAMA tests were rapidly becoming a widely used research tool, despite the obvious shortcomings of the tests and caveats in the LLAMA manual (Meara, 2005). We think that the sudden upsurge was largely due to the fact that we made the LLAMA tests freely available, and their publication coincided with several influential volumes on the subject, which generated additional research interest (Grañena et al., 2016; Reiterer, 2018; Wen et al., 2019). Other, better resourced and more reliable tests were available at the time, but licenses were very expensive, and the developers of these tests often restricted their use. This made it difficult for many researchers to use the more established tests, and they seem to have turned to LLAMA as a free, no-strings attached alternative.

In practice, few users sent us the evaluations that we were hoping for, although we were very grateful to those who did. Feedback from researchers using the tests suggested they were reasonably user-friendly, and by 2010 they were regularly being cited in the research literature; according to Google Scholar, there were more than 700 citations between 2005 and 2010. This interest has grown in the intervening years. By 2021, the LLAMA aptitude test battery and its sub-components have been referenced over 4000 times according to Google Scholar, though it is worth noting that the LLAMA manual (Meara, 2005), which includes a number of important cautions about how the tests should be used, has been cited fewer than 400 times. This has caused us a degree of concern as caveats in the manual may not have been heeded by some researchers. In particular, we advised against the use of the tests in high-stakes situations, yet anecdotally we know of several instances of

the tests being used in such cases, for example, as a gateway to further study or for entry to prestigious exchange programmes.

Supporting a Windows- based suite of programs was more limiting than we had expected. Many users hoped to use the programs on laptops that used other operating systems and were disappointed when the tests could not work. There was also some demand for a version that would work on tablets, and we had regular requests for a version that could be used over a network. Moreover, Microsoft began a series of revisions to the Windows operating system that required us to reprogram the tests on a regular basis. The original tests ran comfortably on Windows XP, but later versions of Windows became increasingly complex, and expensive. As Windows XP was replaced by a succession of new systems – Vista in 2007, Windows 7 in 2009, Windows 8 in 2012 and the several versions of Windows 10 – it became increasingly difficult to keep the programs up to date, and we were faced with a succession of compatibility issues.

LLAMA v.2: Creating the first online versions of the LLAMA tests

The obvious solution to the problem of having the programs as downloadable Windows-based programs appeared to be that we should reprogram the tests so that they ran over the Internet, rather than as stand-alone programs. This approach had a number of advantages, not least that it gave us a mechanism for updating the programs as users identified problems with the coding (and there were rather a lot of these). In 2015 we began the process of converting all the LLAMA programs to web-based tasks; a web-version of LLAMA_B appeared in Meara & Miralpeix (2016) and working versions of all four programs were uploaded to the lognostics web site as LLAMA v.2.0 in 2018.

At this point, several new and exciting problems started to emerge. One problem was that we had underestimated the amount of traffic these new versions would generate, which was much greater than the use of the original Windows-based downloadable versions had suggested. This caused issues for our data-storage plans. A second problem was the variety of software used to collect data using the tests: we tested the programs on the main browsers (Chrome, Firefox, Safari), but the incompatibilities between these different browsers caused us some problems. Android, the main operating system on mobile phones, was a particularly difficult nut to crack. We kept the language neutral buttons or glyphs (e.g. arrows, (un)happy faces) from the original LLAMA tests but this proved difficult to maintain across different browsers and users reporting early exits from tests was a common occurrence. The result of these failures was that LLAMA v.2 was not entirely reliable.

At the same time as the LLAMA v.2 online tests were being created, we developed a parallel version of the LLAMA tests -dubbed the ALPACAA tests (Applied Linguistics ProgrAmmes for the Computerised Assessment of Aptitude). This version used the OpenSesame reaction time software (Mathôt et al., 2012), which allowed us to look at individual performance data, to consider various background variables (including working memory capacity) and to establish what test-takers were doing in the learning phases of the tests. This followed up our previous work investigating individual factors that might influence LLAMA test performance (Rogers et al., 2016; Rogers, Meara, et al., 2017). The results from Rogers, Meara, et al. (2017) were similar to Grañena (2013), who found that the LLAMA tests were language neutral and not impacted by gender or education level. We did, however, suggest caution with groups of mixed L2/Ln learners (see also Bokander (2020) regarding mixed L1 groups) and with younger learners. In a follow-up study, we

also confirmed Grañena's finding that the tests may be testing two different types of aptitude with LLAMA D (sound recognition) loading on a different factor to the other three tests, namely vocabulary learning, sound-symbol correspondence and grammatical inferencing (Rogers, Galvin, et al., 2017).

In the course of re-programming these tests, we fixed some of the problems of the previous downloadable versions of the LLAMA tests: in LLAMA_D, we corrected a scoring error; in LLAMA_E we changed a sound file to make it more distinct from another sound; and in LLAMA_F, we corrected an error in the choices on one item. As the original aim of the LLAMA tests had been to teach research methods skills, the tests had not been examined for reliability or validity. As increasing numbers of researchers were using the tests to examine the effects of aptitude, we felt it imperative to correct these obvious flaws (e.g. in scoring and items) to improve the reliability and validity of the tests. However, changing the online tests limits some of the comparisons between the different versions of the tests.

Overall, feedback from users suggested that LLAMA v.2 worked fairly well in terms of the interface, number of test-takers and the data generated. However, bringing together the ALPACAA reaction time data and data from LLAMA v.2 allowed us to look more closely at how test-users performed on the tests. Two of the programs, LLAMA_E and LLAMA_F, generated data that were less good than they could have been with skewed distributions and low internal reliability scores (Rogers & Meara, 2019). Our initial reliability work was also mirrored in a recent article by Bokander & Bylund (2020), based on the original downloadable versions of the LLAMA tests, which reported weak item analysis results for LLAMA_D, LLAMA_E and LLAMA_F. Initially, we thought the LLAMA_D finding may have been exacerbated by an error in the test, that meant it did not include all the test items. This error has now been corrected as mentioned above. However, in Rogers & Meara (2019) we still found low Cronbach's alpha scores for the revised LLAMA_D / ALPACAA1 (α=.544) but the sample was perhaps slightly underpowered (n=123). In the case of LLAMA_E and LLAMA_F, both tests used a binary forced-choice response method, which meant that random guessing would be expected to produce a score of 50%. This severely restricted the effective scoring range of the tests: given the penalties for getting an item incorrect, test-users needed to get more than half of the items correct in order to score above zero. Moreover, for LLAMA_E, many test-takers scored very highly, suggesting that the test had clear ceiling effects and did not discriminate well. This resulted in a narrow, somewhat skewed range of scores that needed to be improved and so we revised the layout of answers in the test phase to give 20 options instead of a binary choice. In contrast, LLAMA_B v.2 and LLAMA_D v.2 (now out of 100) both worked well and generated a range of scores that was pleasingly normally distributed.

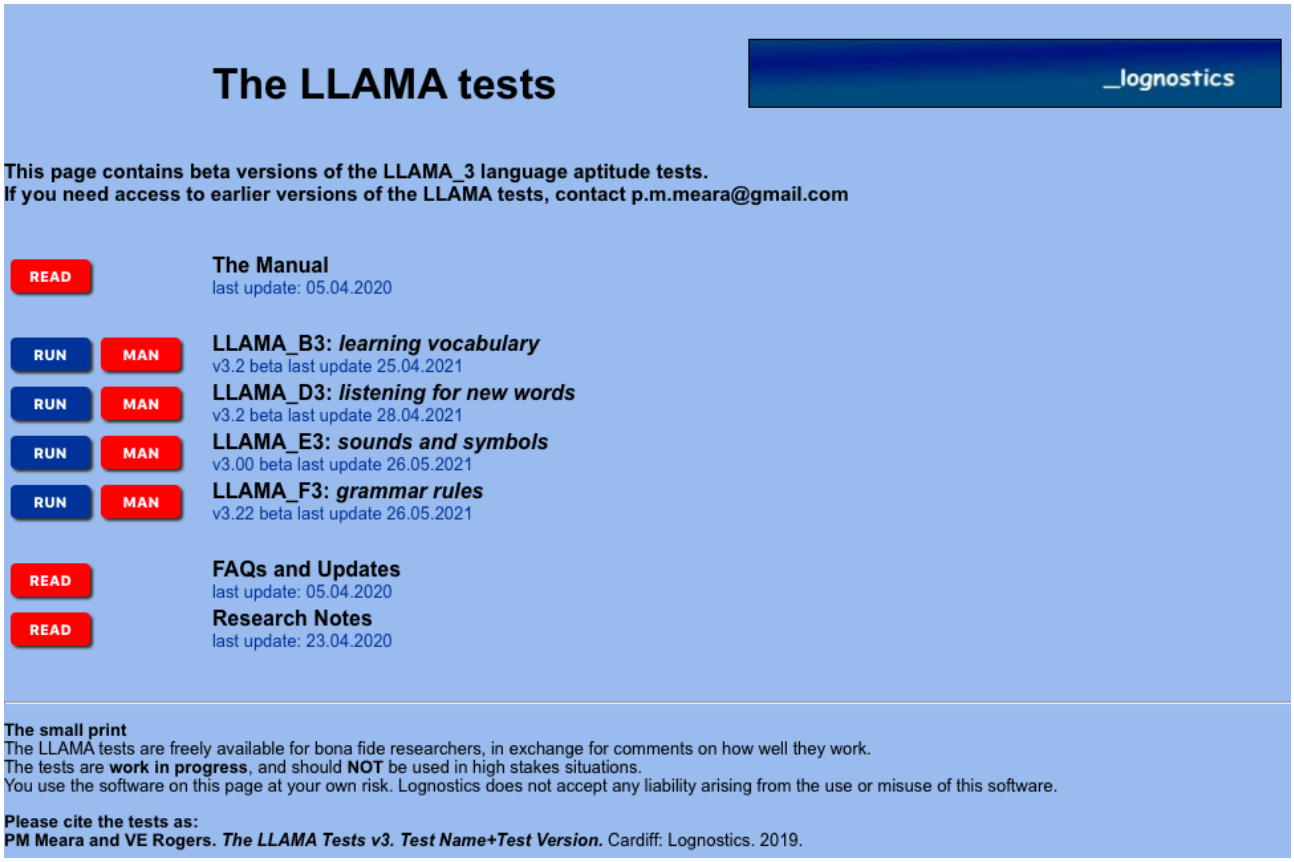
Some further exploratory work with the ALPACAA tests suggested that our revised layout for LLAMA_E did not achieve the desired effect. Our initial item analysis, reported in Rogers & Meara (2019) found that giving test-takers a larger choice of answers (in this case 20, rather than 2) made the test very difficult, and many test-takers now scored poorly (M=32.3, SD=24.329) but the larger choice gave an improved Cronbach's alpha score (α=.883). We felt that providing an increased number of test options was a move in the right direction, but we also felt that some further revision in the layout of LLAMA E and scoring method was required. We also wanted to revise LLAMA F to avoid a binary choice in this task. These considerations motivated a larger revision of the online tests to improve their internal reliability, construct validity and ease of use for the test-taker.

