UiT The Arctic University of Norway

Faculty of Humanities, Social Sciences and Education

Aptitude in the Classroom: an empirical study of the pedagogical functionality of the LLAMA test battery in an upper secondary school

Morten Skillingstad Larsen

Master's thesis in English Linguistics and Education, ENG-3982, May 2021

Abstract

This thesis explores the suitability of using aptitude testing and the *LLAMA aptitude tests* in a Norwegian upper secondary school class, and the potential pedagogical advantages such testing can have. Aptitude testing entails measuring language learners' specific talent for learning foreign languages and this is an individual difference that exhibits considerable variation between learners (Dörnyei & Skehan, 2003).

An empirical study was conducted on 22 participants of an upper secondary school class to see how the LLAMA, an aptitude test battery developed by Paul Meara (2005) would function. The testing was followed by a student questionnaire and two separate teacher interviews, created to investigate the experience and attitudes the teacher and the pupils showed towards aptitude testing and the LLAMA, as well as the potential pedagogical advantages this testing might have.

The results showed that both the teacher and the pupils viewed the LLAMA as a suitable aptitude battery and that the age group was appropriate. The teacher was also positive towards the notion of aptitude testing. Several pedagogical advantages were found and could, with some effort from the teacher, help inform and individually adapt the teaching to each pupil, based on their aptitude profiles.

From the findings of this project, I conclude that aptitude testing and using the LLAMA could help Norwegian teachers individually adapt their teaching and that this is something we should strive to use. I also suggest that there are several pedagogical advantages if the results from the testing are used accordingly and if a functional framework for how to use these results are developed.

Acknowledgements

I would like to use this opportunity to direct my sincere gratitude towards everyone that has stood by me during this project and helped me in my writing process. The support from teachers, friends and family has been important to me during the time I have written this thesis.

I would like to give a special thanks to my two supervisors Christopher Loe Olsen and Natalia Mitrofanova for invaluable advice and feedback, as well as partaking in the dozens of master meetings throughout the whole writing process.

I would also like to thank my family and friends for being patient and supportive in a period where time has been of shortage to me, and where my availability to others has been sparse. My partner has also been writing a master's thesis in the same period, and the many discussions with her have helped me a lot.

In addition to the people mentioned above, I would like to thank the teacher and the pupils who have set aside valuable time from their work and education to provide me with the necessary empirical data for this project.

Table of Contents

List of Figures

List of Tables

1. Introduction

Language aptitude has been a familiar concept in language learning circles for decades, it has recently experienced a resurgence of interest in pedagogical settings, although its pedagogical use has been controversial. The topic of this master's thesis is language learning aptitude, and the concept will be discussed with pedagogical implications and aims.

Language aptitude is defined as a "specific talent for learning foreign languages that exhibits considerable variation between learners." (Dörnyei & Skehan, 2003, p. 613). It is an important individual difference in the study of SLA and has been viewed as one of the most important factors for language learning success. The LLAMA is a test battery that measures participants aptitude level. It has been developed by Paul Meara at the University of Swansea (Meara, 2005). This is an interesting invention in the field of SLA since the test represents one of the newest and most modern aptitude test batteries. This is also an important addition, as many of the older test batteries like the MLAT and PLAB seems to become more outdated in relation to pedagogy, because of the new ways of teaching, where oral interaction and communicative activities predominate (Robinson, 2002). In my previous course works, I compared the LLAMA with the MLAT and the PLAB and discussed differences and strengths and weaknesses with the three aptitude test batteries. In light of this, my motivation for writing this thesis is both personally, as I caught interest in the LLAMA and aptitude testing, and practically grounded.

A renewed interest in the study of aptitude in recent years has also emerged in the SLA research society. Studies have shown that aptitude is the single best predictor of subsequent language learning achievement of all the individual differences (Dörnyei & Skehan, 2003; Sawyer & Ranta, 2001). Because of this, I found aptitude and the LLAMA to be an interesting starting point for a master's thesis and I further believe that these factors can contribute to enhancing the education of Norwegian pupils, if some of the necessary elements discussed in this paper are in place.

1.1 Aims and Scope of the Thesis

The main areas of investigation for this thesis are to look at how the LLAMA functions in the context of an upper secondary school class. The thesis will also connect aptitude and the LLAMA to pedagogy by investigating how aptitude potentially can inform teaching as well as

how the results from the current study inflicted on the choices made in the teaching of the teacher. The attitudes of the teacher will also be viewed as severely important to reach the goal of this thesis. The goal is to discuss if the *LLAMA Language Aptitude Tests* is suitable for an upper secondary school class and how the results can be used actively in teaching English to optimize the education every pupil receives, based on the individual differences and language learning abilities they possess.

The justification of this study is grounded in practical reasons and seeks to inform and help teachers in Norwegian schools to see the LLAMA and aptitude testing as a potential tool that can help adapt and inform their teaching. The study is written as a starting point for further research on how to apply the results from the LLAMA test battery to language education.

1.2 Research Questions and Research Design

In order to investigate the connection between aptitude testing and the potential practical pedagogical implications these results might have, several factors were important for the study. These factors derive from questions of whether the aptitude test is suitable for the participants of the study and also what attitudes those involved in the testing procedure have towards the concept. The research questions for this thesis were developed on the basis of what questions I believed to be of importance to enlighten and discuss, when it came to the idea of enrolling the LLAMA into being a practical mapping tool in the education of Norwegian pupils. Several other studies that discuss the connection between aptitude testing and pedagogy have also inspired me and given me a theoretical basis for developing research questions (see Erlam, 2005; Granena, 2013; Psochner, 2018; Rogers et al., 2017; Wen, Biedroń & Skehan, 2017). The first two research questions are aimed at investigating the suitability and functionality of the aptitude in an upper secondary school class. Research question 3-6 aim at investigating the pedagogical relation the test can have to instructional teaching of English. The thoughts surrounding these ideas have resulted in the following six research questions for this thesis:

RQ1: Is the LLAMA a suitable aptitude test battery for an upper secondary school class?

