LANGUAGE APTITUDE IN YOUNG LEARNERS: THE ELEMENTARY MODERN LANGUAGE APTITUDE TEST IN SPANISH AND CATALAN

Tesi doctoral presentada per

Maria del Mar Suárez Vilagran

com a requeriment per a l'obtenció del títol de

Doctora en Filologia Anglesa

Programa de Doctorat *Lingüística Aplicada*
Bienni 2002-2004
Departament de Filologia Anglesa i Alemanya
Universitat de Barcelona
Barcelona, 2010

Directora: **Dra. Carmen Muñoz Lahoz**

UNIVERSITAT DE BARCELONA

Language Aptitude in Young Learners: The Elementary Modern Language Aptitude Test in Spanish and Catalan

Maria del Mar Suárez Vilagran

CHAPTER 2: MEASURING LANGUAGE APTITUDE: THE MODERN LANGUAGE APTITUDE TEST (MLAT) AND THE MODERN LANGUAGE APTITUDE TEST – ELEMENTARY (MLAT-E)

2.1. Introduction

A recent growing interest in the study of aptitude has enticed linguists to design, with more or less success, new measures of aptitude with the aim of overcoming the lacks in already existing tests (see section 1.3). One test which has prevailed since its creation and which is still in use, at times in combination with other tests released more recently, is the Modern Language Aptitude Test (MLAT). This test peaked in popularity in the 1960s and 70s. After some popularity ups and downs, the MLAT has survived the passing of time and has recently been computerised, keeping the same structure and contents it had in its origins.

Very similar to the MLAT, but simplified and adapted to younger children, the MLAT-Elementary (MLAT-E) was published in 1967. This test, not so popular as its parent, is attested by the extensive use of the MLAT, although in this case not many published pieces of research certify its usefulness or validity. Indeed, the study of aptitude in young learners is, on the whole, rather scant, although some other measures have been designed to measure aptitude in young learners.

After describing in depth the MLAT (section 2.2.1) and how it was originated (section 2.2.1), this chapter includes the statistical information and the standardised norms that appear in the test's *Manual*, which served as the basis for the design of the MLAT-E (section 2.2.3). The MLAT, along with the different adaptations and translations, has been the aptitude battery most widely used in SLA research (section 2.2.6), yet several factors, such as its rather oldish norms as well as its supposed ineffectiveness to determine FL success in some FL learning contexts, have been questioned (section 2.2.7).

This chapter continues with the description of the MLAT-E (both in English and Spanish) and the process of their creation, along with the statistical information available that support the use of these aptitude tests along the years (sections 2.3.1 and 2.3.2). Other than in the norming studies, the MLAT-E or the MLAT-ES have not been used much for research purposes. Those studies in which the MLAT-E or some derivate have been used are reviewed in section 2.3.3. In that section other uses of the

MLAT-E that also apply to the MLAT-ES are also described. From the little research available, and as FL instruction has inevitably changed with time, some improvements to these tests can already be suggested. Indeed, it is a must to revise the already existing tests as the constructs they are meant to measure may not be as relevant as they were the very moment they were created due to the changes in language teaching methodology. These and other questions appear in section 2.3.4.

The MLAT-E and the MLAT-ES are not the only tests that have been designed to measure aptitude in young learners, although there are not many more. The measures designed to be used with young learners as well as the theoretical principles on which they are grounded are exposed in section 2.3.5.

This chapter finishes with the justification for this dissertation, explaining the situation that makes it worthwhile and the general objectives of the study, stated in three main research questions.

2.2. Measuring language aptitude in adults: developing the MLAT

2.2.1. Carroll and Sapon's preliminary study

The MLAT was constructed on the basis of a model which derived from the variables remaining from the results of factor analyses (Carroll & Sapon, 1955; Carroll, 1958) of a large number of individual characteristics believed to contribute to L2 learning in an audiolingual methodology context. The battery of tests designed at this stage was administered to two groups of learners of Mandarin Chinese as a FL at the beginning of the course. Then a factor analysis was run in order to eliminate those variables that were redundant. Finally, intercorrelations were run between the remaining aptitude tests and some end-of-course language achievement and proficiency tests.

From this preliminary study seven factors could be easily identified, although Carroll (1958) only dared to label five of them: verbal knowledge, sound-symbol association ability, immediate rote memory for FL vocabulary, grammatical sensitivity or syntactical fluency, and inductive language learning ability of artificial language structure. The two remaining factors were not labelled for different reasons. One of them (Factor B) was, as Carroll himself admits, difficult to interpret. He even hinted that factors may not have been matched as they should have been. Factor G did not receive a name of its own either, because it was found not to be closely related to language learning and thus, irrelevant for research into FL aptitude purposes.

Analysing the weight that each factor had as compared to the others, Carroll could identify the most relevant factors that exert some influence on FL learning. Surprisingly, verbal knowledge was not found to be as salient as the rest. This does not necessarily mean that it is not relevant at all. It should be taken into account that the subjects in this preliminary study were at the first stages of learning a FL. Consequently, verbal knowledge may not have proved to be important at this stage but could obviously be so at later stages, when the comprehension of difficult vocabulary materials of abstract discourse is involved in the language learning process. That is why this factor was kept in further research despite not seeming relevant in this preliminary study.

Having taken into account the loading of each factor, Carroll (1965) finally considered that the MLAT should contemplate four relatively independent underlying factors: phonetic coding ability (auditory capacity and sound-symbol relations), grammatical sensitivity, associative memory, and inductive language learning (for a definition of each factor, see section 1.4). However, inductive language learning ability is not represented purely in any of the five subtests of the by-product of these series of factor analyses, the MLAT (see section 1.4.2).

2.2.2. MLAT final version: description

The MLAT consists of the five subtests relatively uncorrelated and with consistent validity which turned out useful after the try-outs carried out during the preliminary study summarised in the previous section. These are:

Part 1. Number Learning: This subtest is aimed at measuring rote memory and auditory alertness. In this part, the names of numbers are taught in a new artificial language. Then, test takers have to write down the series of tape-recorded numbers. It has 43 items.

Part 2. Phonetic Script: Scores on this part are to be considered a measure of sound-symbol association ability, i.e. the ability to learn correspondences between speech sounds and spelling. It may also show some kind of relationship with memory for speech sounds and it also shows high correlations with mimicry of FL speech sounds and sound combinations. This part may come in especially useful to detect FL problems in dyslexics. This part has 30 items.

Part 3. Spelling Clues: Besides measuring sound-symbol association ability up to a certain extent, the results on this highly speeded subtest depend on the test taker's vocabulary knowledge, as the correct synonym of a disguised word has to be chosen from the five choices suggested. The maximum score possible is 50.

Part 4. Words in Sentences: It is intended to measure grammatical sensitivity. As not all schools teach grammar explicitly or focusing on forms and/or using metalinguistic terms, Carroll designed this test using analogies in which subjects had to identify which of the components underlined in one sentence corresponded to the highlighted element in another sentence in terms of sharing grammatical function. Carroll (1990:19) claims that "there is nothing in the test that requires technical knowledge of grammatical structure or terminology". Indeed, when acquiring one's L1, we learn the grammatical structure of our mother tongue implicitly; it is only at school that we become aware of the grammatical structure of our L1. Because of this, it remains unknown how much scores on this part depend on grammatical training. Yet the degree of grammatical awareness is supposed to vary depending on our aptitude for learning information about grammatical structure, no matter the amount of exposure to grammatical instruction we receive. This part has 45 sentences in which the word in capitals is to be matched with its analogous in the sentence underneath.

Part 5. *Paired Associates*. It is intended to measure associative rote memory capacity by making the test taker memorise as many words in Kurdish as possible. These words are presented with their English equivalents. Later their equivalent is to be chosen from among five other possible equivalents. The maximum score is 24.

The MLAT can be administered in two ways: the complete test, with 192 items, which takes approximately from 60 to 70 minutes and for which a tape recording is needed; and its short form, which takes about 30 minutes and for which the cassette tape is not needed. However, using the tape is highly recommended so that administration conditions (i.e. instructions delivery and test-doing timing) are kept consistent in all administrations. The short form consists of Parts 3, 4 and 5. For both the short and long form, each student should be supplied with a test booklet, an answer sheet, a practice exercise sheet and a pencil.

All parts but Part 1 follow a multiple-choice format in which test takers have to tick only one answer. If they choose more than one option, the item is invalidated. In Part 1 Number Learning, what is taken into account is accuracy in writing the numbers from dictation. If the examinee fails to transfer the answers to the answer sheet, but has written them in the booklet, they are to be taken into account anyway.

2.2.3. MLAT standardisation and norms

Approximately 1,900 students between grades 9 and 12 and 1,300 college and university students took part in the standardisation stage as well as three adult groups. These were a group of 177 Air Force enlisted men learning Mandarin Chinese, 77 men enrolled in an intensive course at the Foreign Service Institute of the Department of State and a third group of 781 men who were learning a FL at the Army Language School. The requirement to be a valid subject in this study was to be novice at learning a FL. Tables of percentile norms and information about the performance of students at various grades levels can be consulted in Carroll and Sapon (1959) as well as in the edition by the Second Language Testing, Inc. (2000). The population appears divided taking into account their grade levels, sex and language of enrolment. Students at the Army Language School were the ones who obtained the lowest mean scores while men in intensive language training and men in their fresher year at college obtained the highest means. Regarding the test takers in grades 9 to 11, males appear to consistently obtain lower means in all parts but for one group in grade 9. Some of the data in this table, Carroll warns, should be taken with caution, as some of the groups are too small so as to generalise the results.

2.2.4. MLAT statistical information

The concurrent validity of the MLAT was established using criterion measures such as the students' actual performance in the FL as measured by the teachers' marks and ratings or standardised language proficiency tests. These measures were chosen on the grounds that they were supposed to correspond to what the MLAT claims to measure. For grades 9 to 11, the validity coefficient ranged between *r*=.25 and .69 in its long form and from *r*=.21 to *r*=.83 in its short form. For college students, the validity coefficient went from .13 to .69 in its long form and from .21 to .68 in the short one. As Carroll explains, lower validity coefficients appeared in combinations of language groups and sex groupings where the N of such groups was too low to be computed separately. Besides, in comparison to the validity coefficient obtained from those students in intensive courses, the validity coefficients in the high school and college groups is possibly more variable, Carroll presumes, due to the kind of language training, the pace of the courses, and the motivation these students had. Furthermore, Carroll casts doubt on some teachers' marking system that may have affected the validity of the MLAT.

On the other hand, the highest coefficients were found in those groups in which students were following an intensive course, as the subjects would devote most of their time to language training, which could have had an influence on their performance in the MLAT test. Except for one group for which the correlations were low (*r*=.27 in the total test and *r*=.26 in its short form), the other correlation indexes ranged from *r*=.42 to *r*=.73 in the long form and from *r*=.35 to *r*=.69 in the short one. Nevertheless, the longer complete test appears to consistently have a higher validity for these courses and may also be so for secondary school and college courses.

The *MLAT Manual* also offers expectancy tables illustrating the relationship between MLAT total scores and course marks in specific languages, thus giving us the probabilities with which subjects obtaining specified scores on the test may be expected to attain a certain level of achievement in particular FL courses.

Validity coefficients are found not to be very much affected by the FL being learned. Although the highest validities are to be found in those individuals learning European languages using the Roman alphabet, the other two groups of languages involved (both non-Indo-European and Indo-European languages not using the Roman alphabet) obtained higher and lower validities alternatively in different groups of subjects. This, though, does not override the fact that the larger the distance between one's L1 and the FL learned, the higher the difficulty to learn the FL. These coefficients were not affected by the teaching methodology used in the courses the subjects of the study followed. High validities were obtained both in “intensive” courses, in which oral work was emphasised, and in “traditional” courses, in which the stress was laid on grammar and translation.

Reliability data were obtained by running the split-half technique. This technique involves the division of the items into two halves, making sure that each half is matched in terms of item difficulty and content. For reliability to be high, the scores obtained in each half should be highly correlated to one another. When applying this type of reliability statistical analysis, it is assumed that the test administered can be split into two matching halves. Otherwise, the items should be matched according to whether they appear in an even or in an odd position in their test appearance. The split-half technique is also appropriate for tests that are not speed but power measures, which is the case of all parts in the MLAT but Part 3 Spelling Clues. For this part, what was measured was not internal consistency as such. Instead, a correlation corrected with the Spearman-Brown formula was run after splitting the test into two sections and making the subjects take them separately. The reliability of the whole test was calculated both including and overriding Part 3 with no significant differences in the final reliability measure obtained. In the end, both the short and long version of the test proved to have consistent high reliability coefficients, ranging from *r*=.83 to *r*=.93 in the short form and from *r*=.90 to *r*=.94 in the long one.

The reliability coefficients of the subparts tend to be slightly lower. The one test that is particularly salient for not reaching a desirable reliability coefficient by itself is Part 3 Spelling Clues. The highest reliability coefficient reached is *r*=.80 for girls in grade 9. The lowest coefficients are for boys in grade 9 (*r*=.55), for girls in grade 10 (*r*=.67) and for the men in the Air Force group (*r*=.60). In addition to this, Carroll (1990) himself admits that this part usually presents a negatively skewed distribution, which is the statistical information that allows us to say that the test is too easy and, therefore, not discriminating enough at upper levels of ability. He also considers the possibility of lengthening the tests so that the reliability coefficients among the subtests are higher. However, making the tests longer would diminish the practicability for testing due to time constraints.

The intercorrelation of parts was also run so as to check that none of the subtests was redundant. The coefficients obtained were low enough to keep all the subtests in the MLAT as it has remained until present.