LLAMA v.3[2]

LLAMA version 3 is the new, revised iteration of the LLAMA tests. Substantial changes have been made to both LLAMA_E and LLAMA_F, and these will be immediately obvious to users familiar with the earlier versions. Other changes will be less obvious: these are all documented more fully in the notes that follow.

LLAMA v.3 is still an online test and is compatible with the main browsers, such as Chrome, Firefox and Safari. As we do not collect any personal information from the test-takers and therefore avoid any compliance issues with European General Data Protection Regulations, test organisers need to separately collect and manage any biographical data that is relevant to their own research. As in LLAMA v. 2, an overall manual and individual manuals for each of the sub-tests are provided via a clickable link labelled 'MAN' as shown in Figure 1. We hope this will help test organisers to administer the tests and understand the tests' limitations. Three changes have been implemented across all four LLAMA programs to improve the overall face validity of the tests. These changes affect the ID input screen, cosmetic changes to the layout of the tests, and some changes to the way the way the data are recorded.

Figure 1 Home page for the LLAMA tests including manuals

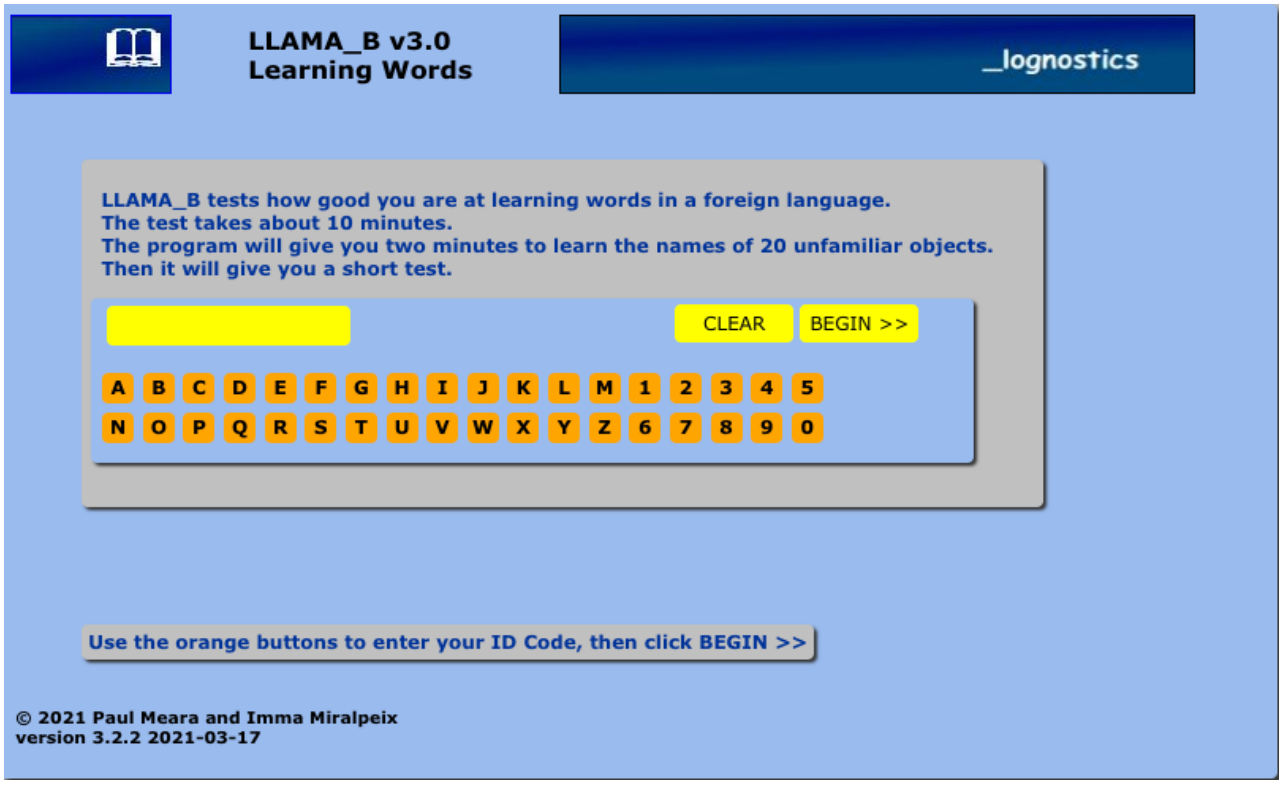


[2] *LLAMA v.3 is currently available on the lognostics website: https://www.lognostics.co.uk/tools/LLAMA_3/index.htm. However, the website is due to move to a new domain name in 2023: https://www.llamatests.org*

The first change is that the initial ID entry screen from v.2 has been changed so that only uppercase letters and digits can be used as part of a test-taker's ID. This is mainly in response to the very large number of test-takers who input their ID in non-alphabetic format or included punctuation marks, which causes problems for the scoring procedures. These options are now excluded. The new format also means that there is more consistency in the way that test-takers identify themselves (whether SWA21pete is the same person as Swa21Pete, for example). Test-takers can still run the tests without providing an ID, but cases of this sort with not be included in the data analysis graph on the report screen, as outlined below.

Figure 2 shows the input screen for the LLAMA_B test. The other tests follow this format, but with appropriate changes to the test title and short instructions. The screen includes two control buttons labelled CLEAR and BEGIN>>. These are coloured yellow on the website. CLEAR resets the ID box. BEGIN >> starts the main program.

Figure 2 The common ID entry screen.



The second change that we have made is that we no longer use language neutral symbols for instructions. In the original downloadable LLAMA tests and the LLAMA v.2 online version, we tried to avoid English language instructions, which are obviously biased against L1 speakers of other languages. Instead, we used a series of glyphs along with a reference list that could be reproduced in any number of languages. In practice this did not work well: the glyphs were not reliably reproduced by different operating systems, which resulted in serious confusion for test-takers. In LLAMA v.3 we have reverted to English language instructions. We will adapt the current versions to include other languages as this becomes necessary in response to user feedback. In addition, we have made some cosmetic changes to the control buttons, to make it more obvious

which button users should click on. Data buttons are now coloured red or blue, while controls are coloured yellow. All buttons are deactivated unless they are to be clicked. This should prevent the premature exits that characterised both previous versions.

The third set of global changes that we have implemented is much less obvious than the first two. All four programs are now scored out of 20 points, rather than as percentages. More importantly, all four programs now store the data that they collect. Each program stores two data lines for each test-taker. The first line is a record of which keypresses the test-taker makes, and looks something like this:

HADRIANVI:ABCDEFGHIJKLMNOPQRST 2021/03/17

This line tells us that HadrianVI took the test on 17th March 2021 and recorded 20 key presses listed after the colon.

The second looks like this:

HADRIANVI,15

This line tells us that HadrianVI scored 15 points on this test.