RQ2: How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?

RQ3: Does the LLAMA aptitude test correspond with the teacher's perception of the pupils' aptitude?

RQ4: What pedagogical advantages can aptitude testing have in an upper secondary class in English?

RQ5: What attitude does the teacher have towards aptitude testing and the LLAMA?

RQ6: Will the teacher make pedagogical changes to the English education based on the aptitude results, and if so, what type of changes?

The overall research design for this thesis will be an empirical case study. Chapter two of this thesis will explain the conceptualizations of the term language learning aptitude, as well as the history and development of aptitude testing. The LLAMA aptitude test will also be presented. Afterwards, relevant theory and studies where the LLAMA has been central to the research that was conducted, will be presented. Chapter three will present the methodology applied in this thesis. The chapter will inform the reader about how the study was conducted and what tools were used to carry out the intervention. Chapter four will present the results from the study and different kinds of analysis' will be provided to understand the results. Chapter five will apply theory and results to discuss the research questions asked in this thesis. Finally, chapter six will conclude by using the findings from all the research questions to answer how the overall goal of thesis was reached and what my final assertions about the research questions asked are.

To answer the research question, several methods of investigation will be applied. RQ1 will mainly be answered through carrying out the LLAMA in the target class and by using a questionnaire, as well as a teacher interview. The student questionnaire will answer RQ2 by giving me insight into the experience the pupils had with using the LLAMA. RQ3 will be answered by using a frame for predicting the pupils' aptitude, as well as the LLAMA test results. RQ4, RQ5 and RQ6 will be answered by using the results from the two teacher interviews.

1.2 Ethical Considerations

The ethical consideration of this project was of the highest importance, as a considerable amount of personal data was processed. Since the participants were 15-16 years old, it was important that they would be protected and properly informed about what they agreed to when they entered the project. Permission from the Norwegian Centre for Research Data (NSD) was granted, and a letter of consent (see Appendix A) was handed out to all of the participants, before the intervention started. The form required a written consent from the participants, and they were encouraged to show the letter of consent to their parents before they signed. This letter of consent was given to the participants several days before the intervention started so that they would have enough time to inform their parents. Since the pupils were all above 15 years of age, permission from the parents was not required. The letter of consent informed the participants about every aspect of the project, data storing, the length of the project and how the data would be used. The results were code marked and stored on servers with password protection so that no others than me would have the opportunity to acquire them.

2. Theoretical background

This section will be introduced by briefly presenting the main concepts of individual differences in second language acquisition and provide an overview of what concepts of individual differences are included in this study. Furthermore, a closer look into the main area of interest for this study, namely aptitude and aptitude testing will be done. Some central conceptualizations of the term aptitude will be presented to highlight the intricateness of the term. Secondly, a presentation of the LLAMA language aptitude tests and the research and work behind this new aptitude test battery will be provided. This section will also present some of the recent research that has been done on aptitude testing using the LLAMA. Some of the latest research conducted by using the LLAMA will also be presented and results from studies regarding language teaching and methods, language learners and feedback will be provided.

2.1 Individual differences

The field of second- and foreign language acquisition is a vast research area with many important areas of study that can explain the factors and stages that take part in a person's acquisition of a new language. Out of all of the research areas that go into this particular aspect of linguistic studies, the study of individual differences is key to understanding why the success among second language learners varies so greatly. The understanding of how the characteristics of individuals are related to their ability to succeed in learning a second language is of great interest to both educators and researchers (Lightbown & Spada, 2018). The number of relevant individual differences vary from study to study and according to which conceptualisation of the term is used and by which researcher. Still, one of the more common ways to study the phenomenon is to divide the concept into the following individual differences: *intelligence, language learning aptitude, learning styles, learning strategies, personality, motivation,* and *learner beliefs* (Dörnyei, 2005; Dörnyei & Skehan, 2003; Hummel, 2014; Lightbown & Spada, 2013; Skehan, 1989). By studying these factors one can gain an understanding of the underlying structures that explain the development in the learning process of a second language and the resulting proficiency of the learner. This paper will not go further in discussing other individual differences than language learning aptitude, as this is the focus of the current study.

2.2 Language learning aptitude

The term language learning aptitude is one of the most debated and discussed factors of the individual differences listed above, both in terms of defining the concept, but also in terms of measuring it (See for example Ameringer et al.; Carroll, 1981; Granena, 2020; Li, 2016; Skehan, 1989; Skehan, 2002). John B. Carroll and Stanley B. Sapon (1959), the creators of the popular and widely used aptitude test battery, the Modern Language Aptitude Test (MLAT), proposed that the concept of language learning aptitude consists of "basic abilities that are essential to facilitate foreign language learning" (p. 14). More recent scholars have defined language aptitude as a "specific talent for learning foreign languages that exhibits considerable variation between learners" (Dörnyei & Skehan, 2003, p. 613). Several scholars have argued that language aptitude is the most reliable and important predictor of second language success (Dörnyei & Skehan, 2003). The concept is tied closely to a person's specific talent for learning new languages and the ability to understand and analyze new language constructs. Aptitude is often the one factor that sets learners apart even though they learn the same language, with the same intentions and with the same instructor and previous knowledge. To understand the term aptitude a conceptualization of the underlying constructs is often important.