2.2.5. Other versions of the MLAT

The MLAT is usually administered in the subjects' L1, although it is meant to measure FL learning ability in any language, which does not have to belong to the same linguistic family as the subjects' L1. Actually, the subjects who participated in the validation phase of the MLAT were learning different foreign languages, some Indo-European languages using the Roman alphabet (Czech, French, German, Spanish, Latin, Polish and Romanian), some Indo-European languages not using the Roman alphabet (Bulgarian, Greek and Russian) and some non-Indo-European ones not using the Roman alphabet (Chinese, Japanese and Korean) .

The relevance of the MLAT in the study of SLA has been (and still is) such that it has been adapted and/or translated into several languages such as Italian (Ferencich, 1964), French (Wells, Wesche & Sarrazin, 1982), Japanese (Murakami, 1974), Hungarian (Ottó, 1996), Spanish, Turkish, Indonesian and Thai. In Sasaki's (1996) Language Aptitude Battery for Japanese (LABJ), there is a translation of the Paired Associates subtest from the MLAT and other tests inspired in other subtests of the MLAT. According to Stansfield and Winke (2008), of all these translations, only a few of them (French, Japanese and Hungarian) can be located at present and only the French version is commercially available. They also inform that "Hebrew, Polish and Chinese versions are currently being developed for research purposes" (Stansfield & Winke, 2008:83).

The MLAT can also be used in its shortened version which, despite omitting parts I and II, should yield similar results. However, this does not always hold true, as it happened in the case study by Steinman and Smith (2001). Besides, individual parts of the MLAT are often used to adjust to research purposes (e.g. Erlam, 2005; Harley & Hart, 1997; Roehr, 2008). Carroll intended to construct and standardise an alternative form, but he did not manage to accomplish the task (Carroll, 1981).

It has also been adapted for children (the MLAT-Elementary - MLAT-E) and there also exist German (Correll & Ingenkamp, 1967) and Spanish versions of the MLAT-E, the MLAT-ES, presented in March, 2004, at the 26th Annual Language Testing Research colloquium held in Temecula, California, as well as another version adapted for blind people, developed by R. Gardner (1965). At present, the SLTI is developing an MLAT-E in Korean and a computer version of the MLAT (Stansfield & Winke, 2008).

2.2.6. Use of the MLAT

Despite its flaws and the criticism that it has received (see section 2.2.7), the MLAT has been widely used since its release. Its main use has been the prediction of any individual's success in learning a FL in a given amount of time and under given conditions. The results obtained should not be assumed to be informative of any facet other than the way this ability is involved in how well a FL is learned. That is, the MLAT cannot be used to infer whether this ability has been affected by any external factors such as previous language training or inheritance. In fact, Carroll (1959/2000:21) suggests that “since it is likely that as grade level increases there is more experience with foreign language training, the failure of the medians (in the median scores for three of the upper norm groups) to show greater change suggests that language training has little effect on scores”. Yet what is unquestionable is that previous FL learning experience is a factor which should not be completely disregarded. Moreover, if any of the abilities has been trained in some way, this could also affect the score obtained in the MLAT, though not necessarily transfer to language learning itself (Carroll, 1971a).

The MLAT has been used to predict different types of FL learning objectives under different methodologies and contexts. Whatever the teaching methodology, in principle, the validity of the test has been upheld. As mentioned in 2.2.4, the MLAT proved to be valid for both oral-work oriented and grammar-oriented courses, yet not enough information is available as to the actual orientation of these courses. This validity was reinforced by the fact that no different levels of achievement were found when comparing students who were following a course based on grammar-translation methodology and those who were in intensive courses aimed at reinforcing speaking skills (Agard & Dunkel, 1943; in Carroll, 1959). Nevertheless, the predictive results of the MLAT cannot be applied, Carroll (1959) supposes, when FL learners are at an advanced level, as the correlation between oral and written performance may appear to be weaker at some point. That is, the MLAT is a powerful test to predict rate of acquisition of only the basis of a FL.

Researchers have made use of the MLAT in different learning contexts, such as focus-on-forms classrooms, communicative classrooms and laboratory learning contexts. Some subtests have proved to be more informative than others depending on the contexts, for instance, the subtest measuring grammatical sensitivity has been considered especially useful for predicting FL ultimate attainment and critical period effects in SLA.

Following its release, the MLAT was found to be a predictor of success in several studies conducted in form-oriented classrooms. Besides the unpublished studies that Carroll (1981) reports, Gardner and Lambert (1959, 1965, 1972) also report the findings from studies carried out in French classes in the US in which measures such as IQ tests, motivation questionnaires and L2 achievement tests were distributed along with the MLAT. The overall results were that both aptitude, especially the Words in Sentences subtest, and an IQ factor were strong predictors of FL achievement as well as general and academic achievement outside the FL class. Later on, Gardner et al. (1976) confirmed that aptitude is more strongly bonded to class marks than to communicative skills, with which the correlations showed to be weaker. Moreover, aptitude was more strongly related to performance in those learners who were at an advanced FL level, which could be a sign of aptitude being influenced by FL learning experience. In a piece of research carried out by Bialystok and Fröhlich (1978), Words in Sentences was once again found to be responsible for most of the variance on the grammar and reading tests. However, it was not so powerful when related to the listening test, which was the one which required the least explicit grammatical knowledge of the FL.

The MLAT has also been revealing in communicative classrooms. Ranta (2002) found high verbal analytic ability to be useful for 6-graders in communicative language learning programmes while this ability proved lower in learners who were not as successful. Ranta concluded that language analytic ability is not neutralised in communicative language programmes, as this ability, together with strategic competence, was indeed helpful.

Reves (1983, in Ranta 1998) also studied the role of aptitude, motivation, cognitive style and learning strategies in formal and informal situations. He found that, in informal situations, what was the most effective predictor of grammatical accuracy, oral fluency and course marks in both Hebrew (the community's L1) and English (the language taught formally) was aptitude as measured by an Arabic adaptation of the Number Learning from the MLAT. She also used an Arabic adaptation of the Words in Sentences, but it only explained a small part of the variance on the ratings of oral accuracy in Hebrew and in English and on the English final mark.

Horwitz's (1987) hypotheses were that social cognitive abilities would be closely related to the learners' communicative competence while aptitude would be associated with grammatical competence instead. Her hypotheses proved to be true only up to a certain extent: the MLAT, especially the Words in Sentences subtest, correlated moderately and significantly with the grammar test as did the test measuring social cognitive abilities with communicative competence. However similar correlations were also found between aptitude and communicative competence as well as between social cognitive abilities and grammatical competence.

In a study whose scope was not only aptitude but other IDs, Ehrman and Oxford (1995) found that the MLAT and a faculty rating questionnaire were the measures that correlated best with speaking and reading proficiency measures. The subjects who took the MLAT were 282 US government employees who were taking part in an intensive course which blended communicative and audiolingual approaches to language teaching. It is worth mentioning that their mean score in the MLAT was one standard deviation higher than the mean for the norms in the *MLAT Manual* (Carroll & Sapon, 1959), probably because the population in this study was highly educated. This means that previous training or education could certainly have some effect on aptitude scores.

DeKeyser (2000) found that those adults scoring high in the Words in Sentences were the ones who also scored in the same range as did early starters. DeKeyser concluded, therefore, that adults should receive some kind of explicit focus-on-form instruction so as to assure success in their FL acquisition process. Erlam (2005) also found that learners with both high analytic ability and working memory capacity benefited most from this approach as shown in writing proficiency test scores.

In addition to its uses in research to elucidate the construct of language ability, the MLAT has also been used for diagnosis and to stream and match FL learners to the curricular option that would be more convenient for them by taking into account the score they obtain in the MLAT subtests and so outlining their aptitude profile (see section 1.5). On a similar note, FL learning disabilities (FLLDs) have also been detected by using the MLAT along with other measures. One such study is that of Gajar (1987), who administered the MLAT to both regular and language-disabled university students, the latter scoring below the former on all subtests, principally in parts 4 and 5. On this same line, Sparks, Ganschow and colleagues have widely used the MLAT to detect FLLDs (see section 1.6.1). Thanks to this application of the MLAT, FL teaching programmes can be tailored to fit the students' needs by adapting the pace of materials presentation, addressing the students' weaknesses and fostering their strengths or by carefully selecting the teacher in charge of a course to meet the students' needs.

Other suggestions for FL programme accommodations are more radical and what they recommend is substitute courses for students with FLLDs or other at-risk students (Shaw, 1999; Stansfield & Winke, 2008). Sparks, Javorsky and Ganschow (2005) expressed their disagreement with this idea, as FLLDs could be wrongly diagnosed due to misinterpretations of the FL aptitude concept. These are caused mainly because of having disregarded the Carroll model of school learning, which also considers both instruction and individual learning differences other than aptitude, including intelligence and motivation. Actually, some students diagnosed with a FLLD have never been given the chance to even start a FL course (e.g. Sparks, Philips & Javorsky, 2002, 2003), so their rejection from enrolment in a FL course is, perhaps, unfounded. In fact, students with LD with similar MLAT scores have actually obtained different outcomes in FL courses. Apart from that, students diagnosed with a LD may drop out before the completion of the course because of, for instance, their lack of persistence and motivation, not of their LD strictly speaking (Sparks et al., 2002).

As dyslexics have great difficulty in succeeding in Part 1 Phonetic Script, this part can be used to detect future difficulties related to dyslexia. Former dyslexics could also face some difficulty because of their misperception of language segmentations and their correspondences with graphemic symbols. According to Carroll (1990:17), "(T)hese difficulties carry over into foreign language learning activities – mimicking sounds accurately, learning the segmentation and spelling of foreign words, and controlling the order in which phonemic units are uttered. This can be one reason why phonetic coding tests turn out to be highly valid in many foreign language learning situations." Notice that Carroll says "many", not "all of them". Indeed, Wesche (1981) found that phonetic coding was necessary for success in the audiovisual method but not so much in the analytical one and not all dyslexics are so due to auditory-phonetic causes, as some neuropsychology studies have shown (Carroll, 1990).

To sum up, the MLAT has repeatedly proved to be a very powerful measure on its own. However, it is highly recommended not to use it as the only measure but along with other measures such as FL or L1 assessments or the FL learning history of the learner. Also, we should always bear in mind not the total score, but the scores obtained in each subpart as well as the relationship these scores establish with the instructional context and, naturally, other FL IDs such as age, motivation, anxiety, learning styles, learning strategies and personality.

2.2.7. Criticism towards the MLAT

Despite its widespread use, the validity of the MLAT has been questioned on several occasions for several reasons in relation to the design of the test itself and to the interpretation of the results obtained when using it. When the MLAT was developed, in the late 1950s, it proved to be a valid and reliable measure, but both the conceptualisation of aptitude and the learner populations to which it was initially administered have changed over the years, while the MLAT has remained the same. Actually, norm samples of tests are estimated to have a validity of approximately 15 years (Salvia & Ysseldyke, 2003), that is, the norms of the MLAT expired more than 40 years ago, which could certainly threaten the interpretation of the analysis carried out when using it at present, although Ehrman (1998) found similar validity coefficients to those from 1958.

Regarding the test design, Carroll (1990:12-13) himself remarks that he would have liked to find the time to create at least one alternative version of the test. In fact, in an unpublished paper Sawyer (1993:4) affirms that “one serious problem concerning the MLAT’s reliability is that it exists in only Form A; a Form B or beyond was never developed”, and he continues: “test-retest reliability estimates cannot be performed (...) since the aptitude measure can be administered reliably one time”. Sawyer also remarks that, as a consequence of the fact that it is only possible to administer it once in normal conditions, it is very difficult to state that aptitude, as measured by the MLAT, does not increase along with language achievement. Carroll (1990) also mentions some other minor design defects that could be easily fixed. For instance, the numbers in the Number Learning subtest bear an “unfortunate correspondence” with the alphabetical order of their names and in Part 3, Spelling Clues, the instructions, apart from being a bit obscure for some test takers, do not sufficiently emphasise that it is a speeded test. These flaws, he admits, may also be present in Parts 4 and 5. Shaycoft (1965) also mentions that the directions do not instruct explicitly whether or not the students are penalised for answering at random.

Pencil-and-paper tests are still in use at present. However, it cannot be denied that the data collected in any study are much more user-friendly for the researcher if they come in a computerised way. In order to solve this handicap, the SLTI are working in a computerised version of the MLAT.

Besides the format faults mentioned above, when reviewing the MLAT, Carroll (1990) mentions ways in which it could be improved or complemented regarding its use for specific purposes. For instance, the Phonetic Script test should be complemented with tests that measure the underlying factors in this subtest, such as general intellectual ability and memory for phonetic material. Regarding the Words in Sentences test, as it has been criticised for being closely linked to explicit grammatical knowledge, Carroll suggests that scores on this test be compared to tests of formal grammatical knowledge and terminology which would confirm or refute the criticisms it has received. Carroll also warns that Part 5 Paired Associates should not be taken as a measure of general memory but of a special kind of rote-learning ability. On top of that, he admits he has never been too confident about its validity, as it ranged from zero to rather substantial validity. Nevertheless, he decided to include it in the final version of the MLAT because he believed it would be useful in some foreign language contexts. Unfortunately, Carroll does not state which ones. Undeniably, results in both Part 5

Paired Associates and Part 1 Number Learning, which is also thought to measure memory somehow, should be contrasted with other measures of memory.