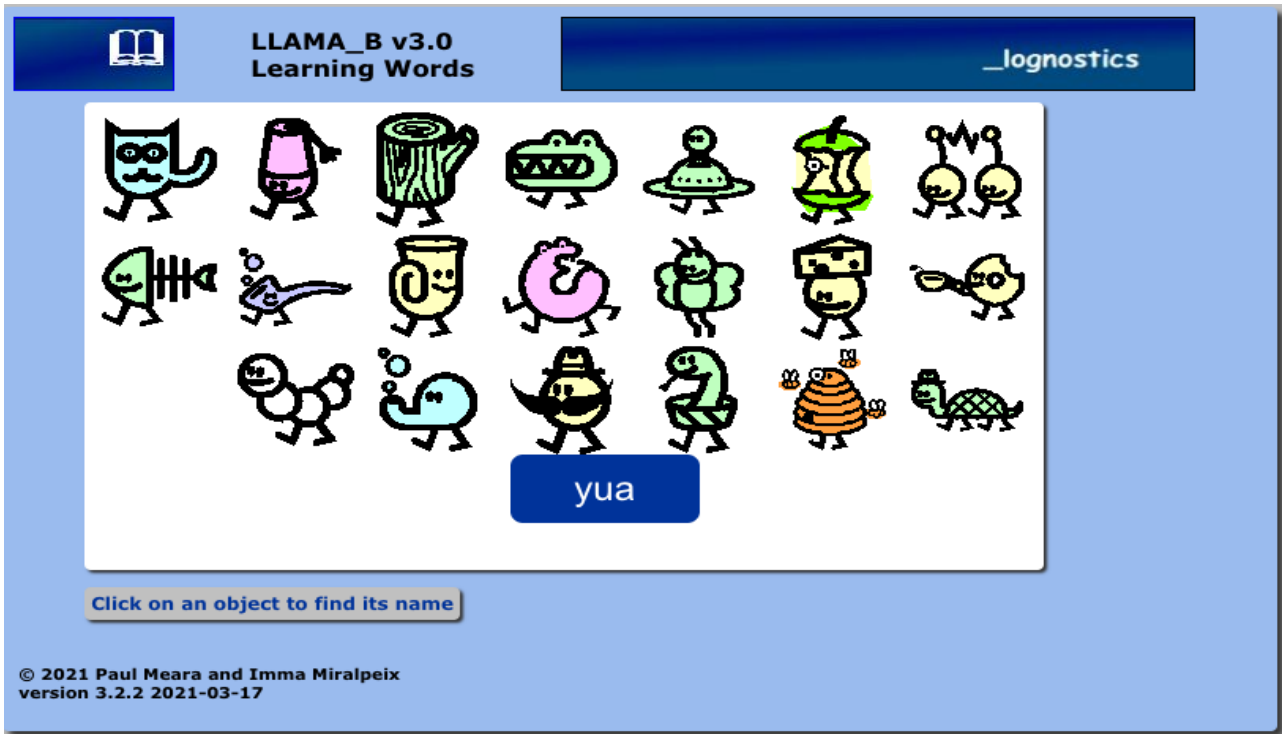
This minimal record keeping is in line with the European Data Protection Directive, in that no personal data are stored and individual test-takers cannot be identified by third parties. However, it does allow us to identify homogeneous groups within the data (for example, if they all start with the same initials), and to weed out data from test-takers who are just playing the system. (For example, test-takers who identify themselves as AAAA and score 0 are probably not serious, so we discard them when the data are reviewed.) It also means that we will be able to carry out item analyses on the different tests to allow for further refinement. In addition to these global changes, the following sections detail the specific changes that we have implemented in the separate programs.

LLAMA_B Learning words

The LLAMA_B test is largely unchanged from the earlier versions, though some cosmetic changes have been implemented in addition to the general changes outlined above. These changes are explained below. LLAMA_B has consistently performed well in previous validation studies and therefore extensive changes were not needed (Bokander & Bylund, 2020; Rogers & Meara, 2019).

The main changes concern the learning screen, which now looks like Figure 3.

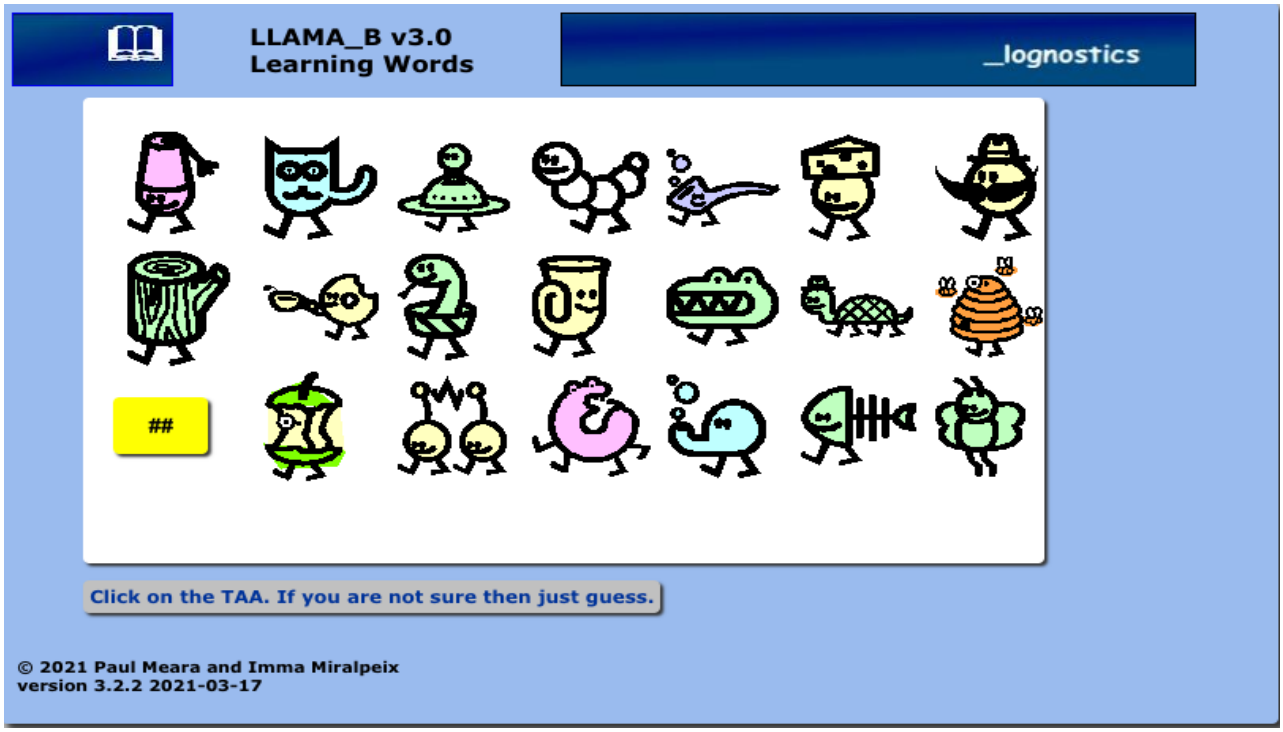
Figure 3 The LLAMA_B v.3 learning screen.



The layout of this screen is slightly different from the earlier versions, and some of the images and the names for the images have been replaced. Instructions appear in the small bar underneath the main display. Clicking on any of the objects causes its name to be displayed in a small box (e.g. *yua* in Figure 3). Test-takers have two minutes to study this material, and there are no constraints on how they do this. After two minutes, the program automatically moves on to the testing phase, as shown in Figure 4.

Again, some cosmetic differences have been implemented in the test screen in version 3. The main difference is that the test items are moved around on the screen so that they appear in different positions. This prevents test-takers from achieving a high score if they use a strategy based on the location of the original material.

Figure 4 The LLAMA_B v.3 test screen.



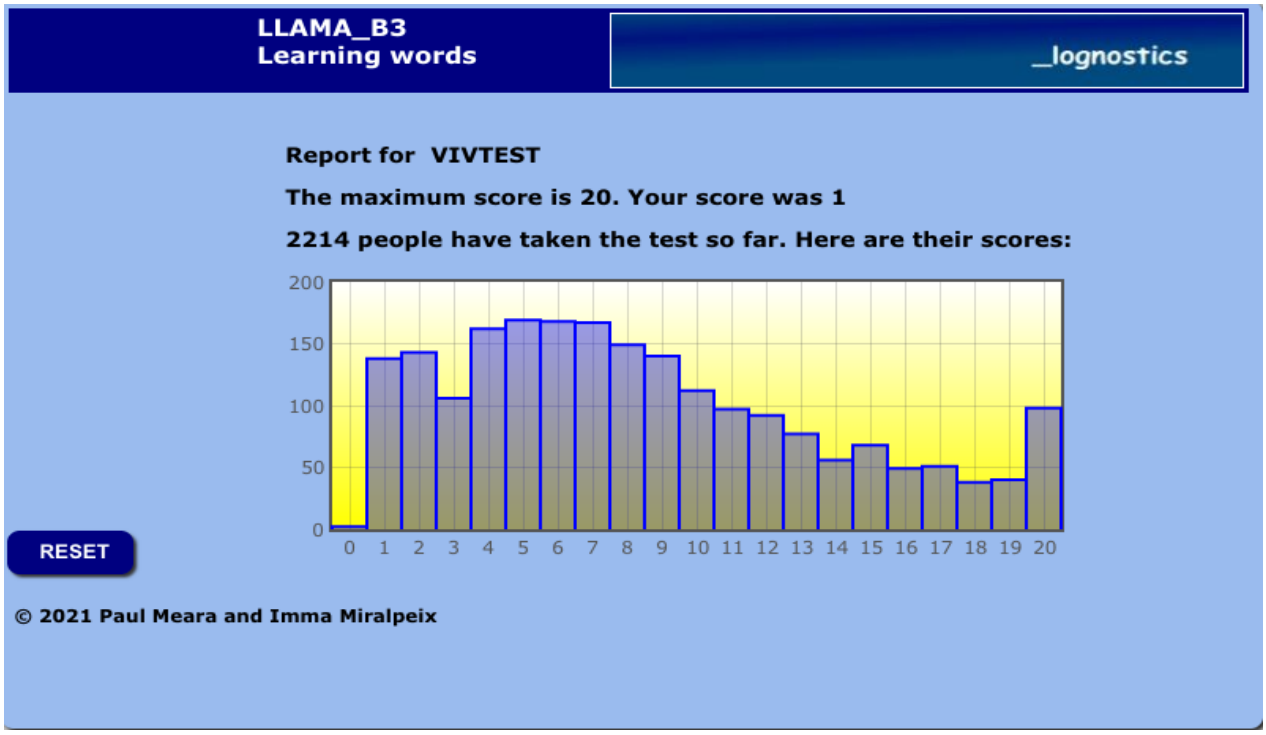
Instructions for the test appear in the small box underneath the main display. Clicking the button labelled ## displays the name of one of the 20 objects, and an instruction to click on the object that has this name. The program asks test-takers to identify the 20 objects in this way and collects a record of their key presses. After the last test item, the objects disappear, and test-takers are asked to click a new button (labelled >>) in order to see their results.