There are several different conceptualizations of language aptitude, both older and more recent, where different aspects of the term are emphasized, and certain elements are included or excluded. Many newer conceptualizations have been proposed in recent years, but as the LLAMA test is quite heavily based on the works by Carroll and Sapon (1959), Carroll (1981) and Pimsleur (1966), this framework is what will be presented in this study. Carroll's (1981) conceptualization of aptitude is still frequently used as the basis for understanding the term and its underlying structures. In his work, language aptitude is described to consists of the four components *phonetic coding ability* which is to identify distinct sounds and associate them with certain symbols, *grammatical sensitivity* which is to recognize grammatical functions, *rote learning ability* which is to learn associations between sounds and meaning and *inductive language learning ability* which is to infer or induce the rules governing a set of language materials (Carroll, 1981). In more recent research Skehan (1998) redefined two of these components, namely grammatical sensitivity and inductive language learning into one new concept called *Language Analytic Ability* (LAA). He defined this new construct of aptitude as "the capacity to infer rules of language and make linguistic generalization or extrapolations" (Skehan, 1998).

Paul Pimsleur (1966), the creator of another aptitude test battery called Pimsleur Language Aptitude Battery (PLAB) has also conceptualized the term by dividing it into three concepts, namely *verbal intelligence* which describes the abilities and the knowledge of words and verbal analysis and reasoning, *motivation* which shows how motivated the learner is and *auditory ability* which means to receive and process information through the ear. It is interesting that Motivation is included into Pimsleur's conceptualization of language aptitude, as most researchers define motivation and aptitude as two separate individual differences in language acquisition (Hummel, 2014). This also contrasts with the findings of Li's (2016) meta-analysis of 66 studies examining the construct validity of language learning aptitude. The conclusion was that aptitude was independent of other individual differences such as motivation (Li, 2016). With exception of the motivational factor these basic conceptualizations of aptitude have great relevance for this study and the creation of the LLAMA. A subtest of the PLAB called *Sound Discrimination,* which measures the learner's ability to recognize sound patterns, has been widely known for having influenced many later aptitude batteries (Skehan, 2002).

As these conceptualizations of the term aptitude get clearer an interesting question regarding whether aptitude is a dynamic or a static trait in the language learning abilities of humans arises. The question posed concerns whether aptitude is something is inherit and unable to change or whether it is an individual difference among learners that can change and develop through the language learning process. Wen, Biedroń and Skehan (2017) discusses the relation between the two terms *ability* and *aptitude.* Carroll (1993) viewed abilities as traits that exhibits "stability and permanence even over relatively long periods of time" (p. 7). Aptitude can easily be recognized as one of these traits a human FL learner can possess. The trait will then be something that describes a given talent for language learning. In line with Carroll, Dörnyei (2005) also point out that even though these are two different terms, based on the contexts in which they are used "in typical practice the two are used synonymously" (p. 32). This was clearly also the view of Carroll as he concluded by viewing aptitude as a latent trait. As such, Carroll regarded aptitude as a sort of ability, namely a latent trait that is relatively stable and relatively resistant to training, and which refers to the potential for achievement provided instruction is optimal (Wen, Biedroń and Skehan, 2017). As such, there might exist a certain agreement among several researchers on foreign language aptitude, that language learning aptitude is a stable trait which is little or not affected at all by other conditions such as language education.

Even though there seems to be a unison agreement regarding the idea of aptitude as a stable and latent trait, the field of research has developed through the last 15 or so years. Zhisheng Wen, one of the researchers mentioned above, and several others, have turned more towards a view of language aptitude that suggests the concept is moving towards being considered a more dynamic trait for language learning. Newer research suggests that working memory is a central aspect of predicting language learning success and thus this component of the mind can be included in the area of FL aptitude. According to Wen, Biedroń and Skehan (2017) this incorporation of working memory means that "[i]t is fair to argue then that the concept of FL aptitude has developed from being seen as a stable and unitary fixed trait (Carroll's time) to being considered as a more dynamic and multiple sets of abilities which interact with other internal or external factors." (p. 23). This makes the discussion of pedagogics and language aptitude much more interesting. In earlier research by Carroll and other SLA researchers of his time, language aptitude was seen as a tool for merely predicting and explaining language learning success. With working memory included into the conceptualization of language aptitude and a new way of thinking about aptitude as a more dynamic and alterable trait, the field of research can move more towards incorporating aptitude as a part of relevant pedagogical research and more empirical data can be collected in the context where aptitude is envisioned more in the direction of pedagogical execution (Wen, 2012).

There has also been conducted new research to suggest that the term aptitude might not be as cohesive as first envisioned and that the term might actually consist of two constructs. Epstein (1990) and later also Pacini and Epstein (1999) proposed a new way of understanding aptitude where the concept was divided into what was labelled *implicit aptitude* and *explicit aptitude.* These two types of aptitude are a part of a theory where a dual-processing system of learning is beinig used. The two styles differ in the sense that they tap into different cognitive learning systems in order to acquire a new language. Implicit language aptitude is a system of learning where the learner uses a more nonconscious, holistic, effortless and faster method of acquiring new knowledge (Granena, 2020). Explicit aptitude entails a style of acquiring a new language where more slow, conscious, analytical and effortful techniques of learning are being used. This style of learning contrasts the implicit way of learning in every way and essentially represents the opposite style of learning a new language (Granena, 2020). This new way of understanding language learning aptitude also lays way for new pedagogical implications for aptitude as it becomes a more dynamic term by viewing it in this way. Another important

development in the field of aptitude research when aptitude is viewed in this way, is that learning strategies, teaching methods and other relevant factors to the language learning process can be linked up to the different cognitive styles and the different language learning aptitudes.