Inductive language learning ability is also included among the factors that shape aptitude. However, this ability is not measured as such in the MLAT. It is tapped only weakly in Part 1 Number Learning. Carroll (1990) mentions he had actually designed one such test, but it was too difficult to administer and so it was not included in the final version of the MLAT. He encourages anybody willing to further investigate this ability and use the materials he created with Sapon (Sapon, 1955) to administer it so as to complement the MLAT with this measure that it lacks at present.

The generalisability, reliability and scope of the results obtained have also been a target of criticism towards the MLAT which dates back to the very piloting study. To start with, it has been suggested that the MLAT only predicts reading and writing, not speaking performance, as oral ability involves much more than the ability of learning sound and grammar systems (R. Ellis, 1986). Actually, Brecht, Davidson and Ginsberg (1993, 1995, in Ehrman, 1998) did not find the MLAT predictive of overall oral proficiency — though the test was predictive for reading proficiency in language training in Russian — nor did they report any relationship between results in the MLAT and oral gains in a study abroad context.

Although, as mentioned in 2.2.6, no significant differences in the validity were found concerning the focus of the teaching methodology that the students were following, there remains a tinge of doubt whether, at the time of its piloting, the differences in methodology were such or not. Besides, at no moment was the test takers' L1 considered, which could also have affected the MLAT predictability results (Fisher & Masia, 1965). Surprisingly, the MLAT has continued to be used until present days despite the remarkable changes that have been made in both teaching methodology and language testing. Consequently, the validity and reliability of the results obtained when using it may not be as trustable as they used to.

The population used for the piloting study, besides being rather small and at times not comparable from grade to grade (Shaycoft, 1965), was made up of only native speakers of English, let alone the fact that the test used a natural language, English, and where an artificial language was used, it was inspired in English as well, which led Fisher and Masia (1965:635) to conclude that what the MLAT measures is "the student's ability to recode English" and, therefore, the student's ability in their L1, not the student's ability in learning a FL. No information is given about whether the participants in the piloting spoke or not more languages other than English either, which is a factor which should not be overlooked, seeing that bilingualism seems to be beneficial for the acquisition of FLs in some studies (see section 1.6.5).

It is claimed that the MLAT can be used both in its long and its short form, consisting of Parts 3, 4 and 5. The use of the short form is plausible, as intercorrelations among the parts are low enough, which suggests that they measure different aptitudinal aspects, and partial reliability is also high enough to be trusted. However, neither the correlation between the short form and the long one nor norms or validity coefficients are present in the *Manual*. That would further warrant the use of the short form (Shaycoft, 1965).

The use of the MLAT to identify FLLDs (see section 2.2.6) and thus, to select or reject students for FL study programmes has also been called into question. Even if the main aim in using it would be detecting students who do not seem to present a natural endowment for FL learning ability and also those whose aptitude is suitable for FL learning, it should not be forgotten that the performance in this test could always be faked if the test taker is not interested in learning a FL (Reed & Stansfield, 2002, 2004). On the other hand, if low aptitude is interpreted as slow learning, a low score in the MLAT is not decisive to turn down FL learners, as low FL rate could be overcome by making an extra effort to learn the FL (Goodman, Freed & McManus, 1990). Despite using the MLAT in their research into FLLDs, Sparks, Javorsky and Ganschow (2005) do not relate this disability to a lack of FL aptitude, but to a lack of L1 mastery as well as cognitive skills. On top of that, this research team also criticises the MLAT for having outdated norms which should be renewed.

Using other measures or repeating the testing is advisable in order to fully explain unexpected scores on the MLAT. Apart from that, since other individual factors are required to fully succeed in FL learning besides FL aptitude, any score in the MLAT alone, no matter whether high or low, should not be considered to be a decisive measure to decide upon the acceptance or rejection of anybody’s enrolment in a FL course. Instead, it should be used along with other tests and questionnaires involving age, intelligence, styles and strategies and especially motivation, which is even thought to override the effect of aptitude (Dörnyei, 2005). Having access to other concomitant data of not only the individuals themselves but also their learning environment is especially important regarding the detection of FLLDs, whose diagnostic can certainly involve other factors and so the MLAT score may as well be of incidental interest only (Sparks & Javorsky, 2000).

The MLAT has been proved to be a reliable measure to predict language aptitude to learn languages that use the Latin alphabet. However, it remains unknown whether its predictability power is such when it comes to learning languages which use other writing systems, such as the logographic or the syllabic, among others. Besides, the MLAT itself uses the Latin alphabet and is to be taken on the test taker's L1.

Despite the amount of criticism the MLAT has been accumulating along the years, it is only recently that new tests have started to be tried out, such as CASL's HiLAB, (see section 1.3) although this test is aimed at measuring highly-skilled FL learners and not the average FL learner. It is, perhaps, time to design new aptitude tests that are as reliable and long-lasting as their predecessors.

2.3. Measuring language aptitude in young learners: the MLAT-Elementary (English and Spanish versions)

Several versions of the MLAT have been designed and validated and further development of versions in other languages is still being carried out. This is not the case, though, of the MLAT-E, which has received far less attention from researchers than the MLAT. Carroll (1981) explains that, thanks to the support of the Carnegie Corporation in the 1960s, he and his colleague Sapon could develop an adaptation of the MLAT, the MLAT-Elementary, for children in grades 3 to 6 with selection, guiding and placing purposes in the Foreign Language in the Elementary School.

In the early 2000s, the Spanish version of the MLAT-E (from now on MLAT-ES) was developed by the SLTI, which is a reputed test development company with a wealth of experience in the field. The subtests are exactly the same as the ones of the English version, so all the abilities that the MLAT-E in English is meant to measure are supposed to be tapped in this and other subsequent adaptations made in different synthetic languages which use the Latin alphabet. If constructing test items is a task fraught with difficulty, as "being able to draw valid and reliable inferences from a test's scores rests in great measure upon attention to the construction of the items or exercises that comprise it" (Osterlind, 1989:1), test translations and adaptations are not easy ventures either. In the process, both intrinsic and extrinsic factors to the test are to be taken into account. Some of the extrinsic factors, such as cultural background, the changes in teaching methodology since the first version of the MLAT-E was normed, or the test takers' native language, among others, will be broached in the discussion of this dissertation.

The intrinsic factors related to the translation and adaptation of tests include, certainly, keeping the test format and thus ensuring that the constructs measured in the resulting version are the same as the ones in the original version. Another intrinsic factor is the target language and its regional varieties. This factor is crucial in tests that are language-based, as are the MLAT-E and the MLAT-ES. English and Spanish are all languages that use the Latin alphabet. Yet they come from different Indo-European branches and do differ in aspects such as the correspondence between sound and grapheme, especially relevant in Parts 1 and 3 and in the fact that Spanish is mainly synthetic while English is more analytic as far as linguistic typology is concerned. This affects especially such aspects as the strictness of word order in the sentence, which can be relevant in Part 2.

In the following sections the reader will find the description of both the MLAT-E (section 2.3.1) and the MLAT-ES (section 2.3.2) taking into account the intrinsic factors mentioned above, which will also be tackled in subsequent chapters, as well as the statistical data of the published norming studies of both tests.

2.3.1. MLAT-E English version (MLAT-E)

The MLAT-E owes its theoretical framework to the MLAT. The theoretical assumptions of the adult version, explained in section 2.2.2, were extrapolated to design the subtests of the MLAT-E, with the structure of three of its subtests identical to three of the five subtests of the MLAT. Obviously, they were rendered easier so that they could be taken by younger examinees, as explained in the *MLAT-E Manual*, which is the main source used for describing its four parts (see appendix A for an overview of sample items on the MLAT-E obtained from the SLTI website).

2.3.1.1. MLAT-E: description and purpose

Part 1. Hidden Words: This part, which contains 30 items, corresponds to Spelling Clues of the MLAT, but presents a less difficult vocabulary. It measures not only knowledge of the English vocabulary, but also sound-symbol association ability. Phonetic Script, used in the MLAT to measure these abilities along with memory for speech sounds, was not retained in the MLAT-E because it was found to be too difficult at the lower level.

Part 2. Matching Words: This was called Words in Sentences in the MLAT. Although it is designed to measure sensitivity to grammatical structure, the terminology of formal grammar is not used. Taped instructions and examples teach the pupils to recognise the job that a particular word does in a sentence and to find in another sentence the word that does a similar job. It has 30 items.

Part 3. Finding Rhymes: This is a part which was not in the MLAT. On it, an attempt is made to measure the ability to hear speech sounds by asking the examinee to select words that rhyme. It has 45 items.

Part 4. Number Learning: As in the MLAT, the test taker learns the names of numbers in an artificial language, and after some practice in recognition and in putting numbers together, the test taker listens to 25 numbers in the new language and writes them down. This part aims to measure the memory component. At the higher level, in the MLAT, it was found that “the part also has a fairly large specific variance, which one might guess to be a special ‘auditory alertness’ factor which would play a role in auditory comprehension of a foreign language.” (Carroll & Sapon, 1959).

Part 1 Hidden Words corresponds to the Spelling Clues subtest in the adult form and is meant to measure phonetic coding abilities. These abilities are, though, so close to spelling abilities, as show the correlations with batteries used to predict phonetic coding, that they may actually be identical to them. Consequently, as Carroll (1993) concludes, they could be measured by dictation tests or misspelling-recognition tests. This would also imply that phonetic coding would make use of implicit knowledge of conventional spelling rules and phoneme-grapheme correspondences, and that spelling ability does not involve immediate memory for visual forms of words. Therefore, more research into this relationship is needed to determine if phonetic coding is a characteristic of auditory-visual memory or “merely a reflection of individual differences in the learning of grapheme-phoneme correspondences” (Carroll, 1993:174).

Part 2 and 4 in the MLAT-E function exactly the same way as they do in the MLAT for adults. Therefore, it could be assumed that they are to be used to measure grammatical sensitivity and memory respectively, as demonstrated by the use of these parts of the adult version in several studies.

While the construct underlying Parts 2 and 4 can be supported by the design and subsequent use of these subtests in research, the author of this dissertation believes there is no source available that details how the Finding Rhymes subtest, which is the one created especially for the Elementary version, is supposed to function. It could be inferred that, since there is no correspondence between English orthography and its sound system, examinees are expected to process a number of different sound-symbol correspondences, which has a straightforward relation with phonetic coding ability.

The directions of all the parts include examples which allow the test taker to become familiar with the test procedure. These directions are pre-recorded on a cassette or CD recording, which also determine the time allotted for each part. It takes 61 minutes to administer the test, including instructions. To these 61 minutes, one must add the time it takes to hand out the test booklets and fill in the identifying information of the test taker. The adult version is heavily speeded and, according to the *Manual*, so is the Elementary version. Surprisingly, in the *Manual*, Carroll points out that "a statement was added in the test manual to instruct test administrators to remind students to work carefully but quickly" (Carroll & Sapon, 2002:11) because just one child did not have the time to complete the first part of the test and most of the answers she had given were correct.

2.3.1.2. MLAT-E: standardisation and norms

More than forty years have passed since the MLAT-E was standardised. In this norming study, more than 4,000 pupils (approximately 1,000 in each of four grades, from 3 to 6) participated, of whom two-thirds were receiving some FL instruction (mainly in French or Spanish). They attended either public or parochial elementary schools. The present situation regarding FL teaching methodology, types of learners and resources is different, so the context surrounding the norming process should never be overlooked when interpreting these results.

The *Manual* provides the percentile norms table for raw total scores of the test by sex and grade (see Table 2.1). The means and standard deviations for this sample can also be found in this table.

Table 2.1. Norms for students in grades 3, 4, 5, and 6 on the MLAT-E. Raw total scores corresponding to designated percentiles (adapted from Carroll & Sapon, *MLAT-E Manual*, 2002:6)

Percentile	Grade 3		Grade 4		Grade 5		Grade 6	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
99	111-130	112-130	120-130	121-130	123-130	124-130	126-130	126-130
97	105-110	107-111	115-119	118-120	120-122	122-130	124-125	125
95	98-104	105-106	109-114	114-117	116-119	119-121	122-123	123-124
90	92-97	94-100	104-108	110-113	113-115	116-118	120-121	121-122
85	86-91	89-93	100-103	106-109	109-112	113-115	118-119	120
80	81-85	84-88	96-99	102-105	106-108	111-112	116-117	118-119
75	78-80	81-83	93-95	99-101	104-105	109-110	115	117
70	74-77	77-80	89-92	96-98	101-103	107-108	113-114	115-116
65	70-73	73-76	86-88	94-95	98-100	104-106	111-112	113-114
60	66-69	69-72	84-85	91-93	96-97	101-103	109-110	112
55	63-65	66-68	81-83	89-90	93-95	99-100	107-108	111
50	59-62	63-65	77-80	86-80	90-92	97-98	104-106	109-110
45	56-58	60-62	74-76	82-85	87-89	94-96	101-103	107-108
40	53-55	57-59	70-73	79-81	84-86	91-93	99-100	104-106
35	49-52	54-56	66-69	76-78	82-83	89-90	95-98	102-103
30	44-48	49-53	62-65	72-75	79-81	86-88	90-94	99-101
25	40-43	45-48	58-61	68-71	75-78	82-85	85-89	96-98
20	36-39	40-44	53-57	62-67	71-74	78-81	81-84	91-95
15	30-35	35-39	48-52	56-61	64-70	73-77	74-80	86-90
10	24-29	30-34	40-47	48-55	58-63	63-72	65-73	76-85
5	21-23	26-29	31-39	36-47	47-57	53-62	55-64	59-75
3	17-20	22-25	26-30	27-35	37-46	44-52	41-54	45-58
1	0-16	0-21	0-25	0-26	0-36	0-43	0-40	0-44
N	493	528	505	510	495	500	670	640
Mean	61.1	64.4	76.3	83.5	88.9	94.7	99.5	104.5
SD	24.7	23.6	23.3	22.8	20.6	19.2	20.4	18.1

From the information in this table, it can be seen that girls were found to consistently obtain higher scores than boys in all grades. Consequently, sex may be an issue worth taking a look at when studying aptitude as measured by using the MLAT-E. Surprisingly, though, at no moment is this difference in scores regarding sex mentioned in the *Manual*.