Clicking the >> button sends a data string to the scoring program as outlined above. The scoring program awards one point for each item that is correctly identified by the test-taker, with no deductions for errors. Scores can range from 0 to 20. Random guessing should result in a score of 1.

The scoring program converts this data into a report, which looks like Figure 5. This screen reports the ID supplied by the test-taker, and their score. The program then saves another data line that records this information, as outlined above. The program reports this data along with a chart that shows how previous test-takers have performed. This allows test-takers (and researchers) to interpret the meaning of their scores. At the time of writing, we have data for over 2000 test-takers, with a wide spread of results so we can be fairly confident that the report is giving us meaningful information. The example in Figure 5 shows that VIVTEST[3] scored 1 mark on the test. This is the kind of score someone who randomly guessed all the answers might get.

[3] Please note these example ID scores were generated for this paper and are not examples of good IDs.

Figure 5 The LLAMA_B v.3 Report screen.

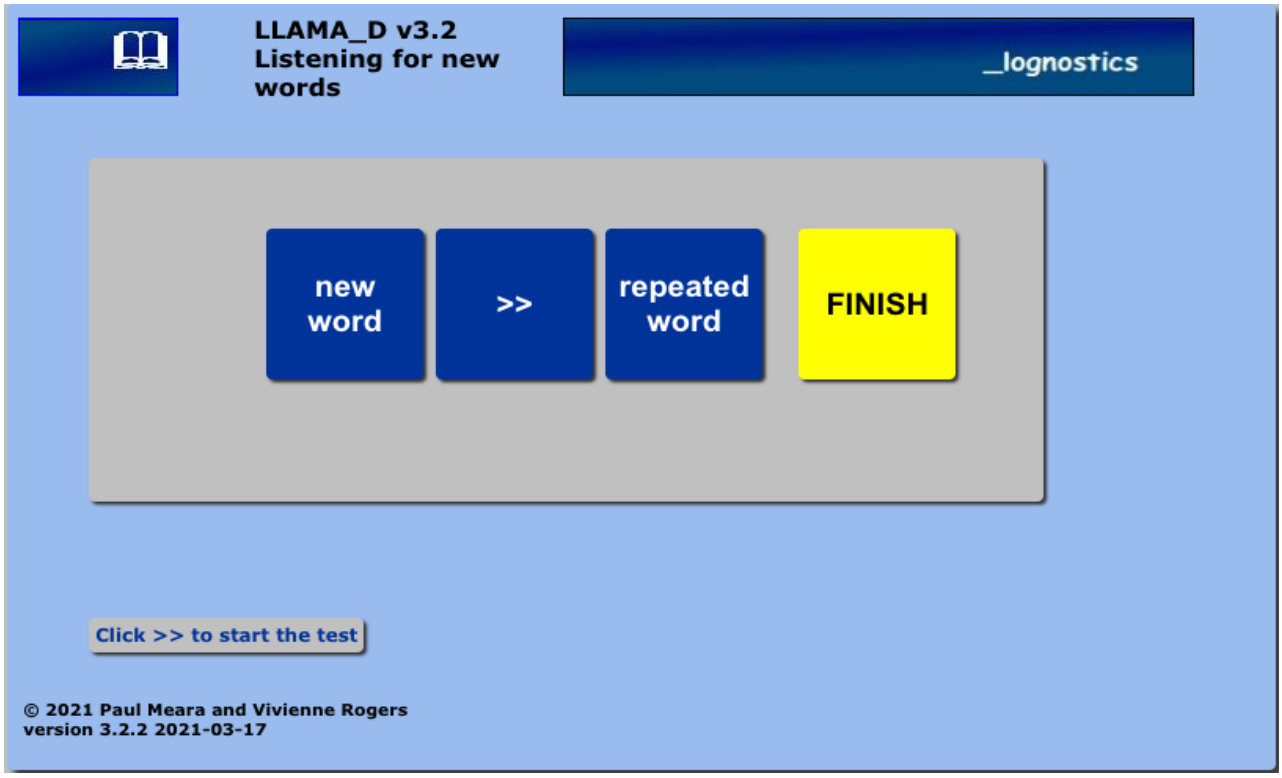


LLAMA_D Listening for new words

The new version of LLAMA_D includes two major changes from the original download version: the number of test items and the format of the learning phase. In the original download version, there was a distinct learning phase with 10 items and separate test phase with 30 items. LLAMA v.2 (and the ALPACAAs) kept this distinction between learning and test phase albeit with 40 test items instead of the original 30. However, the program now consists of a single phase with 50 items in LLAMA v.3, comprising 10 for the 'learning phase' and then 40 'test' items. Each of the original 10 items in the 'learning phase' is presented twice in the 'test phase' along with 20 new words that only appear once. We also changed one of the distractors (new words) as it sounded quite similar to one of the test items. The 50 sound files were generated by selecting a set of nouns from a North American Pacific Coast language, which were then read aloud by a text to speech generator set to expect French input. This process produces a rather unusual sound set, which test-takers are not likely to have encountered previously. The program plays the sound files one at a time, and test-takers must decide whether each sound is a new one that they have not heard before, or a repeated sound that has already appeared in the list. They signal this by clicking the appropriate button as shown in Figure 6.

The additional changes implemented in version 3 in comparison to version 2 are cosmetic. They are principally designed to make the test easier to use, and to avoid accidental keypresses that caused earlier versions to exit prematurely. LLAMA_D also incorporates the new ID-input page, and it saves data in the format described in the previous section. The main test page has been significantly re-designed, and the current version is shown in Figure 6:

Figure 6 The LLAMA_D v.3 test screen.

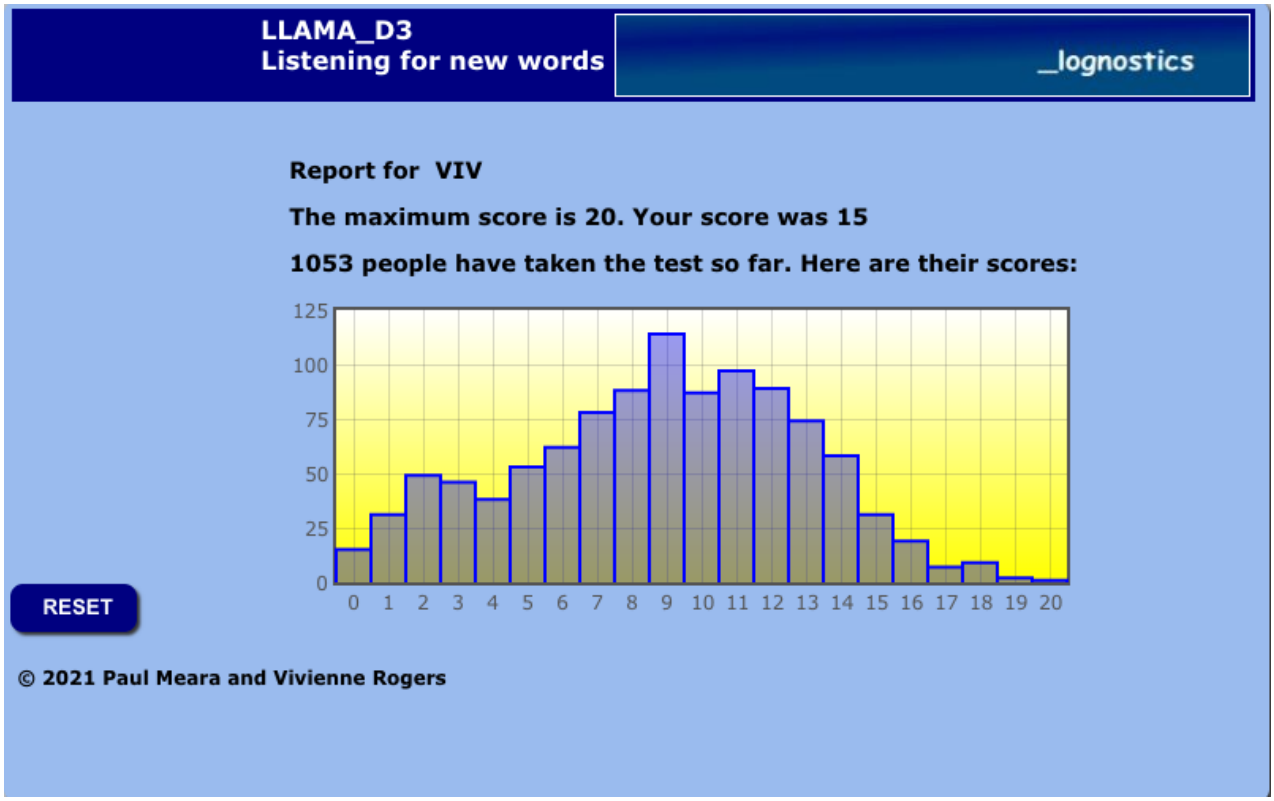


The only changes here are that the buttons have been made easier to navigate, and we have replaced the wordless button labels, which had happy and unhappy faces, with a more transparent, explicit description of what each button does. Feedback from users suggested that the use of happy and unhappy faces caused confusion for some test-takers as they were not clear that a happy face meant that you recognised the word. Buttons can only be clicked in the correct sequence, and the yellow FINISH button remains deactivated until the test is completed.

Clicking the FINISH button at the end of the test submits a data string for scoring. As mentioned above, the program treats the first 10 responses as training or learning data. The remaining 40 items are awarded one point for each item that is correctly identified as a new or repeated word. One point is deducted for each incorrect answer. There are 40 items in the test sequence, so the scores can range from 0-40. Negative scores are possible, but scores of this type are normalised to zero. The final scores are halved so that this test is consistent with the other tests in the suite, which all have a maximum score of 20 points. Guessing behaviour should produce scores close to zero. If you consistently pick one answer, then you will also score zero. We are currently modelling different scoring methods using both raw and adjusted scores.