2.3 The LLAMA Language Aptitude Test

The LLAMA Language Aptitude Tests is an aptitude test battery that was created after a series of projects carried out by students of English Language and Linguistics at the University of Wales Swansea. The LLAMA is a revised version of the Swansea LAT battery that was developed a few years earlier and included the subtests Lat_A, Lat_B, Lat_C, Lat_D and Lat_E (Meara, Milton & Lorenzo-Dus, 2003). Paul Meara (2005) made use the Lat-battery and revised it into what is now the LLAMA, due to the fact that some of the subtests from the Lat were less satisfactory and were not as applicable to aptitude research as firstly hoped. As mentioned above, the current test battery is loosely based on the works of Carroll and Sapon (1959) as well as other research on the field of language learning aptitude test batteries. Since the LLAMA is free to use and easily administered, the test has gained more attention in recent years and several studies have used it to measure aptitude (Artieda & Muñoz, 2016; Granena, 2013; Granena & Long, 2012; Kourtali & Révész, 2019; Rizvanovic, 2018; Saito, 2017; Yalcin & Spada, 2016). Still, the manual states that the creator of the test cautions its users in using the LLAMA in high-stake situations, as the test has not been properly validated and standardized yet (Meara, 2005).

The LLAMA language learning aptitude test is free to use from (www.lognostics .co.uk/tools/llama/index.htm) and it is easily administered by using computers and hearing devices. It includes four subtests: LLAMA_B is a vocabulary learning task, LLAMA_D is a sound recognition task, LLAMA_E is a sound-symbol correspondence task and LLAMA_F is a grammatical inferencing task (Meara, 2005). The test takes about 25 minutes to administer and can be used in several language situations as there is also evidence to support that the test is language neutral (see Rogers et al., 2017). The sounds and words in the subtests are gathered from quite unknown languages such as a native Central American language and a British Columbian Indian language and combines these with images and symbols, instead of English words (Meara 2005). This creates a language neutrality and learners do not have to understand or master English or any other known language to make use of the test. Each of the subtests

have a score range of 0-100. What the score of each subtest indicates varies some but the scores are described mostly in the same point ranges. The descriptions of achieved aptitude on the test have a varying score range, but the manual (Meara, 2005) describes the area around 0-15 as a "A very poor score," the area around 20-40 as "An average score; most people score within this range," the area around 45-70 as "a good score," and the area around 75-100 as "an outstandingly good score. Few people manage to score in this range." Correct answers are indicated with a light "ding" sound, whereas wrong answers are indicated with a "bleep" sound.

2.3.1 LLAMA_B

The LLAMA_B is a simple vocabulary learning task, which measures the learner's ability to learn relatively large amounts of vocabulary in a relatively short space of time (Meara, 2005). It is loosely based on the vocabulary learning task called *paired associates* from Carroll and Sapon's (1959) MLAT, but with a new interface. In this subtest, learners are given 120 seconds to learn the names of twenty different symbols by clicking on them. After the timer has run out (2 minutes by default) learners are given one of the previously introduced names and then has to click on the corresponding object/figure on the screen. The names of the figures are taken from a Central American language, and they are randomly assigned to each figure (Meara, 2005). This measures learner's ability to acquire a new vocabulary of a target language.

Figure 1 - LLAMA_B

2.3.2 LLAMA_D

The LLAMA_D is a sound recognition task that does not to appear in the works of Carroll and Sapon (1959). The subtest measures the learner's ability to recognise short stretches of spoken language that the learner was previously exposed to a short while beforehand (Meara, 2005). This subtest is loosely based on the works by Service (Service, 1992; Service & Kohonen, 1995) and Speciale (Speciale, Ellis & Bywater, 2004). The LLAMA_D also resembles the PLAB subtest *Sound Discrimination*, but this is not mentioned by Meara (2005) in the LLAMA manual. In this subtest, learners are orally exposed to ten words in an unfamiliar language. After, they have to listen to a number of words and identify if the given word was among the originally learnt ten words or not. The words in this test are names of objects in a British-Columbian Indian language. This measures the learner's ability to recognize patterns, particularly in spoken language.

Figure 2 - LLAMA_D

2.3.3 LLAMA_E

The LLAMA_E is a sound-symbol correspondence task which measures the learner's ability to work out the relationship between sounds and a given writing system (Meara, 2005). This subtest shares some traits with the sound-symbol correspondence test in the works of Carroll and Sapon (1959). The LLAMA_E is a revised version of the original Lat_E from Meara, Milton and Lorenzo-Dus (2003). In this subtest students are given 120 seconds to learn the spelling of an unfamiliar language. This is done by clicking on 22 different buttons with syllables on and then hearing how they are pronounced. After the practicing phase (usually two minutes by default), students listen to a word and are then asked to choose, from two options, the grammatically correct spelling of that word, based on what they learned about the fictional language in the practice phase. For this subtest, learners are allowed to take notes. This measures learners ability to connect sounds to symbols and thus their phonetic coding ability.