2.3.1.3. MLAT-E: statistical information

The *Manual* of the MLAT-E supplies tests users with detailed and valuable information about the validity, standard error of measurement, reliability and intercorrelation of parts of the test.

The data available on validity are based on the results of seven schools and were obtained by comparing the scores with performance as measured by course marks or the FL teacher's criterion. These data are considered concurrent validity, as the criterion measures were obtained only between two and three months after the administration of the MLAT-E. However, they could be considered data of predictive validity on the grounds that the data obtained with the MLAT about a year after the test was administered yielded similar validity coefficients. The validity coefficients range from *r*=.26 to *r*=.89. Such a wide range of coefficients is due to the inaccuracy of the criterion measures used, which are subject to contextual circumstances (teachers rating using different criteria, students not working at the limit of their ability, etc.). Despite finding low coefficients, many high validity coefficients were found, as 73% of them are above .45 and 25% are above .60.

The reliability data were obtained from 980 subjects from four schools (see Table 2.2), which is a smaller number of subjects than the one that appears in the percentiles table as reproduced in section 2.3.1.2. Carroll and Sapon considered Parts 1, 2 and 3 highly speeded so, in order to obtain a reliability coefficient of these parts, they printed booklets on this purpose containing halves of these parts with approximately the same difficulty based on previous item analysis. Part 4, which is a paced test, was scored separating the odd- from the even-numbered items, thus obtaining half-scores, too. Reliability was then calculated using split-half correlations, and corrected by the Spearman-Brown formula, which takes no account of the standard deviation of items.

Standard errors of measurement (SE_M) are also offered in the *Manual*. These indicate more precisely the test score, as it is computed independently of the variability of the group. It is not to be confused with the standard deviation of scores on a test taken by a group of students. Instead, the SE_M refers to the standard deviation of test scores that would have been obtained from one single student had that student been tested repeatedly. These indexes, very closely related to reliability, tell us how much variability there is between an observed score and what would be the true score if one were tested multiple times. In this case, it is assumed that errors are normally distributed and that this variability is constant. An estimate of the SE_M is computed by multiplying the test score standard deviation by the square root of 1 minus the test score reliability.

Table 2.2. Reliability coefficients and standard errors of measurements of total raw scores on the MLAT-E (adapted from Carroll & Sapon, 2002:8)

Stats	GRADE 3		GRADE 4		GRADE 5		GRADE 6	
	BOYS	GIRLS	BOYS	GIRLS	BOYS	GIRLS	BOYS	GIRLS
r_{1I}	.96	.95	.96	.93	.95	.94	.94	.96
SE_M	4.6	4.9	4.6	4.8	4.4	4.3	4.3	3.3
Mean	64.0	65.9	80.0	91.3	94.3	96.8	100.9	105.0
SD	22.8	22.3	22.9	18.5	20.2	17.9	17.9	16.6
N	112	105	113	88	112	109	167	174

The *Manual* also presents us a table with the intercorrelation of parts (see Table 2.3). Ideally, all correlations should be from low to moderate, as they would show that each part taps different aspects of aptitude. In this case, the intercorrelations that tend to be higher appear when contrasting Parts 1 and 3. The part that most consistently has low intercorrelations with all other parts is Part 4 Number learning, followed by Part 2 Matching Words, whose intercorrelations are a bit higher and reach moderate levels in grades 5 and 6. From these results, it can be concluded that Part 4 (and possibly Part 2) are truly tapping at only one aspect of aptitude while the other subtests may overlap in the construct they are supposed to tap.

Table 2.3. Intercorrelations and reliability coefficients of parts of the MLAT-E (based on data from four schools). (Adapted from Carroll & Sapon, 2002:9)

		Grade 3							
		Boys (N=112)				Girls (N=105)			
		1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1.	**Hidden Words**	**.88**				**.70**			
2.	**Matching Words**	.51	**.83**			.46	**.77**		
3.	**Finding Rhymes**	.81	.42	**.94**		.77	.43	**.94**	
4.	**Number Learning**	.45	.32	.39	**.90**	.43	.45	.37	**.90**
	Mean	14.7	11.2	31.0	7.2	14.7	11.4	31.9	7.9
	SD	6.3	5.4	10.9	5.9	5.3	5.5	10.8	6.4
		Grade 4							
		Boys (N=113)				Girls (N=88)			
		1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1.	**Hidden Words**	**.79**				**.73**			
2.	**Matching Words**	.51	**.84**			.47	**.82**		
3.	**Finding Rhymes**	.72	.47	**.92**		.62	.40	**.83**	
4.	**Number Learning**	.54	.51	.53	**.94**	.45	.44	.32	**.94**
	Mean	19.8	13.5	36.0	10.7	21.3	16.8	39.1	14.2
	SD	5.3	5.8	9.7	7.3	4.5	5.6	6.6	7.4
		Grade 5							
		Boys (N=112)				Girls (N=109)			
		1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1.	**Hidden Words**	**.74**				**.74**			
2.	**Matching Words**	.49	**.84**			.66	**.88**		
3.	**Finding Rhymes**	.72	.50	**.93**		.55	.61	**.85**	
4.	**Number Learning**	.44	.54	.49	**.95**	.37	.47	.29	**.94**
	Mean	22.6	17.0	39.8	14.8	22.4	18.2	41.1	15.2
	SD	4.6	6.0	7.1	7.3	4.2	6.2	5.3	7.2
		Grade 6							
		Boys (N=167)				Girls (N=174)			
		1	2	3	4	1	2	3	4
1.	**Hidden Words**	**.70**				**.76**			
2.	**Matching Words**	.58	**.82**			.56	**.83**		
3.	**Finding Rhymes**	.64	.60	**.79**		.63	.59	**.90**	
4.	**Number Learning**	.50	.51	.41	**.94**	.41	.40	.39	**.92**
	Mean	23.8	19.2	41.2	16.7	24.7	21.4	42.2	16.7
	SD	4.0	5.6	5.5	7.0	3.8	4.8	6.0	6.6

Note. – Reliability coefficients are indicated in bold-face type.

The results presented in this section are displayed dividing the population into sex (boys and girls). Girls consistently appear to perform better on all the parts of the test but for Part 1 in Grade 3 and Part 4 in Grade 6, where boys and girls obtain the same mean. This would support the research into sex and language aptitude which supports the females' advantage over males (see section 1.6.4).

2.3.2. MLAT-E Spanish version (MLAT-ES)

In the absence of an instrument to measure aptitude in young learners whose L1 is Spanish, Stansfield and Reed adapted Carroll's MLAT-E in 2005, thus creating the possibility of studying language aptitude in young learners of many Spanish-speaking countries in South and Central America as well as in Spain. The resulting product is the *Modern Language Aptitude Test – Elementary: Spanish Version*, which was released in 2005 after being standardised and validated.

The sections that follow are devoted to the description of the parts of the Spanish version of the test, focusing on the main similarities and differences between the Spanish version and the original one in English (see appendix B for an overview of sample items on the MLAT-ES obtained from the SLTI website). Thus, in section 2.3.2.1, the four parts that form the MLAT-ES are contrasted with those of the English version paying special attention to the strategies that have been followed to adapt the MLAT-E to the Spanish language. Following are the data regarding the standardisation and norms of this new version of the test that are supplied in the *Manual*. Finally, in section 2.3.2.3, the reader can find a summary of the relevant information in the *Manual* related to the validity, standard error of measurement, reliability and intercorrelation of parts of the test that are necessary to support this new version of the MLAT-E as compared to its predecessor.

2.3.2.1. MLAT-ES: description

Following is the description of the different parts of which the MLAT-ES consists as compared to the way they were designed in the original version in English.

2.3.2.1.1. Parte 1: Palabras ocultas

This part is aimed to measure not only knowledge of Spanish vocabulary, but also sound-symbol association ability. In it, test takers have to decipher the word that is hidden under a misspelling disguise and match it with the word that means the same out of four options. Like the English version, this part in the Spanish version has 30 items, although the piloting version contained 31. The missing item was eliminated because it did not provide any additional information to the test (Stansfield, personal communication, 2006). The item eliminated was item 3, whose stem and options were:

3[12]. vallena □ con mucha gente □ animal marino enorme
□ barbilla □ peluca

Apart from statistical grounds, which are not explicitly stated in the *Manual*, this item could probably have been dropped due to the fact that "*vallena*" can be interpreted as "packed with people", which is the meaning of the distractor "*con mucha gente*".

Consonants are approached in the same way in the English and the Spanish version of the test: they are altered in such a way that, when read aloud, the stem sounds like the word they represent. The consonants involved in the MLAT-E are, among others, <k> for correctly spelled <c> or <ck>, <s> for unvoiced /s/ spelled as <ce> pronounced /s/ or /θ/, <s> for /ʃ/, or missing consonants for silent sounds like just <n> for spelled <kn> or just <s> for spelled <swe>.

Table 2.4. Examples of consonant sound-spelling correspondences in MLAT-E Part 1 Hidden words

Graphemes involved	Phonemes involved	Stem	Hidden word	Definition
<k> for <c>	/k/	kmfrt	comfort	ease
<k> for <ck>	/k/	nikl	nickel	a five-cent coin
<s> for <ce>	/s/	silns	silence	quiet
<s> for <ce>	/sə/	resnt	recent	not long ago
<sh> for <ce>	/ʃ/	oshn	ocean	the sea
<n> for <kn>	/n/	nif	knife	a sharp tool
<s> for <swe>	/sə/	ansr	answer	reply

In Table 2.5, one can see the strategies used to hide the consonants in the MLAT-ES such as the use of <k> for <c> pronounced /ka/, <q> for <cu> /ku/, <b> for <v> and vice versa pronounced /b/, <z> for <s> sounding /s/, <z> and <s> for /θ/ and <j> for graphic <g> and vice versa pronounced as /χ/; the use of double consonants for simple ones (e.g. <rr> for /r/ or <tt> for /t/) or the other way round (e.g. the use of <y> for /λ/, which is usually spelled <ll>).

[12] When specific items of the tests are explained apart, they will appear as in the printed version of the tests, that is, in Times New Roman, font size 12, or smaller if they appear as table feet.

Table 2.5. Examples of consonant sound-spelling correspondences in MLAT-ES *Parte 1 Palabras ocultas*

Graphemes involved	Phonemes involved	Stem	Hidden word	Definition
<k> for <c>	/k/	skushar	escuchar	oír
<q> for <cu>	/ku/	qlevra	culebra	reptil
<b> for <v>	/b/	bakka	vaca	da leche
<v> for <b>	/b/	kveza	cabeza	parte del cuerpo
<z> for <s>	/s/	rmozo	hermoso	bonito
<z> for <c>	/θ/	hozeano	océano	mar
<s> for <c>	/θ/	cilnsio	silencio	sin hablar
<j> for <g>	/χ /	solójico	zoológico	parque con animales
<g> for <j>	/χ /	geff	jefe	el que manda
<rr> for <r>	/ r /	cirrena	sirena	vive en el mar
<tt> for <t>	/t/	zkeletto	esqueleto	huesos
<y> for <ll>	/λ/	gayina	gallina	ave

Vowels are necessarily approached in a different way in the Spanish version as compared to the English one due to the big differences between both systems. In English there exist only 5 graphic vowels, but they can be pronounced in 12 ways as a monophthong (there not being a direct correspondence between the graphic sign and the sound). Besides, one single vowel can also represent a diphthong, as is the case of, for instance <i>, which in some words is pronounced /aɪ/ or <o> in “cold” /əʊ/. Consequently, with only 5 graphic vowels, many different sounds can be “hidden” in this test. Apart from that, many unstressed vowels, whatever their graphic representation, are pronounced as a schwa /ə/ or as a short <i> /ɪ/. Also, since some unstressed vowels are almost imperceptible, these have also been avoided graphically in some stems. Below are some examples of the vowel changes just explained.

Table 2.6. Examples of vowel sound-spelling correspondences in MLAT-E Part 1

Graphemes involved	Phonemes involved	Stem	Hidden word	Definition
<e> for <ee>	/i:/	nedl	needle	something used for sewing
<e> for <ea>	/i:/	egl	eagle	large bird
<e> for <y>	/i/	nme	enemy	not a friend
<i> for <i>	/aɪ/	frit	fright	fear
<u> for <oo>	/u:/	ruf	roof	top of a house
<-> for <a>	/ə/	buflo	buffalo	a kind of animal
<-> for <e>	/ə/	rivr	river	large stream of water
<-> for <o>	/ɒ/	kmfrt	comfort	ease

These phenomena do not appear as such in the Spanish version, as in Spanish the number of graphic vowels coincides with the phonemes they represent. Consequently, the only strategy used in this part of the test regarding vowels is omitting them by taking advantage of the names of the consonants which, when read, supply the vowel sound corresponding to the grapheme omitted (see Table 2.7).