The program computes an overall score for each test-taker and stores this data in the usual format. The program then makes a report in the usual format (cf. Figure 7 below).

Figure 7 The LLAMA_D v.3 report screen.



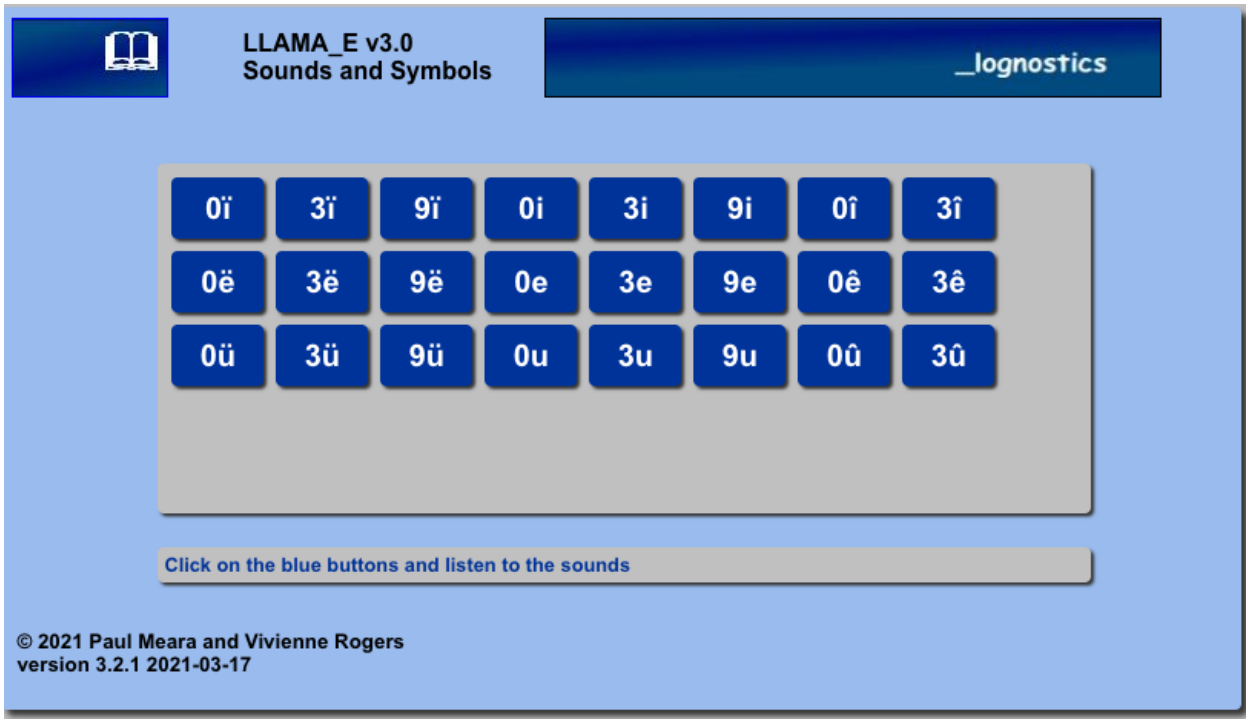
Here the report indicates that VIV scored 15 points, a very good score on this test. At the time of writing, just over a thousand people have taken the LLAMA_D test, and the data appears to be normally distributed.

LLAMA_E Sounds and Symbols

The global changes outlined in the previous sections have also been implemented in version 3 of the LLAMA_E program. This program now has the standard ID input screen, and the shape and position of some of the controls has been changed to make accidental fatal keypresses more difficult. However, LLAMA_E has also undergone more significant changes. The data we collected from earlier versions suggested that test-takers found LLAMA_E either very difficult, or fairly easy – depending to some extent on their experience with formal phonetics (Rogers, Meara, et al., 2017). LLAMA_E was scored using a binary forced choice test, and this meant that test-takers could score 50% by guessing randomly. The new version has been designed to eliminate these issues.

The LLAMA_E learning screen remains largely unchanged and looks like Figure 8.

Figure 8 The LLAMA_E v.3 learning screen

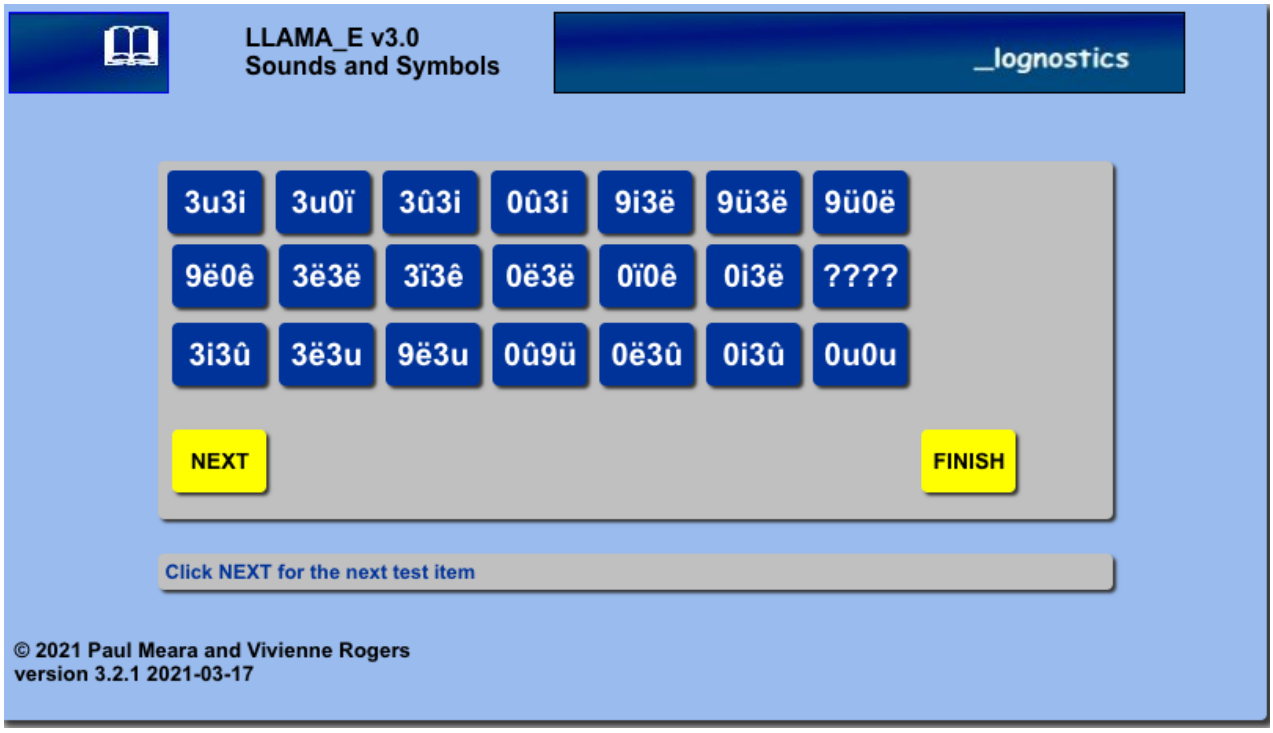


Clicking on any of the buttons plays a sound file corresponding to the letters displayed on the button. Test-takers are allowed two minutes to study this material. The instructions do not specify how the test-takers should do this since there is no obvious way of checking how test-takers use the materials. Test administrators may wish to instruct their participants according to their own research needs.

Each sound is a short Consonant+Vowel syllable, and test-takers who are familiar with phonetics will probably notice that the lay-out of this screen in Figure 8 is not arbitrary. Syllables that begin with voiceless sounds - [p], [t] and [k] - are all positioned on the left of the screen; syllables beginning with voiced sounds – [b], [d] and [g] – are positioned in the three central columns. The final two columns on the right of the screen are all sounds beginning with nasal sounds. Moreover, the rows are all consistent in the types of vowels presented. This gives a slight advantage to test-takers who are familiar with formal phonetics.

This screen is displayed for two minutes. Once this learning time has elapsed, it is replaced by a testing screen that looks like Figure 9 below. On this screen there are 20 buttons, each labelled with a two-syllable word made up from the individual syllables that were used in the learning phase of this program. For example, in the learning phase, the participant may have heard /di/ and /ma/ separately but in the test, they need to identify the combination /dima/. These buttons are not organised randomly (as they were in the original ALPACAA version): they are grouped so that words ending with similar sounds appear together. There is also a button labelled ??? which records a DON'T KNOW response.