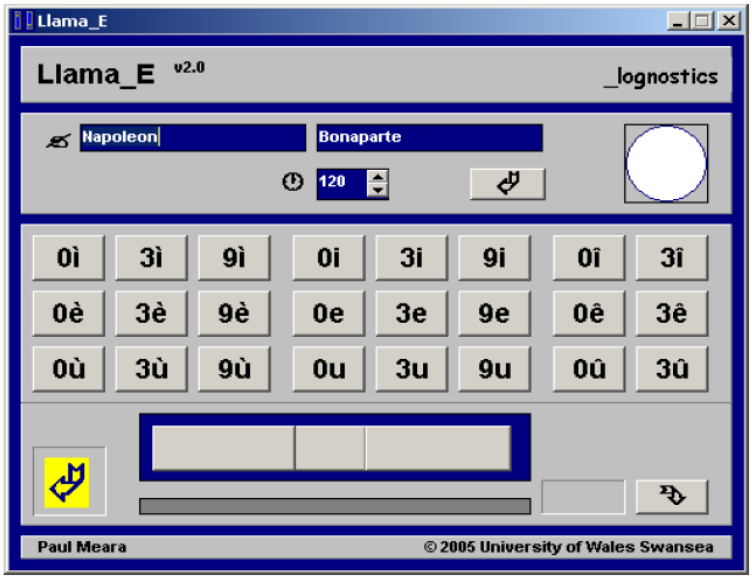


Figure 3 - LLAMA_E

2.3.4 LLAMA_F

The LLAMA_F is a grammatical inferencing task which measures the learner's ability to work out the grammatical rules of an unfamiliar language. It is a revision of the Lat_F and the subtest has been good at identifying learners with outstanding analytical linguistic abilities (Meara,

2005). The last subtest gives students 300 seconds to learn as much as possible about a new language. The learners are shown an image of a figure when they click on one of the different buttons on the test panel. The figure is also provided with a corresponding sentence that describes the image. After the practice time is up, students are shown an image and two sentences, where one is correct and the other has one major grammatical error. The learners are asked to choose the correct sentence. This task measures the grammatical abilities of the learners.

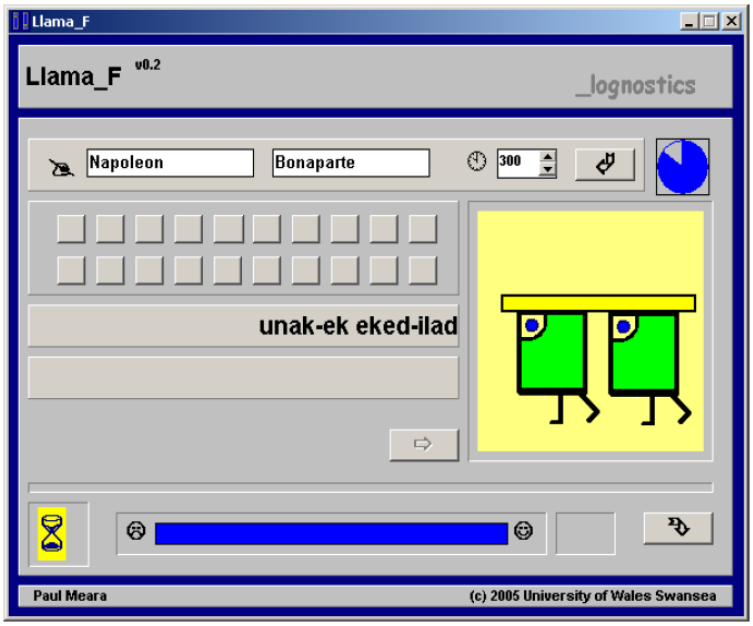


Figure 4 - LLAMA_F

2.4 Validating the LLAMA

As mentioned, since the LLAMA aptitude test battery was first published in 2005, a version quite different from the current shape it has today, the test has not been properly validated and standardized. Meara (2005) himself clearly states in the manual of the test that "[t]he materials provided to you are exploratory versions of on-going research, and they should NOT be used in high-stakes situations where accuracy and reliability are at a premium" (p. 21). The need for validation and standardization is something that takes time and effort from several studies from different researchers in a range of linguistic milieus. Granena (2013) explains that "[t]he LLAMA has not been extensively standardized as it has not been administered to a large variety

of ages, racial- and socioeconomic backgrounds" (p. 113). The LLAMA has become very popular among researchers and with over 700 citations since 2013, it seems to be a preferred choice for many empirical studies on language aptitude (Rogers et al., 2017). Still, there has been done a lot of validating studies on the LLAMA in recent years to prove that it is on its way to become a reliable and validated aptitude test.

Granena (2013) conducted an exploratory validation study on the LLAMA that assessed the reliability of the test and explored its underlying structures with a sample of 186 participants. The results showed that the internal consistency and stability in time were acceptable. Furthermore, a study by Meara himself and other colleagues (Rogers et al., 2016), also concluded with an acceptable level of reliability for the LLLAMA by investigating how several factors such as gender, language and education affected the test results. The only important factor that did inflict substantially on the test was age. Younger L2 learners would probably need other norms for testing than older learners. The study investigated as much as 229 participants from several countries, with different linguistic backgrounds (Rogers et al., 2016). A continuation of the study was also done by some of the same researchers one year later. The basis of test subjects was then raised to 404, giving the study a substantial base of corpus data. The results supported the findings on the impact of age found in Rogers et al. (2016) and pointed towards the LLAMA as being a reliable aptitude test battery. An important finding also suggested that previous L2 instruction could affect the results heavily (Rogers et al., 2017).

In addition to studies on the validation and reliability of the LLAMA, some research has also been done to examine the relation between LLAMA and age. Rogers et al. (2017) conducted a study where the LLAMA was tested on as many as 240 participants, with various linguistic backgrounds. One of the research questions for the study was to examine what effect age had on the LLAMA test scores. By conducting thorough testing, Rogers et al. (2017) came to the conclusion that "the current LLAMA tests are not suitable for use with younger learners." This conclusion is based on results showing that two groups of older learners aged 30-70 and 20-21, outperformed a group of younger learners aged 10-11 when testing the group's aptitude with the LLAMA_B and LLAMA_D. These results should suggest caution when using the LLAMA tests with younger learners and that is the reason I chose to safeguard my project by choosing the oldest possible learning group in the Norwegian Educational System, where every pupil still had the same prerequisites to perform on an aptitude test like the LLAMA.