Table 2.7. Examples of vowel sound-spelling correspondences in MLAT-ES Parte 1 Palabras ocultas

Graphemes involved	Name of the letter involved	Stem	Hidden word	Definition
<l> for <ele>	ele	tlbizión	televisión	se ve en una pantalla
<q> for <cu>	cu	qlevra	culebra	reptil
<d> for <de>	de	ddo	dedo	está en la mano
<k> for <ca>	ca	kveza	cabeza	parte del cuerpo
<n> for <ene>	ene	nmigo	enemigo	contrario
<t> for <te>	te	trror	terror	susto

2.3.2.1.2. Parte 2: Palabras que se corresponden

In this part, the stems are sentences with a word in capital letters. Test takers are expected to find the word that performs the same function in the sentence below the stem. The target functions in the English version are the same targeted in the Spanish one (subject, verb, adjective and direct object). In Table 2.8 are some examples which are very similar (though not exactly the same) in both the English and the Spanish versions of the test and illustrate all the functions at work. The word whose function is to be found in the sentence below appears in capital letters while the word in the sentence performing the same function is underlined.

Table 2.8. Examples of sentences of Part 2 in the MLAT-E and the MLAT-ES

Function	Language	Example
Subject	**English**	Did YOU buy the nice picture?
		Tomatoes grown on a vine.
	Spanish	¿Compró USTED una foto bonita?
		Las naranjas crecen en un árbol.
Verb	**English**	Henry THREW the heavy stone.
		Sally rides a bicycle.
	Spanish	Enrique TIRÓ una piedra grande.
		Sandra monta en bicicleta.
Adjective	**English**	Jill wore a GREEN dress.
		Alex wanted a new sled.
	Spanish	Juanita llevaba un sombrero VERDE.
		Álex quiere un par de patines NUEVOS.
Direct object	**English**	The dentist pulled my TOOTH today.
		Fred wrote a long letter.
	Spanish	Esta tarde el dentista me sacó una MUELA.
		Alfredo escribió una carta larga.

In the published English version, which has 30 items, 8 sentences aim at the subject, 7 at the verb, 7 at the adjective and 8 at the direct object. In the piloting version of the MLAT-ES, there were 31 items, 8 aiming at the direct object, the verb and the adjective and 7 at the subject. From the original 31 items, item 22 was dropped due to a miskey which has to do with the misinterpretation of word order and a false translation from English. The item eliminated was as follows:

22. ¿Qué te pareció el nuevo COMPAÑERO?

Mi tía salió y no apagó el televisor.

□ □ □ □ □ □ □ □

The key showed that "*televisor*" was the correct answer to this item when indeed, the correct answer is "*tía*", as "*compañero*" is the subject of the stem sentence. In Table 2.9, this item is analyzed in terms of the grammatical function and the semantic role of the elements of the sentences. It is important to say that semanticists have not reached an agreement in relation to the number or possible semantic roles that any language can have and that there is no universal truth about the roles of some elements due to the intrinsic characteristics of some verbs in relation to their grammatical function.

Table 2.9. Grammatical and semantic description of item 22 in the MLAT-ES *Parte 2 Palabras que se corresponden*

Stem 22 in Spanish	***¿Qué***	***te***		***pareció***	***el nuevo compañero?***
grammatical function	(?) predicative	indirect object		verb	subject
semantic role	(?)	experiencer		-	content
English translation of stem 22	**What**	**did**	**you**	**think**	**of the new mate?**
grammatical function	(?) predicative	aux.	subject	verb	prepositional object
semantic role	(?)	-	experiencer	-	content
Target sentence in Spanish	***Mi tía***	***(…)***		***apagó***	***el televisor.***
grammatical function	subject			verb	direct object
semantic role	agent			-	patient / theme
English translation of target sentence	**My aunt**	**(…)**		**turned off**	**the television.**
grammatical function	subject			verb	direct object
semantic role	agent			-	patient / theme

In item 22, "*apagar*" (turn off) is a transitive verb that requires a direct object that, in this case, could be argued that has the semantic role of patient because it is affected by the action of the agent ("*mi tía*" – my aunt) or the semantic role of theme because it is the thing being acted upon. In contrast, "*parecer*" (think of) is a psych verb that requires an experiencer. The experiencer role is assigned to the indirect object in Spanish, while in English it is the subject who performs this role. The question mark next to the grammatical function and in the semantic role box of "qué" (what) is a means to indicate that it is a complement the function of which is very controversial, as it shares characteristics of both direct objects and predicative complements. As for its semantic role, it is very difficult to assign one, if any, to it; hence the question mark as well.

Therefore, the miskey in item 22 could have been due to a misinterpretation of word functions in verbs that are similar in meaning in English and Spanish but work in a completely different way at the syntactic level. One common example of this phenomenon is the verb "*gustar*" (like), as shown in Table 2.10.

Table 2.10. Grammatical and semantic description of the verb "like" and "*gustar*"

English	**I**	**like**	**something**
grammatical function	subject	verb	direct object
semantic role	experiencer	-	content
Spanish	***Me***	***gusta***	***algo***
grammatical function	indirect object	verb	subject
semantic role	experiencer	-	content

Both in item 22 and in the example above, the sentences are virtually identical in both languages regarding their meaning, but the syntactic functioning of the elements that compose the sentences varies depending on the language at work.

The order of words in sentences is much stricter in English than in Spanish. While in English the canonical order is SVOCA (Subject, Verb, Object, Complement, Adverbials), in Spanish the verb can precede the subject and verb complements appear in different positions depending on the context of the utterance. Thus, the order of appearance of words in English is very revealing of the function they are performing in the sentence while in Spanish it is not necessarily so. Some ways to disguise this almost one-to-one correspondence in English in the MLAT-E is altering the order of the sentence by, for instance, placing a subordinate clause in front of the subject, as in "When winter comes the BIRDS fly south", or fronting elements such as complements of time or place (e.g. "In bad weather, I always CARRY my umbrella"). Some word order alteration is also present in questions with the verb to be, as in "Is your SISTER still sick?". However, in Spanish, since word order is much more flexible, it allows the speaker to alter the order of appearance of words in the sentence in order to put more emphasis on some parts rather than on others. Hence in item 9, "*A Juan le COMPRARÁN un regalo el lunes*", the speaker is emphasising the fact that Juan, and not someone else, will be given a gift next Monday.

In English, adjectives and determiners, regardless of the type they are, always appear in front of the noun they complement unless they are in a predicative position (e.g. "Children love to play in the COLD snow" versus "The snow in the playground is COLD"). In Spanish, however, adjectives may appear in two positions, either before or after the noun they complement. When they appear after the noun, they are qualifying adjectives, that is, they mention a quality of the noun they complement. For example:

13. Las jirafas tienen un cuello LARGUÍSIMO

Jorge come en una mesa amarilla del parque.

□ □ □ □ □ □ □ □

In item 13 "*Las jirafas tienen un cuello LARGUÍSIMO*" (Giraffes have an extremely long neck), "*larguísimo*" (extremely long) describes the way a giraffe's neck is and so does "*amarilla*" (yellow) as far as the "*mesa*" (table) is concerned in the sentence meaning "George eats at a yellow table in the park".

When adjectives are fronted, they can be either qualifying adjectives or determiners. Let's take a look at item 4:

4. La GRAN mansión del presidente es blanca.

En la clase de matemáticas hay pocos alumnos.

□□ □ □ □ □ □ □

Item 4 "*La GRAN mansión del presidente es blanca*" (The president's BIG mansion is white) has as a target the quantifier "*pocos*" (few) in "*En la clase de matemáticas hay pocos alumnos*" (There are few students in Mathematics class). In the same sentence, the prepositional phrase "*de matemáticas*" ("of Mathematics" if translated word for word) tells us a quality of the noun "*clase*" ("class") in the target sentence, while "*pocos*" informs us of a quantity, not a quality, as "*gran*" does.

The subject is obligatory in English, while in Spanish, being a pro-drop language, it is not, especially if it is a personal pronoun. Consequently, item 24 (23 in the published version) "*¿A qué horas crees tú que llegarás a cenar?*" sounds redundant to the Spanish native speaker, as "*llegarás*" already carries the grammatical information of 2nd person singular. Dropping the subject would be the most common way to say this sentence, but it is not possible in this case because the subject is the target function of this sentence in the test.

As far as the wording of items is concerned, three different strategies have been followed in the adaptation/translation process from the MLAT-E to the MLAT-ES. First, some items have been translated directly from English. For instance, "Last summer my FATHER took me to the circus" – Years ago, people lived in caves - has been translated into "*El año pasado mi PAPÁ me llevó al circo*" – "*Los hombres vivían en cavernas hace miles de años*". Second, some other items have undergone some slight content and word order changes such as Item 4: "Peter WINDS his clock every night" – "In the summer the warm winds blow", which has been translated into "*Pedro PONE el* *despertador todas las noches*” – “*En el verano soplan vientos calientes*”. This is so because the verb phrase “wind a clock” becomes a periphrastic expression in Spanish (“*Pedro LE DA CUERDA al despertador todas las noches*”) while “*pone*”, which could be translated into “sets” is made up of just one word. Finally, the rest of items have been changed completely so there is no correspondence between the items but for the functions at which they aim.

The cultural references in the English version have been eliminated in the translation process. Consequently, “George Washington” is not mentioned any longer in the Spanish version and neither is “Goldilocks”. The odd thing is that most references to the rural world disappear in the Spanish version, too. Thus, the “barn” that appears in item 27 ("We found the box under a table in the old barn") becomes an “*ático*” (a penthouse) in Spanish; item 29, in which a henhouse, a fox and chickens appeared, has been completely changed as has as well item 19 “The farmer’s son carelessly dropped the eggs”.

The translation of some items arose some confusion when the piloting test was administered. Item 9 was worded as “*María y José jugaron fútbol*” ("Mary and Joseph played football”) whereas in peninsular Spanish the translation would be “*María y José jugaron a fútbol*”). Item 12 also had a typos in the diminutive “*trencito*”, which should have been “*trenecito*” in peninsular Spanish, as one-syllable words in the diminutive form add an –e, while two-syllable words do not (*tren* > *trenecito* vs. *Carmen* > *Carmencita*). Finally, item 24 is worded as "*¿A qué horas…?*" (What time…?) while in peninsular Spanish the most common way to ask this question is by using the singular form “*hora*”. These items appeared in this way in the piloting version because, as Stansfield explained, “our chief item writer is Chilean and Chileans drop a lot of prepositions following verb (...).The same applies to item 12, *trencito*. *Trencito* is correct in Chile but incorrect in Spain. *Trenecito* is accepted everywhere, so we'll go with *trenecito*. We have already made these changes.” (Stansfield, personal communication, April 2005).

Two other minor changes which were also suggested by the author of this dissertation to Stansfield and Reed were not introduced. Although the author of this dissertation suggested the removal of two proper nouns, “*Leila*” and “*Perla*”, because some Catalan participants in this dissertation did not know whether they were the name of a girl or not, only “Leila” was replaced for “Susana”. Also, “*luciérnagas*”, which is an insect some participants had never heard about before, was also kept despite its removal was also suggested.

2.3.2.1.3. Parte 3: Palabras que riman

English offers a wider range of options to choose from in order to create the items of Part 3 than Spanish does because of the phonological and spelling system of these languages. Although the test standardisation of the Spanish version proved that this part was reliable and that it fulfilled the requirements of validity, it is not clear whether the score on this part is related with the phonetic coding construct it is intended to tap. If it is so, it should be demonstrated if this relationship is as straightforward as it is when the English version of this test is used. This is due to the fact that in Spanish there is an almost exact sound – spelling correspondence except for some consonant variations or deviations. Consequently, unless there are differences in the stress placement on the word, it is possible to complete the test just by checking the order of appearance of the last letters without actually knowing whether the words rhyme or not. This is, however, not possible in English. In spite of this, there are at least three phonetic phenomena that force test takers to look for the rhyme between words as they cannot rely only on the orthography to choose the answer.

One of these phenomena is the fact that <b> and <v> are homophones in Spanish. This is illustrated in item 22 (*CLAVO – rabo,* i.e. NAIL - tail), item 34 (*CUEVA – prueba,* which are CAVE – proof in English) and item 38 (*VALLA – baya* meaning FENCE - berry of the published version. Two other phonetic phenomena that have been used in part 3 are, on the one hand, *seseo* and *ceceo* and, on the other hand, *yeísmo.*

Seseo and *ceceo* affect the way <s, z> are pronounced. In Peninsular Spanish <s> is pronounced /s/ and the consonant in <z, ce-, ci> is pronounced /θ/. People who *cecea* pronounce <s> as /θ/ while people who *sesea* pronounce <z, ce-, ci> as /s/. The speaker on the CD *sesea* and some rhymes are thought to play with the lack of phoneme-grapheme correspondence that the *seseo* implies. It is worth mentioning that in Spain only half of the *comunidad autónoma* of Andalusia, in the south of Spain, some regions of Murcia and in the Canary Islands, *sesea*, that is, only half of Andalusian people pronounce /s/ for /θ/. In the south-west of Andalusia there also exists the opposite phenomenon, the *ceceo*, which involves pronouncing /θ/ for /s/ so the distinction of these consonants in the different variants would be as follows:

Table 2.11. Sound-spelling representation of the *seseo* and *ceceo* phenomena

Peninsular Spain (except for the South of Andalusia)		Most part of South America, Canary Islands and centre of Andalusia (*seseo*)		Southwest of Andalusia (*ceceo*)	
<s> sounds /s/	repisa /re’pisa/	<s> sounds /s/	repisa /re’pisa/	<s > sounds /θ/	repisa /re’piθa/
<z, ce, ci> sounds /θ/	tiza /‘tiθa/	<z> sounds /s/	tiza /‘tisa/	<z> sounds / /θ/	tiza /‘tiθa/

The items on the MLAT-ES that contain these consonants, due to the phonological changes described above, work differently depending on the variety the test taker speaks. That is to say, while all the rhymes to be found in the items are consonant, those items containing these consonants will aim at finding consonant or vocalic rhyme depending on the dialect the test taker has. This change happens in item 16 (*PROMESA – cabeza,* i.e. PROMISE - head), item 25 (*BRAZO – vaso,* which mean “arm” and “glass” respectively) and item 36 (*PAYASO – pedazo,* i.e. CLOWN - piece) of the published version.