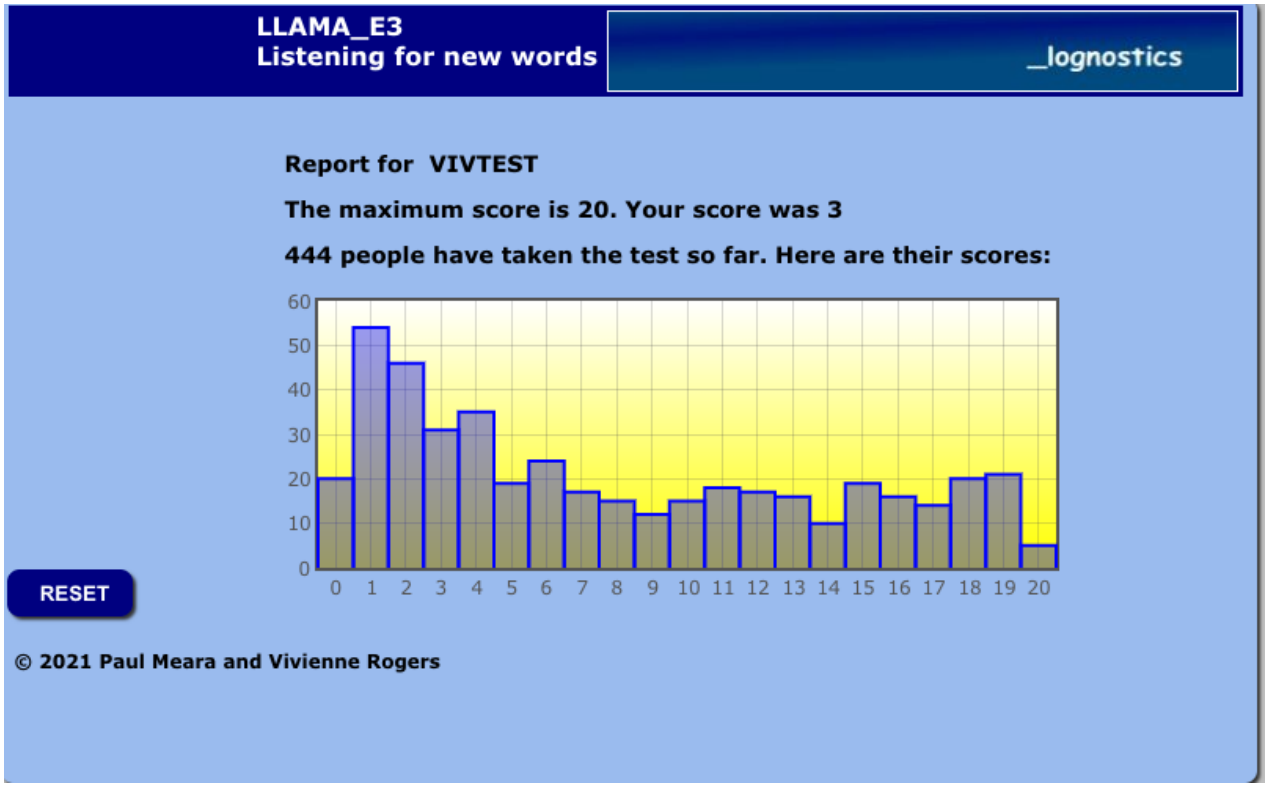
Figure 9 The LLAMA_E v.3 test screen



Instructions appear in the strip at the bottom of the display. First, test-takers are instructed to click the NEXT button. Doing this makes the program play a short sound file that consists of two of the syllables. Test-takers are then instructed to click on the button that corresponds to the sound file they just heard. It is not possible to play the sound multiple times and the programme will wait until a response button is pressed. There are 20 test items. After the last item, test-takers are instructed to click the FINISH button. This task is considerably harder than the binary task used in the original version of LLAMA_E but easier to navigate than the ALPACAA test version we piloted (Rogers & Meara, 2019).

Scores on this test range from 0 to 20. Each test response is given one point if it is fully correct. No points are awarded for partially correct responses. Marks are not deducted for incorrect responses. LLAMA_E scores the data it collects as two strings in the same way as the other tests previously outlined and generates the usual report screen shown in Figure 10. At the time of writing, we have data from over 400 test-takers. The scores appear to be somewhat skewed towards the lower end of the range. We think this is because test-takers who find this test difficult will typically answer a handful of test items correctly, and then resort to guessing. Figure 10 shows that VIVTEST scored three points on this test run, a score that is consistent with this interpretation. There is still work to be done on this test. Looking forward, we will carry out further analysis of the scores in this test and will model giving partial credit for answers to see what improvements, if any, that may make to the distribution of scores and the test's discrimination between participants.

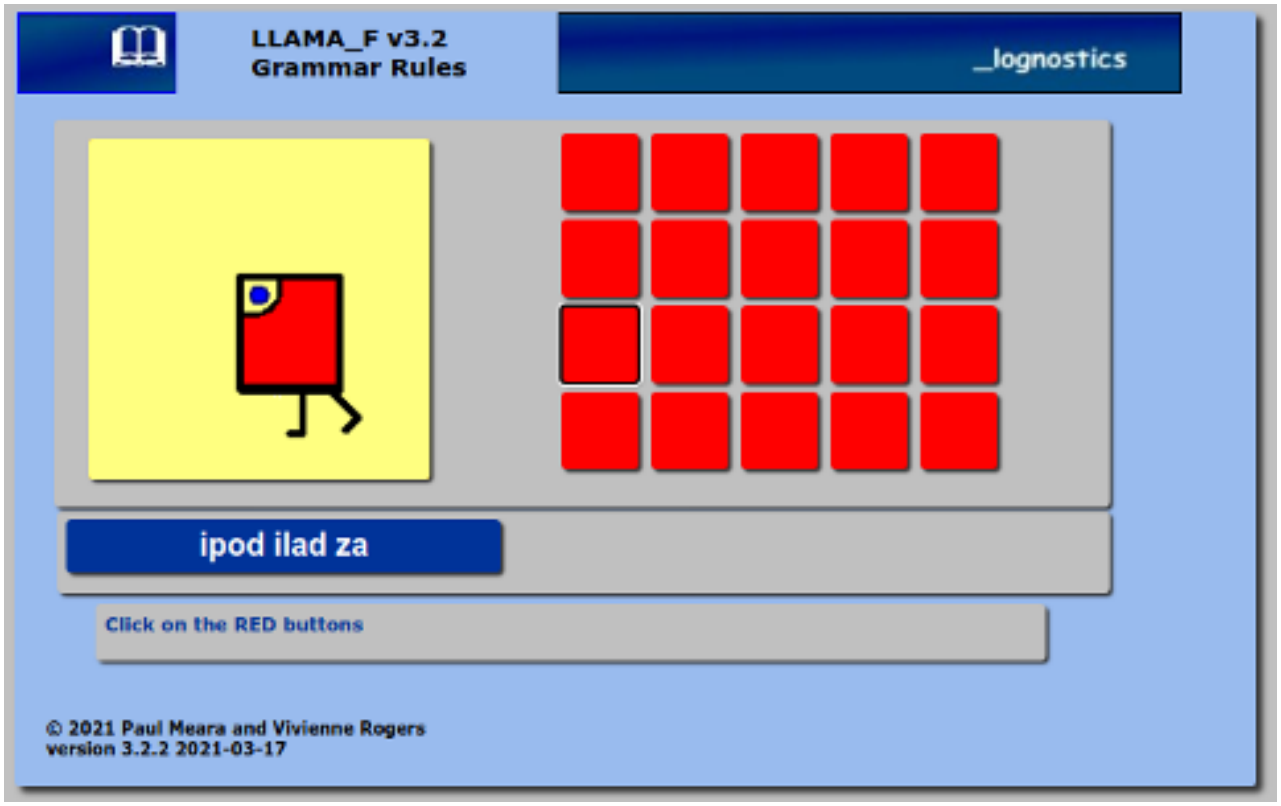
Figure 10 The LLAMA_E v.3 report screen



LLAMA_F Grammatical inferencing

LLAMA_F has also undergone significant changes from earlier versions. In this test, test-takers are tasked with learning the rules of an invented language called *PATSI*. The main display screen is shown in Figure 11 below. Here, the changes from earlier versions are mainly cosmetic.

Figure 11 The LLAMA_F v.3 learning screen



Clicking on one of the square buttons displays an image along with a description of the image in an artificial language, *PATSI,* which has some interesting morphological and syntactic properties. Test-takers get four minutes to study the language: earlier versions allowed five minutes, but the additional time did not materially affect the results as discussed in Rogers et al. (2016). The shorter time has been implemented in version 3 in the interests of brevity. How test-takers do their learning is up to them, and there are no restrictions on how they use the available time. The program tries not to force test-takers to think in a specific way about what the pictures show. For example, the picture in Figure 11 might be interpreted *a red thing,* or *as a square thing,* or *as a walking thing,* or in any number of other ways. Some of these ways will line up with the *PATSI* description of the picture, while others will not. For example, if a test-taker identifies this picture as *a thing with one eye*, or as just *a thing,* then it might be more difficult for them to isolate the correct syntactic features in the other pictures. However, some test-takers will notice that there are three "words" in the *PATSI* description of this picture, and this might indicate to them that *ipod ilad za* is quite a specific description. Precisely what each of these words refers to only becomes clear as more pictures are observed and more *PATSI* descriptions are noted.

PATSI has a small vocabulary - only 20 words - and a fairly simple sentence structure. The main features are:

1: adjectives follow their nouns
2: Sentences are verb (or preposition) initial[4]
3: Nouns take a singular marker, but not a plural marker.

[4] When designing the tests, we viewed these items, for example, *umush*, as verbs. However, many others tend to view these as prepositions. We've kept both options in the text throughout.

4: The singular marker follows its noun.
5: Numerals precede their noun.
6: Affixes have several different forms (similar to grammatical gender).

The pictures are designed to make these grammatical features obvious, and most experienced linguists will have no difficulty recognising them.

The learning materials are common to LLAMA_F version 3 and the earlier versions of the program. We have, however, made some changes to the layout of this page to make it more obvious to test-takers what they have to do, and to make it more difficult for test-takers to exit the test accidentally. Accidentally exiting the test was a particular problem in our data collection with the downloadable LLAMA tests, as many test-takers thought the exit button was a 'continue' button and clicked it with the intention of moving on. They then re-started the program. This meant that some test-takers may have completed some or all of the learning phase multiple times before actually progressing to the test. This is clearly a problem for the reliability and construct validity of earlier versions of this test.