Another important research question the study from Rogers et al. (2017) investigated was whether the LLAMA aptitude tests could be considered as language neutral. The study examined 195 L1 speakers of English, Chinese and Arabic to assess whether language had an impact on the test results of not. They all conducted all four of the LLAMA subtests and results were compared by controlling other individual factors among the participants such as L2 instruction, that could potentially affect the results. The results showed that none of the language groups outperformed the others, even though the main hypothesis was that the English L1 group would outperform the other two groups. These results point toward the conclusion that the LLAMA is in fact language neutral and can be used across different L1s without any interference on the results. This conclusion was also supported by the study from Granena (2013) which examined 187 Chinese, Spanish and English participants. The results from this study also showed that L1 had no effect on the LLAMA results. This contrasts many other aptitude tests, such as the MLAT, which has to be administered and developed into multiple versions in order to be applicable to different L1 groups (Rogers et al., 2017).

Bokander and Bylund (2019) examined the LLAMA data from 350 participants and assessed the data using classical item analysis, Rasch analysis, and principal component analysis within a framework of best practices in educational and psychological test validation. The results showed that only LLAMA_B, of the four, produced satisfactory scores that had sufficient accuracy. They suggest a further careful approach to using the LLAMA and propose that there is a potential for refining the aptitude test battery further (Bokander & Bylund, 2019). There has also been focus on the lack of research on general proficiency in L2 learners and its connection to the LLAMA. Bokander (2019) conducted a study where he tested 93 newly arrived university students with a range of L1 backgrounds. They participated in a Swedish language course for beginners and were tested for how their scores on the LLAMA correlated with their scores on a C-test, measuring general L2 proficiency. The results showed that the LLAMA_D seemed to be a valid predictor of initial L2 learning. Some methodological considerations were also highlighted, such as the effect of applying robust statistics, as well as using tasks of appropriate difficulty when subsets of participants may be expected to perform at different proficiency levels (Bokander, 2019). In all, a weaker validity was found for LLAMA_D, LLAMA_E and LLAMA_F. It seems that LLAMA_D is the subtest with the strongest validity at this current moment, although the other subtests still show correlation, but a weaker significant correlation.

2.5 Aptitude Testing in Instructed Language Learning

Instructed second language acquisition (ISLA) has been defined as "any systematic attempt to enable or facilitate language learning by manipulating the mechanisms of learning and/or the conditions under which these occur" (Housen & Pierrard, 2005). In order to understand the functionality of the LLAMA aptitude tests in an instructed setting it is key that some aspects of the previous research that has been done in such contexts are covered and that the possible learning- and research outcomes of such research is presented here. This section will provide insight into some of the most central research that has been done on aptitude testing (especially using the LLAMA) in instructed language learning.

Historically, language aptitude has been closely connected to language learning and the language education offered to pupils in instructed language learning settings. Carroll and Sapon (1959) created the MLAT in order to predict language learning success. This was done on behalf of schools and the government in order to pick the best language learners for intensive foreign language training and to match pupils in the educational system with the proper language learning classes (Hummel, 2013). Since aptitude was seen as a stable and latent trait in language learning at this time, aptitude testing was not done much in relationship to pedagogy and the idea of language development and formative assessment. Aptitude was seen more as a useful tool that could be used outside the actual teaching and more for the reasons of placement. When it comes to teaching approaches, the MLAT and the PLAB and other older aptitude test batteries, are based more on a task based and audiolingual approaches, making it rely more on the skills needed for achieving language learning success in older conceptualizations of instructed language learning in the Norwegian educational system (Erlam, 2005). According to Peter Robinson (2002) these older aptitude tests are on the verge of being outdated and he explains that "[g]iven the changing nature of classroom instruction since the 1959s and 1960s, however, it is questionable whether these tests are optimally predictive" (p. 117).

As aptitude testing is most often done in some sort of educational language learning situation and the participants of the different studies are frequently learners of an L2 language that are submitted to some sort of educational institution, aptitude testing has frequently been done in order to examine several aspects of teaching and to investigate pedagogical advantages in language teaching. In recent years, the LLAMA has been used as a tool in several studies done on different aspects of language learning and aptitude. The reasons for conducting studies on aptitude in relation to pedagogy is often to find pedagogical advantages by knowing the

learning style and aptitude level of the pupils. Profiling individual differences in cognitive abilities and matching these profiles to effective instructional options is one of the major aims of pedagogically oriented aptitude research (Robinson, 2002). The design of these studies could often be to find alternative methods for teaching the same content, based on the individual learning it, randomizing assignments of students to treatments and initial testing to measure propensities hypothesized to be more relevant to one treatment than another (as cited in Robinson, 2002). Below, I will present some studies that tap into the typical aspects of language learning aptitude research and pedagogy. The studies I have chosen are making use of the chosen aptitude test battery for this paper, namely the LLAMA, as testing method in various degrees and for different purposes.