Another phonological change used in this part is *yeísmo*, which consists in mispronouncing <ll>. <ll> should be pronounced as /λ/ although there is the general tendency to pronounce it as /j/, which is graphically represented as <y>. Catalan speakers do not tend to make this phonological change, since /λ/ is a phonemic entity in Catalan. The use of these phonemes make the items containing them aim at vocalic rhyme, not consonant rhyme, for those test takers who *yeyean.* That is the case of item 38 of the published version (*valla – baya*).

Some other options, although they may not rhyme with the stem, are very similar in terms of the consonant and vowel combinations they have, supposedly in order to draw the test takers’ attention. For instance, item 5 *FLECHA* (arrow) is very similar to the distractor “*ficha*” (card), item 6 *CUANDO* (when) resembles both “*cuánta*” (how much) and “*cuento*” (tale), item 21 *HOMBRO* (shoulder) is similar to the distractors “*hombre*” (man) and “*hambre*” (hunger) or 45 *FLACO* (thin) has the same consonant – vowel combination except for one vowel as its distractor “*fleco*” (fringe).

The piloting version of the MLAT-ES had 46 items, while the published version has 38. The number of items in the Spanish version does not coincide with the published version in English, which has 45 items. In a personal communication, Stansfield told the author of this dissertation that the other eight items had been removed because they were not representative or relevant (they did not provide any new information or affected the item-total score of the part). If we take a look at them, we can see that 4 out of the 8 items removed contained the s/z conflict, 2 other items aimed at recognising consonant rhyme (or, why not, coincidence of the last letters of the words written) without any distractor liable to be chosen at first sight, and the other two aimed at finding consonant rhymes but with some more elaborated distractors which did not work as expected. In one case, two of the distractors (here underlined) rhymed with the stem at the vowel level (*BUENO – tengo – muerdo – heno – pino)*, the correct answer being "*heno*". In the other item removed, stress came into play (see distractor underlined) so the strategy of focusing on the last letters written, if it were not for the graphic stress, could not work in this case. The item removed was *AMIGO – obligó – ombligo – refugio – hormiga,* the correct answer being "*ombligo*"*.* The analysis of the distractors of the piloting study could probably clarify which were the actual reasons to remove these items and not others.

2.3.2.1.4. Parte 4: Aprendamos números

The Spanish version of this part is very similar to the English version. Test takers learn some numbers and how to combine them to form two-digit numbers. The only thing that changes is the name that the numbers to be learned receive and the amount of numbers to be learned. While in the English version test takers are supposed to learn the names of numbers 1, 2, 3, 20 and 30, in the Spanish version they are also supposed to learn the name of the number 10. The name of numbers in the English version are as follows: "*ba*" is "one", "*baba*" is "two", "*dee*" is "three", "*tu*" is "twenty" and "*ti*" is "thirty". Notice that the name for number 20 is very similar to the name of the English number 2 and that number 2 is formed by saying 1 ("*ba*") twice. Therefore, one could remember the name of number 2 by remembering that two times one is two.

In Spanish, test takers are meant to learn 6 numbers and to combine them in the same way as in English. The name of numbers in the Spanish version are as follows: "*co*" is "one", "*vein*" is "two", "*ras*" is "three", "*silca*" is "ten", "*vinca*" is "twenty" and "*rasca*" is "thirty". Notice that the beginning of the name for the number 2 is the same as the number 20 in Spanish ("*veinte*") and that the tens are formed by adding what could be analysed as a suffix (-*ca*).

2.3.2.2. MLAT-ES: standardization and norms

The MLAT-ES was standardised with data collected in the 2004-05 school year. It was administered to 1,186 students from public and private elementary schools in Spain, Mexico, Costa Rica and Colombia. Of the 441 students from Spain, 227 were bilingual speakers of Catalan and Spanish and are the same subjects whose data will be used for the research in this dissertation. The subjects were in grades from 3 to 7 (the sample for the standardisation of the MLAT-E was in grades from 3 to 6 only). In the *Manual*, Stansfield and Reed (2005) admit that the data they present for the standardisation and norms is an initial reference, but they feel a larger sample should be tested to obtain a more accurate picture of the population than the sample referred to.

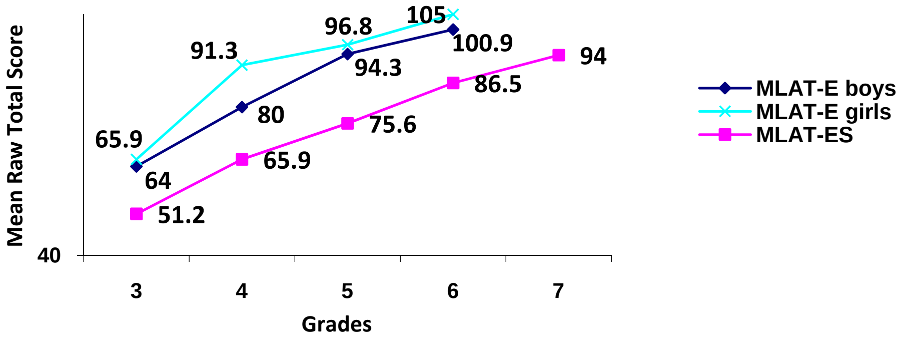
As for the norms, one can find five tables in the *Manual*. The first one presents percentile norms for raw Total Scores on the MLAT-ES by grade only (see Table 2.12), in contrast with the *MLAT-E Manual*, which offered these norms by grade and sex (see section 2.3.1.3). In the *MLAT-E Manual*, the division of the population according to the sex variable did not yield significantly different results, so Stansfield and colleagues could have considered that it is a piece of information of which it seems sensible to dispose. If the results that follow were to be compared to the percentile norms of the MLAT-E, it should also be taken into account that Part 3 has a different number of test items (MLAT-E part 3 has 45 items whereas MLAT-ES has 38), so the total raw score maximum is different (130 vs 123). Consequently, the means and standard deviations for this sample have to be compared with caution too because of the different amount of items. The other tables in the *MLAT-ES Manual* present the percentile ranks of the raw part scores by grade. These cannot be compared to the ones on the MLAT-E either because this information is missing in the *MLAT-E Manual*.

Table 2.12. Norms for students in grades 3, 4, 5, 6 and 7 on the MLAT-ES, Total Score. Raw total scores corresponding to designated percentiles. (Table adapted from Stansfield and Reed, 2005:9)

Percentile	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
99	102-123	117-123	117-123	118-123	120-123
97	97-101	114-116	114-116		119
95	93-96	110-113	112-113	117	
93	91-92	107-109	110-111	114-116	117-118
90	86-90	104-106	108-109	113	115-116
87	80-85	99-103	106-107	110-112	114
84	76-79	95-98	104-105	108-109	113
81	72-75	93-94	101-103	107	112
78	71	88-92	99-100	105-106	109-111
75	67-70	95-87	96-98	103-104	108
72	64-66	82-84	94-95	102	106-107
69	62-63	79-81	91-93	101	105
66	61	76-78	89-90	99-100	104
63	59-60	74-75	87-88	98	103
60	54-58	71-73	85-86	95-97	101-102
57	51-53	68-70	82-84	93-94	100
54	50	66-67	78-81	92	98-99
51	48-49	64-65	75-77	90-91	97
48	43-47	60-63	73-74	88-89	96
45	41-42	58-59	70-72	86-87	95
42	40	55-57	67-69	84-85	93-94
39	39	53-54	65-66	81-83	88-92
36	37-38	50-52	62-64	79-80	86-87
33	34-36	48-49	59-61	76-78	85
30	31-33	46-47	55-58	73-75	83-84
27	29-30	43-45	52-54	70-72	81-82
24	28	40-42	51	67-69	80
21	27	38-39	47-50	64-66	77-79
18	26	35-37	46	59-63	75-76
15	23-25	32-34	42-45	56-58	72-74
12	20-22	30-31	39-41	51-55	64-71
9	16-19	23-29	37-38	45-50	57-63
6	13-15	20-22	33-36	40-44	52-56
3	11-12	17-19	26-32	35-39	44-51
1	0-10	0-16	0-25	0-34	0-43
N	207	206	289	306	178
Mean	51.2	65.9	75.6	86.5	94.0
SD	25.3	28.0	25.9	23.0	19.4

Regarding the information in Table 2.12, and comparing it with the norms table of the MLAT-E, at first sight one sees that the means are lower than the ones in the MLAT-E of the norming sample even if there were more items and so the range of scores is wider. The evolution of scores follows the same pattern in both cases: a sharper increase between grade 3 and 4 and not so sharp in the other grades, although in the *Manual* it is said that "mean total scores increase in a uniform way from one grade level to the next, while the standard deviation and standard error of measurement decrease for each grade level" (Stansfield & Reed, 2005:14). Besides, the tendency is for the raw total score to increase throughout all grades, but not at such high rate between grade 6 and 7. In Figure 2.1 below, one can compare the evolution of the raw total mean scores of both the MLAT-E and the MLAT-ES norming.

Figure 2.1. Raw total scores on the MLAT-E and on the MLAT-ES per grades (from the data in the *Manuals*)



2.3.2.3. MLAT-ES: statistical information

Like the *Manual* of the MLAT-E, the *Manual* of the MLAT-ES supplies test users with detailed and valuable information about the validity, standard error of measurement, reliability and intercorrelation of parts of the test. Since both tests share the same underlying constructs, these data are supported by the previous validation study.

The data available on validity were obtained by the same means as they had for the validation of the MLAT-E, i.e. comparing the scores with performance as measured by course grades or the FL teacher's estimate of achievement. That is, they used criterion-related measures. The coefficients for the total score on the MLAT-ES range from *r*=.26 to *r*=.42, which is a narrower range of coefficients if compared to the range of the MLAT-E. High validity coefficients were not expected by Stansfield and Reed, they say, due to the inaccuracy of the criterion measures used. These involved the numerical scores obtained from a questionnaire administered to the teachers in which they had to rate their students' aptitude for various aspects of FL learning: listening and reading comprehension, speaking, writing, grammar and vocabulary. In addition to that, the teachers should also provide the students' probable mark in a FL that year. The students themselves were also asked to provide the mark they had obtained in the FL subject the previous year. Despite the fact that the correlations were from low to moderate, it is important to say that nearly all of them are statistically significant and that the best predictor of the criterion measures was the total score, as the part scores obtain lower coefficients. Actually, when the correlations across grades were averaged using Fisher's Z-transformation, the total score correlations proved to be 25% higher than the part score ones, which supports the fact that the total score is perhaps the best FL aptitude index as measured by the MLAT-ES.

The reliability data were obtained from four countries and are consistent and high for all grades both for raw total scores of the whole test and for each part. No explanation is given as to how these figures were obtained. That is, the reader does not know if Stansfield and Reed used the same method as Carroll and Sapon did (split-half correlations based on previous item analysis) or Cronbach's alpha indexes. Whichever the method, partial reliability coefficients in these types of tests tend to be very high due to the high number of items, as equations of reliability include the number of items as a factor. Actually, Cronbach's alpha is the most common measure of scale reliability because items can be split in several ways and depending on the way a test is split, one may come up with one index of reliability or another. One possible way to split a test is scoring the even numbered items and the odd ones separately and then examining the correlation between the two halves obtained. However, this implies that each part has fewer items than the original test and, as it has been pointed out, the more items on a test, the higher the reliability index. Consequently, split-half reliabilities underestimate the value obtained unless the Spearman-Brown formula is applied. This formula, which does not take account of the standard deviation of items, can be used to estimate the true level of reliability if a test is lengthened or shortened. Then, though, it is assumed that any item added is parallel to those already on the test. Besides, splitting a test does not guarantee that the parts obtained will be identical. When doing so, many factors come into play, such as the item's facility (IF), the item's index of discrimination or the order of appearance of this item on the test. Generally speaking, Cronbach's alpha can be considered a measure equivalent to splitting data in two in every possible way and to computing the correlation coefficient for each split. In addition to the information on reliability, the indexes of SE_M are also provided in Table 2.13, adapted from the *Manual*.

Table 2.13. Reliability coefficients and standard errors of measurements of total raw scores on the MLAT-ES (adapted from Stansfield & Reed, 2005:9)

Statistics	GRADE 3	GRADE 4	GRADE 5	GRADE 6	GRADE 7
N	207	206	289	306	178
Reliability	.97	.97	.97	.96	.95
SE_M	4.72	4.70	4.67	4.47	4.25
Mean	51.2	65.9	75.6	86.5	94.0
SD	25.3	28.0	25.9	23.0	19.4

The *Manual* also presents us a table with the intercorrelation of parts (see Table 2.14), the *p*- value of item facility and the mean item-total correlation data (which can be interpreted as an index of item discrimination) by test part and grade level. When combining all grades, parts 1 and 3 seem to be moderately correlated (*r*=.57), which is something expectable, as both are meant to tap phonetic coding ability. Parts 2 and 4 are also correlated moderately (*r*=.58). They share, as a target, the ability to recognise the grammatical function of words (especially Part 2) and to discover the relationships between words (especially Part 4), although Part 4 is also supposed to measure some memory component. The intercorrelations obtained, as Stansfield and Reed (2005) state, are very similar to the intercorrelations for the MLAT-E, which supports the idea that the MLAT-E and the MLAT-ES are comparable tests in terms of the psychological constructs tapped. The reliability of all the parts and grades combined was excellent (Cronbach’s alpha .97).