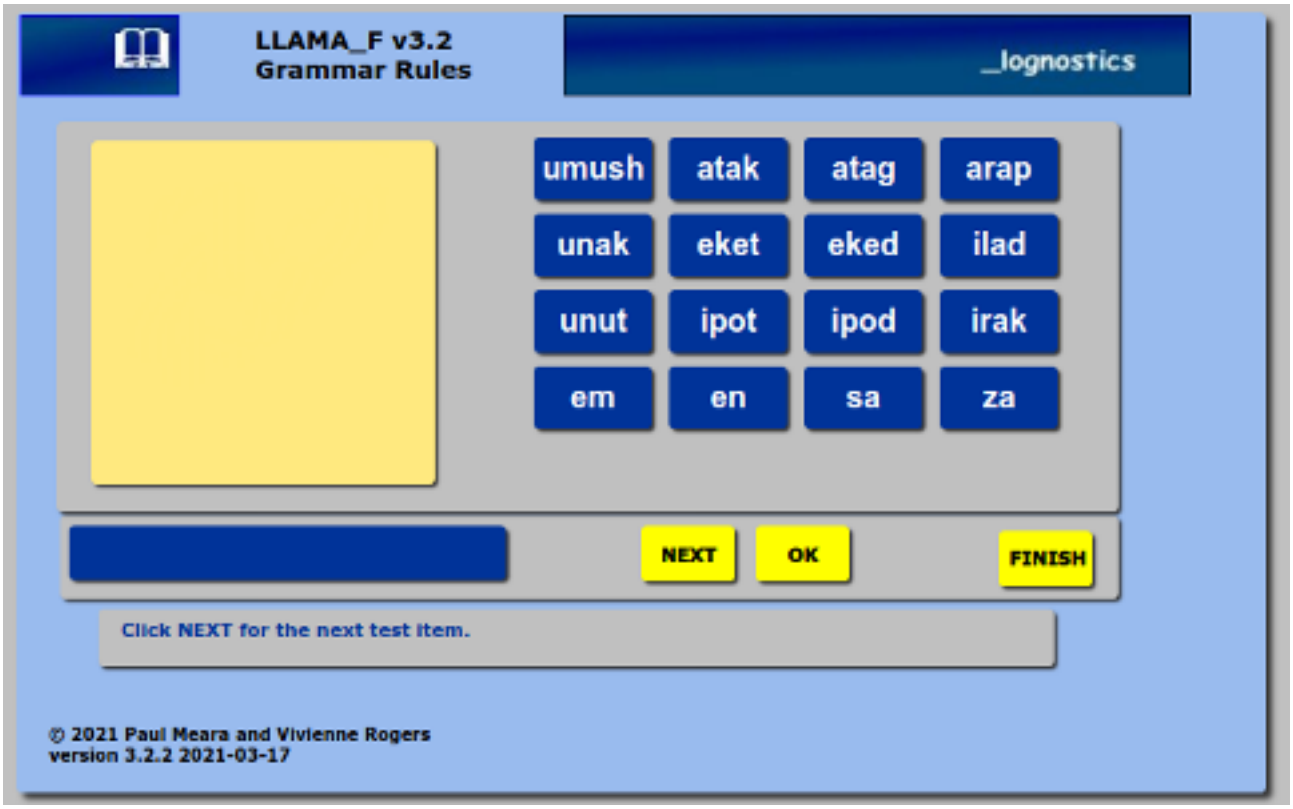
As in LLAMA_E, we have not specified if participants can take notes. Rogers, Meara, et al (2017) report that note taking was not beneficial and resulted in strategic behaviour to work out the rules after the learning phase not during it:

> "We have conducted two versions of the test; one in which our participants could take notes (this study, n = 211) and our previous study in which participants could not take notes (n = 135). A t-test did not show any difference ($t(344) = 0.268$, $p = 0.789$) between participants who were allowed to take notes (M = 41.42, s.d. = 26.28) and those who were not (M = 42.22, s.d. = 28.35). Anecdotally, we noticed that those who were permitted to take notes did so and also made use of the full five minutes of learning time, whereas those who could not take notes did not use the full five minutes. We also noted that quite a few of the note takers wrote out the sentences as a whole and drew pictures. They then tried to work out the rules in the testing phase rather than using the learning phase to do so. This was contrary to the instructions they were given." (Rogers, Meara, et al., 2017, n.6)

This strategic behaviour of test-takers to work out the rules after the learning phase is a limitation that we have not been able to mitigate. As with LLAMA_E, test organisers may wish to instruct their participants according to their own research needs if this issue is problematic.

The test screen, shown in Figure 12 has been completely re-designed from earlier versions. Again, some global changes have been introduced that make it harder for test-takers to exit the program accidentally. We have also coloured the buttons to distinguish their function as previously mentioned.

Figure 12 The LLAMA_F v.3 test screen



More importantly, we have completely changed the way in which test-takers' knowledge of *PATSI* is evaluated. In the earlier versions, each test item consisted of a pair of sentences, one of which was grammatically correct, while the other violated one of the rules that characterise *PATSI*. Test-takers were asked to identify which of the two sentences correctly described the test picture. The main problem with this approach is that guessing behaviour on the part of test-takers will produce about 50% correct answers by chance, and this means that the effective mark scale is very short. Earlier versions of the test contained 20 items, each scoring one point, so the effective scale ran from 10 to 20 points - or 50-100% when scaled up. However, marks were deducted for incorrect answers, meaning a test-taker needed to get above 50% correct in order to score above zero. In practice, the results tended to be bimodal: a few test-takers scored close to full marks, but most indicated a high degree of guessing behaviour (Bokander & Bylund, 2020; Rogers, Meara, et al., 2017; Rogers & Meara, 2019).

The new version of the test attempts to alleviate this problem by taking a different approach to assessing how much test-takers have learned about *PATSI*. Instead of a binary choice (sentence A or sentence B), the new version asks test-takers to actually construct *PATSI* sentences. However, we did not want to penalise test-takers for minor spelling inaccuracies, so we do not ask them to write *PATSI* sentences on a blank slate. Instead, we provide a set of 20 buttons, one for each word they have met in the learning phase. Clicking on one of the buttons adds the appropriate word to the answer bar underneath the word list.

This approach makes it less important for test-takers to remember what the individual words mean, and the number of possible answers that test-takers can make effectively rules removes the

possibility of producing random combinations of responses that are both semantically and syntactically correct. The test should now target the test-takers' ability to generate grammatical inferences and focus less on their ability to remember specific vocabulary thus improving the overall construct validity of the test. However, it does make the scoring much more difficult.

The approach we have adopted here is slightly unusual in comparison to the scoring of the other tests. Unlike in the previous version of LLAMA_F, there are now only ten test items (as opposed to 20), each consisting of a picture like the ones that were presented in the learning phase. This change was introduced to reduce test-taker fatigue but given the changes to the nature of the test noted above and explained more fully below, there should still be sufficient discriminatory power. Initial results support this (see Figure 13). Importantly, the items are not scored for complete accuracy. Each item is scored for two syntactic features, and one point is awarded for each feature that correctly appears in the test-taker's answer. Any other features that appear in the answer, whether correct or incorrect are ignored.

An example might make this clearer. Suppose that we have an item that is designed to test two features:

i. verbs/prepositions are sentence initial; and
ii. sentences with a singular subject end in a singular marker.

Any response that includes a verb/preposition in initial position will score one point, even if the verb/preposition is semantically incorrect. Similarly, any response that contains a singular noun and ends in a singular marker will be awarded one point. It does not matter which singular noun the test-taker chooses, and it does not matter whether the correct singular marker is chosen.

The test-taker's responses are scored by matching them to two Perl Compatible Regular Expressions that encapsulate the syntactic features that we are looking for. It is fairly straightforward to design regular expressions that correspond to the main syntactic features in *PATSI*. For example, if the program knows that responses A, B and C correspond to verbs/prepositions in *PATSI*, then the regular expression ^[A-C] will capture any response that has a verb/preposition in initial position. Similarly, if G, H and I are nouns, while K and L are two different singular affixes, then the regular expression (G|H|I)(K|L) will capture any response which contains a noun immediately followed by one of the two affixes, ignoring whether the correct choice of affix has been made.

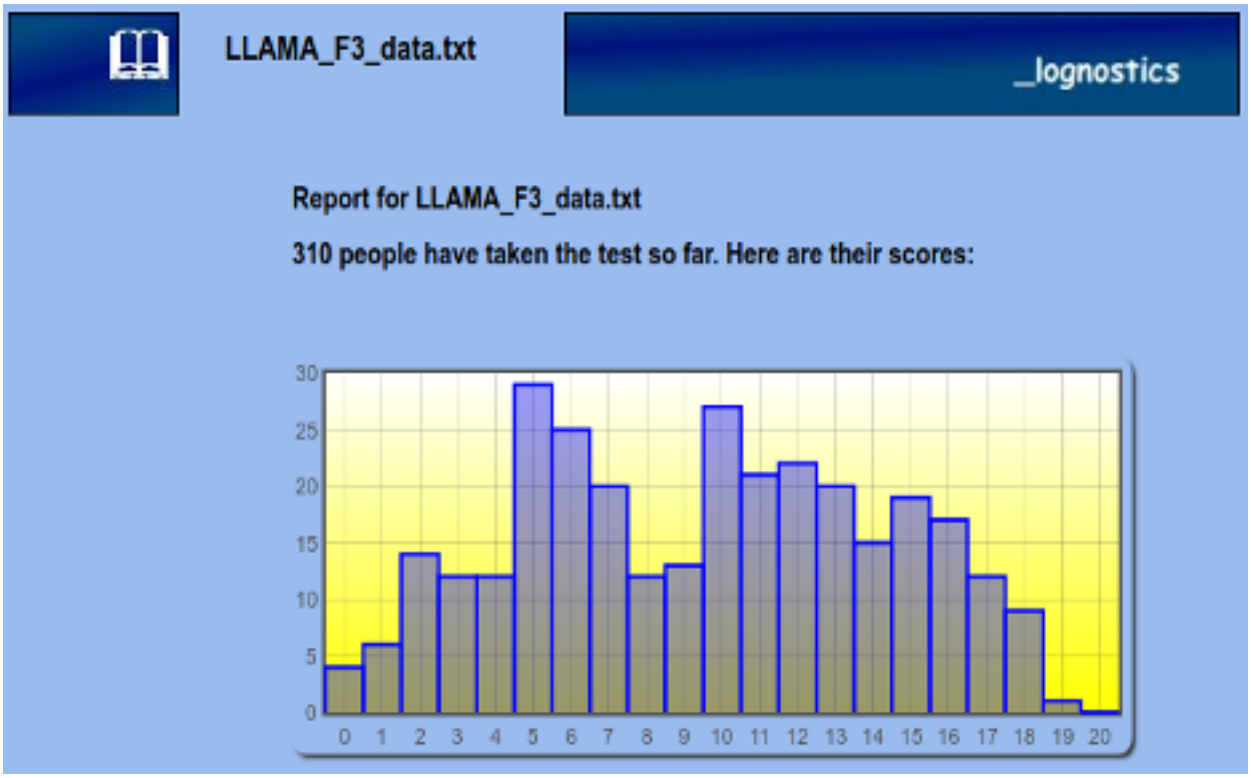
The rules that the regular expressions test are given below and each is tested four times:

Rule 1: Prepositions/ verbs are sentence initial
Rule 2: Adjectives precede their nouns
Rule 3: The singular marker is sentence final and varies with noun class
Rule 4: Numerals precede the noun
Rule 5: Adjectives agree with their nouns

Some of these rules seem to us to be easier to learn than others as the basic word order of adjective-noun is present in all items whereas the singular marker (for example) is not. The new data collection protocols will allow us to examine this hunch in detail, and we expect to be able to improve on our choice of test items once we have collected more data with this version. The

original downloadable versions of LLAMA_F also showed rule-clustering effects (Bokander & Bylund, 2020). We have checked a subset of 56 test-takers, who took the new LLAMA_F as part of an undergraduate dissertation experiment investigating aptitude and memory. We found with an absolute "all or nothing" scoring system, users typically scored 0 or 1 with a maximum score of 4/10 (M=0.48, Mdn=0, SD=0.953, Min=0, Max=4). However, when using the five rule system, the range was much larger, allowing for greater discrimination between participants (M=8.47, Mdn=8, SD=4.509, Min=0, Max=17).

Figure 13 The LLAMA_F v.3 report screen



As with all the LLAMA v.3 tests, each test-taker's data are stored in a file which allows us to analyse the responses in some detail. At the time of writing, we have data from 310 test-takers. Their scores are summarised in Figure 13. Generally speaking, this new version is giving us a better spread of scores than earlier versions, and there is no evidence of ceiling or floor effects. The appearance of a dip in the distribution for scores 8 and 9 is odd, but the raw data files will allow us to investigate further whether this is a reliable effect or not. We are currently collecting data with background biographic information and memory measures as part of our ALPACAA testing to establish if 10 items, each scored twice gives us a more reliable test that is sufficient to discriminate between learners.

Conclusion

This chapter has sought to present the new, revised, online LLAMA tests and the context of their creation. It is probably not an exaggeration to suggest that the development of the LLAMA tests has been a game-changer for aptitude research. The most important factors here are that we have made the LLAMA tests freely available to researchers, and that we treat users as collaborators, rather than

as customers. Most of the changes introduced in version 3 are a response to comments made by engaged users. Iterative improvements to the performance, reliability and validity of tests such as these are an on-going process and most effective when they are informed by users of the tools.

Finally, we should perhaps point out that the LLAMA tests have been much more widely used than we originally intended. Specifically, the LLAMA tests were designed to be used by adult learners, and they may not be appropriate for young learners. Researchers using the LLAMA tests with groups of this sort will need to exercise appropriate caution and report their findings with the necessary caveats. We also repeat our warning that the LLAMA tests should not be used in high-stakes situations. We still consider the LLAMA tests to be work in progress, rather than a finished product, and we hope that users will be mindful of this limitation when they use the tests.

After a long period in the wilderness, aptitude research is once again an active area of research. This is in no small part due to the creation of modern, easily accessible tests, of which the LLAMA tests are a prime example. Despite their provisional status, it is worth noting that published research using the LLAMA tests suggests that they are indeed measuring some interesting features that characterise individual differences in language learning ability. Much further investigation of aptitude in general, and the LLAMA tests in particular, is needed. The relationship between aptitude and other areas of language acquisition, attrition and pedagogy is already recognised, and we hope that the revised, online LLAMA v.3 tests will help meet some of the needs of the research community.

References

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

Ameringer, V. (2018). Cognitive Abilities: Different Memory Functions and Language Aptitude. In Suzanne Reiterer (Ed.), *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (pp. 19–42). Springer.

Artieda, G., & Muñoz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. *Learning and Individual Differences*, *50*. https://doi.org/10.1016/j.lindif.2016.06.023

Bokander, L. (2020) Language aptitude and Crosslinguistic Influence in Initial L2 learning. *Journal of the European Second Language Association*, *4*(1), 35–44. DOI: http://doi.org/10.22599/jesla.69

Bokander, L., & Bylund, E. (2020). Probing the Internal Validity of the LLAMA Language Aptitude Tests. *Language Learning*. https://doi.org/10.1111/lang.12368

Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2010). The role of language aptitude in first language attrition: The case of pre-pubescent attriters. *Applied Linguistics*, *31*(3), 443–464. https://doi.org/10.1093/applin/amp059

Bylund, E., & Ramírez-Galán, P. (2016). Language Aptitude in First Language Attrition: A Study on Late Spanish-Swedish Bilinguals. *Applied Linguistics*. https://doi.org/10.1093/applin/amu055

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test.* Psychological Corporation.

Grañena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test. In G. Grañena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (Vol. 35, pp. 105–130). John Benjamins Publishing.

Grañena, G., Jackson, D. O., & Yilmaz, Y. (2016). *Cognitive individual differences in second language learning and processing*. John Benjamins.

Kourtali, N. E., & Révész, A. (2020). The Roles of Recasts, Task Complexity, and Aptitude in Child Second Language Development. *Language Learning*. https://doi.org/10.1111/lang.12374

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2). https://doi.org/10.3758/s13428-011-0168-7

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

Meara, P., & Miralpeix, I. (2016). Tools for Researching Vocabulary. In *Tools for Researching Vocabulary*. Multilingual Matters. https://doi.org/10.21832/9781783096473

Meara, P., & Rogers, V. (2021). The LLAMA Tests v.3.2. Cardiff: Lognostics.

Reiterer, Susanne (Ed.). (2018). *Exploring language aptitude: Views from psychology, the language sciences, and cognitive neuroscience* (Vol. 16). Springer.

Rogers, V., Galvin, T., Cobner, A., Chisholm, M., Clothier, J., & Greenfield, I. (2017). Investigating the Relationship between Working Memory and Language Learning Aptitude. In *https://viviennerogers.info/wp-content/uploads/2020/09/EuroSLA-Poster.pdf*. EUROSLA conference.

Rogers, V., & Meara, P. (2019). *Turning a LLAMA into an ALPACAA and back again: An initial revised attempt at assessing aptitude*. Language Aptitude Roundtable. https://viviennerogers.info/wp-content/uploads/2020/09/alpacaa_2019_macau.pdf

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

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

Saito, K. (2015). The Role of Age of Acquisition in Late Second Language Oral Proficiency Attainment. *Studies in Second Language Acquisition*, *37*(4), 713–743.

Saito, K. (2017). Effects of Sound, Vocabulary, and Grammar Learning Aptitude on Adult Second Language Speech Attainment in Foreign Language Classrooms. *Language Learning*, *67*(3). https://doi.org/10.1111/lang.12244

Saito, K., Suzukida, Y., & Sun, H. (2019). Aptitude, experience, and second language pronunciation proficiency development in classroom settings. *Studies in Second Language Acquisition*, *41*(1). https://doi.org/10.1017/S0272263117000432

Serrano, R., & Llanes, À. (2015). An exploratory study of the role of age and language learning aptitude in a short stay abroad. *Vigo International Journal of Applied Linguistics*.

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.

Singleton, D. (2017). Language aptitude: Desirable trait or acquirable attribute? *Studies in Second Language Learning and Teaching*, *7*(1), 89. https://doi.org/10.14746/ssllt.2017.7.1.5

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive Individual Differences in L2 Processing and Acquisition*. John Benjamins.

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.

Suzuki, Y., & DeKeyser, R. (2017a). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*, *38*(1). https://doi.org/10.1017/S0142716416000084

Suzuki, Y., & DeKeyser, R. (2017b). The Interface of Explicit and Implicit Knowledge in a Second Language: Insights From Individual Differences in Cognitive Aptitudes. *Language Learning*, *67*(4), 747–790. https://doi.org/10.1111/lang.12241

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

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

Wen, Z. E., Skehan, P., Biedroń, A., Li, S., & Sparks, R. L. (Eds.). (2019). *Language aptitude: Advancing theory, testing, research and practice*. Routledge.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics*. https://doi.org/10.1093/applin/ams044

Yilmaz, Y., & Grañena, G. (2019). Cognitive Individual Differences as Predictors of Improvement and Awareness Under Implicit and Explicit Feedback Conditions. *Modern Language Journal*. https://doi.org/10.1111/modl.12587