Sternberg, Grigorenko and Zhang (2008) researched what they called an aptitude–treatment interaction. In this study, students who were placed in instructional conditions that better matched their pattern of abilities outperformed students who were mismatched. The participants were divided into two groups of different sets of learning and thinking. The first set of learning and thinking that was examined was the ability-based teaching approach. This approach has shown to be beneficial to the language learning process, as pupils will be challenged to use their cognitive thinking in relation to the knowledge presented. Therefore, Sternberg, Grigorenko and Zhang (2008) urge teachers to "teach and assess achievement in ways that enable students to analyze, create with, and apply their knowledge" (for teaching techniques see Sternberg, Grigorenko & Zhang, p. 487). The other set of learning and thinking that was examined in the study is called a personality-based style. This style of learning essentially entails a preference for using abilities in certain ways, i.e. how one likes to use one's abilities when learning. This personalities are roughly divided into three groups: a legislative style, that consists of learners who has a predilection for tasks, projects and situations that require creation, formulation, planning of ideas, strategies, products and the like. The second personality group is called the executively oriented students. This group has a predilection for tasks, projects and situations that provide structure, procedures, or rules to work with. The third personality group is the judicially oriented student that has a predilection for tasks, projects and situations that require evaluation, analysis, comparison and contrast and judgement of existing ideas, strategies, projects and the like (Sternberg, Grigorenko & Zhang, 2008). Another important point made in the study is that a teacher can also have certain personality-based style that can affect the teaching and learning outcomes. The study shows how important it is to take the differences among students into consideration when teaching and applying good teaching

practices and pedagogical methods to the different students by applying adapted and varied education in the classroom (Engelsen, 2012). In other words, when students are taught in a way that fits the way they think, they do better in school (Sternberg, Grigorenko & Zhang, 2008).

A new way of understanding the concept of aptitude has also been proposed in recent studies where aptitude has been divided into two separate constructs. Linck et al. (2013) showed a possible distinction between an explicit language aptitude and an implicit language attitude. The constructs hypothesized to tap into explicit language learning were skills such as explicit induction and rote memory, whereas the constructs hypothesized to tap into implicit language learning were skills like primability and implicit inductive learning ability (Linck et al., 2013). These different aptitude types might be important for which teaching style the teacher chooses and what kind of assignments are given to different pupils. Granena (2013) also made a similar distinction between aptitudes for implicit and explicit learning in a series of exploratory factor analyses. She concluded with an aptitude dimension interpreted as analytic ability, relevant to explicit learning, and sequence learning ability, relevant for explicit language learning (Granena, 2013). The research on these has also been further developed into connecting implicit and explicit language aptitude to different cognitive styles.

Pacini and Epstein (1999) have proposed a framework where two main information processing cognitive styles are being used in language learning. The first cognitive style is the rational-analytical style which is strongly related to "Ego Strength, Openness, Conscientiousness, and favorable basic beliefs about the self and the world, and it was most strongly inversely related to Neuroticism and Conservatism" (Pacini & Epstein, 1999). The second cognitive style was called an experiential-intuitive style and was related to "Extraversion, Agreeableness, Favorable Relationships Beliefs, and Emotional Expressivity, and it was most strongly inversely related to Categorical Thinking, Distrust of Others, and Intolerance" (Pacini & Epstein, 1999). These two cognitive styles were connected to aptitude constructs.

Granena (2016) also used the LLAMA to investigate the connection between cognitive styles and different language aptitudes. The participants were 82 Chinese first language-Spanish second language speakers. The study used the LLAMA, a probabilistic SRT task and the Rational-Experiential Inventory (REI), a test for measuring cognitive style, developed by Epstein (1990). The results from the study were of interest to this study because the different subtests of the LLAMA seemed to tap into different types of cognitive aptitudes. The results

from LLAMA_B (vocabulary learning), LLAMA_E (sound-symbol correspondence) and LLAMA_F (grammatical inferencing) all seemed to correspond with rational engagement (reliance on and enjoyment of thinking in an analytical and logical manner) which suggest that they are related to a more explicit aptitude (Granena, 2016). These abilities draw on the participants analytical skills and could therefore be considered a measure of explicit language aptitude, tapping cognitive abilities such as explicit indictive ability, explicit associative learning ability, and rote learning ability. These are cognitive abilities that can be expected to play a role in inducing rules behind a set of examples in an unknown language (LLAMA_F) and in the learning of associations acquired consciously and intentionally between drawings and word strings or between sounds and symbols (LLAMA_B and LLAMA_E). On the other hand, LLAMA_D (sound recognition) gives participants no time to rehearse and memorize materials and Granena (2016) therefore argues that this subtest relies less on the connection between rational ability and performance, and therefore subsequently also less on the participants analytical problem-solving abilities. This means that the cognitive abilities that are being used when using LLAMA_D relies more on implicit aptitude (Granena, 2016).

There has also been done extensive research on different learning outcomes for different learner types and other pedagogical conclusions drawn from using the results of the LLAMA subtests. Poschner's (2018) recent study on vocabulary learning skills and its connection to language learning strategies is closely related to answering the question of how to apply the results from the LLAMA to practical teaching and pedagogy. The study used the LLAMA_B to measure the vocabulary acquisition aptitude of 19 German native speakers. They were later given a questionnaire to examine the participants preferences for various cognitive vocabulary acquisition strategies. The main findings of the study were that there is no difference between the use of cognitive vocabulary strategies between high- and low scorers of the LLAMA_B, i.e. students with a high- or low vocabulary learning aptitude. The cognitive strategies in question here are mnemonic strategies, learning with pictorial representations, the use of synonyms and antonyms, grouping words together in meaningful groups, and using no specific technique or strategy. The study shows that the low vocabulary scorers might experience great benefits by using these vocabulary learning strategies. Still, they do not use the strategies in their own learning process. It is therefore important for the teacher to focus on these strategies when working on vocabulary acquisition for low scorers of the LLAMA_B. The high scorers are not more aware of these strategies than the low scorers, but they do not seem to have as

much use and as high a learning benefit by using these strategies as the low scorers do (Poschner, 2018).