Table 2.14. Intercorrelations, reliability coefficients, mean *p*-values, and mean item total of parts of the MLAT-ES (based on data from 10 schools) (adapted from Stansfield & Reed, 2005:17)

	Grade 3 (N=207)				**Grade 4 (N=206)**			
	1	2	3	4	1	2	3	4
1. *Palabras que riman*	**.93**				**.92**			
2. *Palabras que se corresponden*	.36	**.88**			.57	**.92**		
3. *Palabras que riman*	.42	.53	**.95**		.59	.52	**.94**	
4. *Aprendamos números*	.34	.53	.43	**.95**	.48	.63	.53	**.95**
Mean *p*-value	.37	.37	.44	.49	.43	.50	.60	.60
Mean item total	.57	.47	.60	.66	.55	.56	.57	.67
	Grade 5 (N=289)				**Grade 6 (N=306)**			
	1	2	3	4	1	2	3	4
1. *Palabras que riman*	**.91**				**.91**			
2. *Palabras que se corresponden*	.45	**.93**			.36	**.93**		
3. *Palabras que riman*	.59	.55	**.94**		.46	.46	**.94**	
4. *Aprendamos números*	.41	.49	.40	**.95**	.33	.47	.44	**.93**
Mean *p*-value	.54	.59	.64	.69	.64	.68	.73	.78
Mean item total	.53	.56	.55	.66	.52	.57	.55	.61
	Grade 7 (N=178)				**Grades combined (N=1186)**			
	1	2	3	4	1	2	3	4
1. *Palabras que riman*	**.93**				**.93**			
2. *Palabras que se corresponden*	.27	**.91**			.52	**.93**		
3. *Palabras que riman*	.37	.30	**.93**		.57	.57	**.95**	
4. *Aprendamos números*	.42	.38	.24	**.90**	.49	.58	.50	**.95**
Mean *p*-value	.75	.73	.76	.82	.55	.58	.64	.68
Mean item total	.57	.54	.53	.54	.58	.58	.59	.67

Note. – Reliability coefficients are indicated in bold-face type.

The mean *p*-values inform us about the difficulty of the tests for each grade, as it is the average percent correct score on a test. Except for grade 7, in which all mean *p*-values are comparable, Part 1 appears to be the most difficult for all grades, followed by the other parts in order of appearance in the test. Mean *p*-values higher than .55 indicate that the test is easy. Taking this value into account, the test starts to be easy at grade 5 and it is very easy in grade 7. Part 4 is particularly easy from grade 5 on. All parts, as the mean item-total correlation indicates, are very good at discriminating across grades.

2.3.3. Use of the MLAT-E and the MLAT-ES

The MLAT-E was developed for children between 8 and 12 years old who are literate in the language of the test. According to the test manual, the uses for the

Modern Language Aptitude Test include selection, placement, guidance and diagnosis of learning abilities. It is recommended to administer it before or right at the beginning of the study of a FL, although it is also informative if it is administered when children have already received FL instruction as it is assumed that, in principle, language training should not affect the scores on the test.

The MLAT-E for selection

The MLAT-E can be used to select individuals who show promise in learning foreign languages. In contrast with the version for adults, the MLAT-E is not used to justify the time and expense of placing them in a language training program, since learning a FL at primary school is determined by the official curriculum. However, it can be used to select the students who may be placed in advanced classes.

The MLAT-E for placement

In situations where there is more than one class or group of students in a language training programme or course, the students can be placed according to their aptitude level so that they can make the most of class time.

The MLAT-E for guidance

Besides aptitude, other factors such as motivation or interest have an influence on FL learning. Having access to a student's MLAT-E scores, it can be more easily determined if failure at learning a FL is due to a lack of FL aptitude or to any other factor.

Diagnosis of Learning Disabilities

The MLAT-E can also be used together with other measures to diagnose a history of FL learning disability. Moreover, looking at a child's score on the different parts of the test can help to match students' learning styles with instructional approaches.

For the official announcement of the MLAT-ES, the SLTI staff distributed a leaflet in which they also listed the following uses, which are supposed to have been adapted to the contexts in which the MLAT-ES can be distributed, that is to say, wherever there are Spanish native speakers who mostly learn English as a FL. The list included the following points:

- Placing new students into educational settings that support their development of English language skills
- Determining how students perform in relation to international norms for Spanish-speaking children
- Creating expectancy tables to show the relationship between language aptitude scores and grades in FL (including EFL or ESL) classes
- Identifying students with low second language learning aptitude
- Identifying students with second language learning disability
- Identifying gifted students, particularly for learning languages
- Developing profiles of strengths and weaknesses to inform the teaching of English
- Developing local norms and a placement system after norms are established
- Investigating issues or hypotheses in language aptitude or cognitive linguistics

The use of the MLAT-E for research into SLA is, obviously, one of its applications, as it is of the MLAT. However, other than two pieces of research, the author has not been able to find any other study in which the MLAT-E has been used as a measure of FL aptitude.

The first one is a study by Hauptman (1971) in which the MLAT-E was used as a measure of language aptitude in children. In this study, two groups of children whose L1 was English were compared according to the approaches they were made to follow in their Japanese as a FL classes. Hauptman labels these approaches as “structural approach” and “situational approach”. In the “structural approach”, the FL grammatical and lexical forms were presented in an ascending order of difficulty, while in the “situational approach” the same materials were introduced by means of dialogues without following any difficulty rationale. The findings in this study as regards the teaching approaches were that the learning results in the situational approach were slightly better than those in the structural approach. As far as aptitude is concerned, those students with high language aptitude and intelligence performed better in the situational approach and there was no significant difference between approaches among students of lower aptitude and intelligence. Therefore, according to this study, high aptitude learners benefit from teaching approaches of communicative nature.

Since the norms of the MLAT-E are for grades 3 to 6, Harper and Kieser (1977) designed a study to provide preliminary standardisation data for the test for grades 7 and 8 and to prove the concurrent validity of the MLAT–E for predicting FL achievement. The population was divided in two groups: one followed an audiolingual approach and the other sample followed what they labelled a “cognitive approach” or conscious-active method, that is, an approach in which writing and speaking practice in both English (L1) and French (FL) are combined with the teaching of formal grammar aspects. The samples contained 144 boys and 133 girls from grade 7 and 145 boys and 137 girls from grade 8. The results obtained on the MLAT-E show that the boys from grade 8 obtained lower scores than the other three samples. The scores were all similar across grades, though, and the means were not significantly different from one another. From low to moderate significant correlations were found between results in the MLAT-E and the FL achievement measures (final mark) no matter the teaching approach. Running a stepwise multiple regression analysis, it was found that Part 2 Matching Words was the part that contributed the most to obtaining concurrent validity, followed by Part 1 Hidden Words. Part 4 Number Learning was the least effective of all the parts in predicting FL achievement. On account of the results of the stepwise multiple regression, Harper and Kieser conclude that administering only the first two parts could result in a reliable FL aptitude measure up to a certain extent.

If the quantity of studies published that use the MLAT-E in English is very scarce, it is even scarcer, not to say almost non-existent, when it comes to the use of the Spanish version. The author of this dissertation is not acquainted with any study that makes use of the MLAT-ES as a measure of aptitude besides the concurrent validity that appears in the *Manual* using criterion variables that may not be as reliable as desirable.

2.3.4. Improvements on the MLAT-E and the MLAT-ES

Research into language aptitude has focused mainly on adults; consequently, the MLAT has been much more used than the MLAT-E. That is probably why, while the MLAT has been extensively criticised in the literature, there is hardly any body of research which fosters or undermines the use of the MLAT-E or the MLAT-ES. What follows below are not published reservations against the test, but the review of the changes made to the test as they appear in the *Manual.* These changes are based on the feedback resulting from the trialling phase together with some other suggestions by the author of this dissertation. The suggestions came from the minor conflicts that arose during the administration process of the MLAT-ES.

In the *Manual* of the 2005 edition, Stansfield explains that they administered the test to three children to improve the flaws that the original version had. Thus, they were able to make some minor improvements related, mainly, to the recording and the layout of the test. For instance, he mentions that one of the children interviewed considered the space provided for writing the name too short. On this same note, I suggest that, since the directions read “My name is (print)” in the English version and “*Mi nombre es (letras mayúsculas)*” in the Spanish one, but they do not make any reference to surnames, it would be wise to rewrite them as "My name and surname are (print)" and "*Mi nombre y apellidos son (letras mayúsculas)*". This is just a practical issue to make the task easier for the researcher if more data is being collected to be contrasted with scores on the MLAT-E or for the teacher using the test so as to be able to identify the test taker more quickly.

On the cover, test takers are asked to write two marks: the mark they had last year in the FL subject and the mark they "*esperan sacar*". In Spanish "*esperar*" means both "hope" and "expect". Therefore, when answering this question, test takers could not be answering the same question as it is posed. Accordingly, it is advisable that this is rephrased into a new sentence that does not present this ambiguity. One possible rephrasing could be "*¿qué nota crees que sacarás?*", which would translate into English as "which mark do you think you will obtain?".

At the end of each page there are instructions such as "Go on to the next page" and "Stop. Do not turn the page". These were made bigger because the main test administrator, i.e., the author of this dissertation, noticed they were smaller than what test design conventions dictate. Further improvement regarding this could be mentioning the number of pages each part has on the CD recording, or writing how many pages are left for each part in the booklet (1 of 2 pages, 2 of 2 pages). The author considers this addition necessary because some of the test takers were confused with the note "*Total de esta página*" that appears right above the instructions related to turning or not the page. This last direction is meant to help the test scorer to mark the items on each page as well as the short lines that appear next to each item, which also caused some confusion among the youngest test takers. Another solution in order to make the instruction of turning the page more noticeable could be to increase the font size of these directions. Actually, the font size used in the text is 12, while test standards recommend 14-point size serif typeface letters for tests to be taken by children up to grade 6.

The children interviewed by Stansfield also noted some inconsistencies between what was being said in the recording and what was written in the booklet. These inconsistencies were solved in the new edition. In the Spanish try-out version, there was also a similar problem: in Part 4, test takers are required to write 25 numbers. However, in the piloting version of the booklet there was space for only 24 items.

The number of items on the MLAT-ES is 123 while on the MLAT-E is 130. Strikingly enough, the total score of both tests is the arithmetic sum of the four parts they consist of. Consequently, the weight with which each part contributes to the total score is not the same. An explanation about this should be given in the manuals so that the test user can know what each part measures and how much weight is attached to it. Otherwise, one same score can have a different explanation depending on the partial scores.

Specifically for the Spanish version, another suggestion to be taken into consideration if the test is to be administered in Spain is the speaker's Spanish variety. The speaker of the directions recorded on the CD *sesea*, that is to say, he pronounces all the dental consonants /θ/ as unvoiced alveolar fricative consonants /s/. This surprised the participants of this study, and made them giggle and make some comments with the classmates sitting next to them when they first listened to the tape. Although the Spanish subjects quickly got used to the accent and got to paying attention at once, it could be a good idea to record as well one CD version with a speaker of Peninsular Spanish to be used when administering the test in Spain.

Something else that caused some giggling in class was the name of number 30 in Spanish, "*rasca*" which, besides being meaningful per se (it means "it scratches"), happens to be the name of a TV cartoon character. Whenever the participants heard "*rasca*" during the explanation of the part, it was inevitable for some students to say "*Rasca y Pica*", which are the names of Itchy and Scratchy in the Spanish dubbed version of the show within the show in the TV series *The Simpsons.* These characters are called Tomy and Daly in the Hispanic-American dubbed show, so it would not be necessary to change the name of this number for this population.

The construction of part 3 in languages such as Spanish and Catalan, as well as in German, among others is an issue that deserves further exploration. Due to the almost transparent correspondence between grapheme and sound in these languages, phonetic coding may not be tapped in the Spanish version of this subtest in the same way as it is supposed to be in the English version. Instead, the test taker could simply adopt the strategy to recognise letter chains without even thinking about the way they sound. Apart from that, in some items there is more than one option that rhymes with the stem. More specifically, test takers are expected to choose the option that rhymes with the stem with a consonant rhyme, and not only because of the coincidence with the vowel phonemes, i.e. where there is a vowel rhyme. However, this is not mentioned explicitly in the directions.

On the whole, it can be concluded that more research is needed that uses the MLAT-E or the MLAT-ES as an instrument of FL aptitude in order to fine-tune the design of these tests as well as to make sure that they are valid measures of FL aptitude that warrant their use for diagnosis, selection, guidance and placement.

2.3.5. Other aptitude measures for young learners

As stated in the previous section, while there is a large body of research into FL aptitude in adults, this is not the case of FL aptitude in young learners. The MLAT-E, published in 1967, has hardly been used and, since its release, there have been scarce attempts at designing new aptitude tests for young learners. Luckily, this situation is starting to change little by little. Besides the adaptation of the MLAT-E by the SLTI, Milton and Alexiou (2004a, 2004b; Alexiou, 2005, 2009) have undertaken several projects on FL aptitude in young and very young learners (from five to nine years old) taking as a model both Carroll and Sapon's MLAT-E and Esser and Kossling's (1986) cognitive tests of aptitude. The preliminary product of their research is a test that measures, first, short-term rote memory, by means of a memory-picture game; second, semantic integration, through a test in which children have to recall the shapes that had been presented to them before and that are now, in a second view, missing; third, the capacity to retain sign pairs, considered to be equivalent to the capacity to retain FL vocabulary; and, fourth, the learners' classification and inductive ability by means of a game that uses an artificial language.

The results obtained by Alexiou and Milton led them to believe that aptitude is a changing ability in young children (contrarily to what some think it appears to be in adults). They believed that analytic skills improve after about the age of six while memory does not, as Gathercole and colleagues (1992) also suggest, but the results obtained are not totally conclusive as far as memory is concerned. This would have implications regarding FL learning and teaching, as young learners should not be considered memorisers who learn FLs mainly implicitly. They could also learn explicitly, although this type of learning is likely to be more effective in older young learners.

Nevertheless, some studies present the opposite results as for WM. A regular increase in WM was observed in subjects between 6 and 19 years old (Siegel, 1994) followed by a gradual decline after adolescence (though not in STM) and has also resulted in the strongest predictor of L2 learning in traditional, grammar-oriented teaching contexts (Ando et al., 1992, in Mackey et al., 2002), especially in older children, while they found the opposite result in more communicative-oriented contexts.

Kiss (2004; in Kiss & Nikolov, 2005) also developed and validated a new aptitude test with a sample of 419 12-year-old children who are Hungarian learners of English. In order to measure language aptitude they took as a model the test used by Ottó (1996), which is based on Carroll's MLAT, and Pimsleur's PLAB, and simplified it so that it suited the needs and cognitive level of the population of the study. The resulting test has 4 tasks. Task 1 is entitled Hidden Sounds and in it, students hear 15 sentences from a tape and have to identify which word is in the sentence they hear from the four words they have heard and also seen written in each test item. Therefore, it is a sound-symbol association task. Task 2 Language Analysis involves studying a set of words and a sentence in a nonsense language together with their Hungarian equivalents in order to, later on, choose the correct translation out of four new Hungarian sentences. The focus of this task is recognising structural patterns. The aim of Task 3 Words in Sentences is, as part 4 of the MLAT and part 2 of the MLAT-E, identifying semantic and syntactic functions. Finally, Task 4 Vocabulary Learning aims at measuring memory for lexical items. Students have to memorise twelve words and expressions in a nonsense language along with their Hungarian equivalents and then, in a multiple-choice test, they have to identify the equivalents of ten of these words.

Once the modifications from the piloting study were applied, they administered the final version of the test. They also gathered data regarding background and motivation as well as English language proficiency measures (listening, reading and writing) and grades in other school subjects. As for the aptitude test, Task 4, whose the focus was memory, was the one that proved to be the easiest, while Task 2 was the most difficult one. This is consistent with previous research that seems to indicate that children make more use of memory than adults (Harley & Hart, 1997) and hence, memory is more developed than other abilities. The correlations among scores on tasks of the aptitude test showed a dissociation between them, and the correlations among the aptitude test and the FL proficiency measures were moderately high (r=.634, p <.01).

Although no significant differences were found between girls and boys in the norming study of the MLAT-E (see section 3.5.5), they do exist in the study by Kiss and Nikolov (2005), as girls outperformed boys significantly in both aptitude scores and proficiency measures. They found that the total scores of the aptitude measure correlated significantly and moderately with language proficiency in general (a composite of several listening, reading and writing tasks). Concerning the relationship between aptitude scores and the students' mark in English and in other school subjects, it was found that participants with higher English marks were also better on the language aptitude test. This same relationship was found between aptitude and course grades in other school subjects but the relationship was not so strong as in the case of FL marks. The quantity of input did not seem to have an influence on the aptitude scores obtained either. In turn, a relationship was found between motivation and aptitude (r=.367), but this was weaker than the relationship between motivation and FL proficiency (*r*=.478) and than the one between aptitude and FL proficiency (*r*=.627). A multiple regression analysis finally indicated that FL aptitude explained over 20% of the variation in FL proficiency while motivation explained almost 8% of it.

For selection purposes, a primary school required a FL aptitude measure for children at the end of grade 2. Kiss (2009) was in charge of developing this aptitude measure in Hungarian, which would be contrasted with FL performance one year later approximately. She designed it by using mostly the task types of the MLAT-E but with fewer items (39) due to time constraints for the administration (only 45 minutes). The test consisted of four parts. The first three parts had 10 items each. Task 1 Find the Matching Word corresponded to Part 2 Matching words on the MLAT-E, Task 2 Rhyming Words, corresponded to Part 3 Finding Rhymes on the MLAT-E, and Task 3 Hidden Words, followed the same rationale as Part 1 on the MLAT-E. In the fourth task, Vocabulary Learning, children had to study a list of 12 words and expressions in an unknown language and their equivalents in Hungarian. After practising the words and expressions, they had to identify 10 of these nonsense words by choosing them from four alternatives. The final version of this task contains only 9 items. The test was piloted in a small group of children (N=40) who achieved a mean score of 56.3%, SD 13.4% and were distributed close to normality. The reliability index was not excellent (Cronbach Alpha .74), but the discrimination power was sufficient.

After the piloting phase, the aptitude test was administered to two grade-2 groups which showed a similar performance on the test. For these two groups, Tasks 1 and 4 were more difficult than Tasks 2 and 3 and task 1 was significantly more difficult for these children than it was for the children in the piloting phase. Intercorrelations between parts showed that each part measured a different construct, as they were significantly low, although some sort of association was found between Tasks 2 and 3, both of which deal with sound-symbol association.

When comparing the 2-graders' performance in the 2009 study with that of the 6-graders in Kiss and Nikolov (2005), it is evident that 6-year-old children obtain much lower scores than 12- and 13-year-old children. Consequently, the construct validity of this test can very much depend on the test takers' age. Once the test was piloted, it was contrasted with the FL proficiency measures of 25 children enrolled in a dual language teaching program. The FL proficiency measures arose from an oral interview consisting of two tasks in which five linguistic aspects were assessed: task achievement, fluency, vocabulary, accuracy and pronunciation. The marks were obtained from a rating scale created *ad hoc*. The scores on the language aptitude measure were correlated with these marks, obtained from three different raters: a trained and experienced teacher who did not know the children beforehand, a teacher who had taught the children for one year and a British native speaker who was teaching the children participating in this study the year the study took place. The aptitude measure only correlated with the marks and ranks given by the teacher who had taught the children participating in the study for one year (*rho*=.404 and 427, *p*<0.05), and did not correlate with the marks assigned by the teacher who was teaching the participants that year nor with the scores provided by the trained rater who did not know anything about the children's FL proficiency. As the number of participants in this study is rather low and the FL proficiency measures are rather limited, the results obtained are indicative of a tendency but should not yet be generalised.

2.4. Using the MLAT-E in Spanish and designing the Catalan versions of the MLAT-E

The study of aptitude in young language learners is still, as explained in section 2.3.5, in an embryonic state, although the prospects seem quite promising. In order for this type of research to reach the same status as the study of FL aptitude in adults, there are two main options to choose from. One possibility would be continuing to design, pilot and validate new aptitude measures. This implies not only the work related to the construction of the items or exercises comprised in a test, but also being able to assure that the test is the instrument that will serve to reflect the psychological construct intended in its design later on. Another option is for researchers to use tests which have already been validated and which have proved to tap the construct they are supposed to measure. It is necessary to translate and adapt these tests to languages and cultures other than the one used originally so that the test can be administered to populations different from the ones participating in the validation stage. Both options require effort, work and time, so it is up to the researcher to choose one option or the other depending on their needs and affordability.

2.4.1. Statement of the problem

Any researcher willing to study the FL aptitude of subjects between grades 3 and 7 in Catalonia in the first decade of the 21st century is in need of an instrument to determine this construct. Other than the MLAT-E translated into Spanish, the MLAT-ES, no other measures of FL aptitude exist at present. From the two options exposed above, the author of this dissertation decided to further validate the MLAT-ES which, due to its recent translation and adaptation, had not yet been used as a measure of FL aptitude with subjects whose L1 is Spanish. Moreover, since this test for young learners has not been so widely used as the version for adults, using the Spanish version of this test would contribute to the study of FL aptitude in young learners as measured by tests with an MLAT-E format. Thus it would be possible to further confirm whether the theoretical basis on which it is grounded, the MLAT, has a clear correspondence in the version for young learners.

The data of this study were collected in Catalonia, a bilingual community in which both Catalan and Spanish are spoken as L1s. This context made it possible to not only further validate the MLAT-ES, but also to adapt it and translate it into Catalan. Thus the test was administered in the two L1s of the community.

Both the Spanish and the English versions of the MLAT-E are accompanied by a manual with a wealth of detailed information on the test. Along with the description and purpose of the test, in the *Manual*, the test administrator can find information about the data of the participants in the piloting study, the steps to follow in the administration of the test, the scoring methods as well as statistical information about the validity, standard error of measurement, reliability and intercorrelations of the parts. Finally, a section entitled “special considerations” provides the test administrator with a short review of the uses of aptitude tests for placement, guidance and diagnoses of FL learning disabilities. The manual of the English version also has a couple of subsections devoted to the changes to test materials as well as the appropriateness of the MLAT-E for seventh and eighth graders. Thanks to this manual, any researcher interested in using the MLAT-E can access data that will be useful to compare their results with those of the norming study in the *Manual*.

2.4.2. The present study

From the statement of the problem, it can be drawn that this study has more than one aim. The first one is the development of a Catalan version of the MLAT-E so that Catalan/Spanish bilinguals in grades 3 to 7 can take this test in Catalan, one of their L1s, and not only in Spanish.

Secondly, using the MLAT-ES in Catalonia makes it possible to analyse the test item functioning in a community different from the one present in the original norming study. As Oakland puts it, “those engaged in adapting tests and using them should be sensitive to views and attitudes that tests are likely to be inherently invalid when used with different groups within one country and especially when used cross-nationally” (2005:76), as happens in this study. The norming study was partially carried out in Spain with monolingual participants from Madrid and only a small sample of test takers were bilinguals from Catalonia.

Having a look at the *Manuals* of the MLAT-E and the MLAT-ES, one can see there are two significant differences. On the one hand, while data for the norming study in the *MLAT-E Manual* only appears up to grade 6, this is not the case of the *Manua*l of the MLAT-ES, which covers grade 7 as well. The reasons for that absence in the English version are unknown. Moreover, from the results in the *Manuals*, it can be seen that the evolution of the means across grades is not linear, but the difference between some grades is larger than between some others. It is, therefore, a third objective of this thesis to analyse how the MLAT-E in Spanish and Catalan work across grades.

In the fourth place, it is also easy to see that the English version of the *Manual* provides the data for the norming divided according to the participants’ sex, while the norming study using the Spanish version does not. However, no explanation is given about the absence of this classification of the data. Therefore, another objective of this dissertation is to find out whether there are any differences in the performance on the MLAT-E according to sex and try to see if these exist in the performance on the MLAT-E in Spanish and Catalan as well.

Finally, one more objective of this dissertation is to check the construct validity of the MLAT-EC and the MLAT-ES. Bearing that aim in mind, the subjects were asked to complete a cloze passage, a multiple-choice listening activity and a dictation in English. The first two differed in type and/or number of items depending on the grade and only participants in grades 5 to 7 did the dictation.

2.4.3. Research questions

The objectives exposed in the previous section result in the following research questions for this dissertation:

Research question 1: To what extent are the MLAT-E in Spanish and Catalan suitable language aptitude measures for learners in grade 3 to 7?

Research question 2: Is there a relationship between language aptitude (as measured by the MLAT-E in Spanish and Catalan) and the subjects' sex?

Research question 3: Is there a relationship between language aptitude (as measured by the MLAT-ES and the MLAT-EC) and the subjects' proficiency in English as a foreign language?

2.5. Summary of Chapter 2

Chapter 2 has described in depth the most widely used aptitude test for adults, the MLAT, in spite of the criticism that it has received along the years. Thus, specific data on the parts of which it consists, its validity, statistical information, mean *p*-values and norms have been reproduced in the preceding pages. The positive results in the norming study have rendered the MLAT as an excellent aptitude measure although, as years have passed, its usefulness has been questioned. The same has been done with the MLAT-E, the version of the MLAT adapted for the use with young learners, and its adaptation and translation to Spanish, the MLAT-ES. The latter has been described comparing and contrasting each part separately with its homologous version in English, already hinting some of the challenges that the adaptation of the English version to a Romance language can pose.

The usefulness of the MLAT-E in young learners is, in principle, backed up by the validity of its parent. Not having been used much so far, some improvements can already be suggested, though, not only to the MLAT-E, but also to the MLAT-ES. Along with the MLAT-E and the MLAT-ES, some other try-outs with other aptitude measures for young learners have been presented. These are the tests designed by Alexiou

(2005) and Kiss and Nikolov (2005), thanks to which aptitude in young learners has started to be studied in this first part of the 21st century.

Having defined aptitude and related it to the factors that may have an influence on it in Chapter 1, and after having explored how aptitude is measured with the MLAT for adults and the MLAT-E and the MLAT-ES, at the end of this chapter, the reasons why this dissertation is useful have been exposed. Lacking valid aptitude measures to be administered in a bilingual Catalan/Spanish context such as Catalonia is what drives the rest of this dissertation, which will aim at solving three research questions: first, if the MLAT-ES and its Catalan version, the MLAT-EC, are valid measures for grades 3 to 7; second, if there are any differences between the aptitude in boys and girls as measured by the MLAT-ES and the MLAT-EC; and third, if these two tests show concurrent validity with the proficiency measures and criterion measures.