The LLAMA_D has also been subject to research on learner outcomes in relation to language learning aptitude. The discussion has often revolved around how aptitude relates to the proficiency one can expect from a participant with different scores on the LLAMA. Carroll and Sapon's (1959) MLAT was initially designed to predict learner success in acquiring a second language. The LLAMA has been subjected to some studies that have investigated the correlation between scores on the LLAMA and oral production in English. Maddah and Reiterer (2018) showed that scores in the LLAMA_D test revealed a significant, positive relationship with the subjects' English pronunciation score ($r = .66$) in their study of 30 Iranian L1 Farsi learners of English. The correlation between LLAMA_D and pronunciation proved that subjects with better short and long-term memories could achieve a higher native-like attainment in the pronunciation of a second language. Still, research has also been conducted on the connection between oral proficiency and language learning aptitude. The conclusion from a study by Yalcin (2012) stated that there seemed to be little or no connection between LLAMA scores and scores on oral performance tasks. The same findings were also presented by Saito (2017) in his study of 50 Japanese EFL learners who were analyzed through a range of pronunciation-, fluency-, vocabulary- and grammar measures. Still, one would expect learners who gain a high score on the LLAMA_D and LLAMA_E to be more precise in their oral production, as these two tests tap into the learner's sound-symbol correspondence understanding and their sound recognition abilities. High scoring LLAMA_D individuals would also typically rely on intuition and a more holistic approach to information processing and may therefore be better at learning complex patterns or hidden covariations in the environment implicitly (Granena, 2016).

Another study has investigated the positive correlation the connection between specific learning conditions and teaching methods and a particular aptitude profile might have. Erlam (2005) shows that there is a strong connection between aptitude scores and pedagogical choices when teaching. She suggests that pupils with a high analytic ability should be taught with an inductive approach and a structured-input method. The pupils with this skill would be those who typically score high on the LLAMA_F. This means that they would benefit from being exposed to examples of the target language and then asked to figure out the rules that govern it. It is a kind of induction that lets the pupils explore the language themselves before it is structured for them. The structured-input method means that the instructor presents input that

is manipulated in ways that push learners to become dependent on form and structure to get meaning (Lee & Van Patten, 2003). Activities that support the structured-input method are supplying information, matching, binary options, ordering/ranking and selective alternatives.

Furthermore, Erlam (2005) also studied the approach that is best used for teaching an entire class where the test results from the LLAMA are somewhat mixed and difficult to group and adapt a more individualized instruction based on the pupil's aptitude scores. If you have a class with a highly mixed level of language aptitude the deductive approach should be used for teaching (Erlam, 2005). That means you should explain the concepts and rules of the language firstly and then later introduce examples and relevant situations in which the previously learned skills can be used. Another general advice for teaching mixed aptitude groups comes from the research of Kourtali and Revesz (2019) who suggest that low-complexity tasks have the capacity to minimize the degree to which learner differences in L2 aptitude predict development in task-based contexts when feedback is available. In other words, when the aptitude of the class is mixed, uncertain or too complex to differentiate the teaching methods of grammar, less cognitively demanding tasks should be introduced so that the development of grammatical knowledge can be more efficient throughout the whole intact language class.

As a final note on the research done with regards to aptitude tested by the LLAMA for instructional language learning purposes, a look at feedback on the language learning process should be presented. Yilmaz (2013) studied the two cognitive factors, working memory capacity (WMC) and language analytic ability (LAA) with 48 native speakers of English who were exposed to an unknown target language. WM has a central role in cognitive SLA research and Engle (2007) define the concept as "attentional processes that allow for goal-directed behavior by maintaining relevant information in an active, easily accessible state outside of conscious focus, or to retrieve that information from an in-active memory, under condition of interference, distraction or conflict" (as cited in Yilmaz, 2013). These learner abilities were measured up against two forms of feedback, namely explicit correction and recasts. The LLAMA_E was used to measure the LAA of the participants. Results showed that explicit correction worked better than recasts only when the learners in the compared groups had high cognitive ability (high WMC or high LAA), i.e. achieved a high score on the LLAMA_F subtest. That means high scorers on the LLAMA_F and probably also LLAMA_E, since this subtest also measures language analytic ability, would benefit more from being explicitly corrected and presented with the correct answer when they are mistaken, rather than being asked to repeat a correct structure or word, uttered by the teacher.

3. Methodology

This chapter will present the method employed in order to realize the aims of this master's thesis. The methodology in this paper is chosen to answer the research questions[1] asked previously in this paper as thorough as possible. The experiment was conducted with permission from the Norwegian Centre for Research Data (NSD), and an informed consent was retrieved from all participants in this study: the teacher of the class and the pupils took part in the testing. The pupils were given a letter of consent (see Appendix A) to fill out with all the information about the test and the master thesis the testing would be used in. The pupils were also urged to inform their parents about this project even though this was not a requirement, as the pupils were all over the age of 15.

This methodology section will first present how the treatment of the experiment was conducted, i.e. what was done in the pilot and experiment group of the study. This section will also present the steps done both before and after the actual testing. Second, the chosen participants and selection criteria of the intervention will be presented. Finally, the measuring tools and procedure of the experiment are described in detail. A rationale for the chosen methods will also be provided for every instrument used in the testing procedure. Every step in the experiment will also be connected to a specific research question for an explanation of why the method in question was used.

3.1 Research Design

The research design that has been chosen is that this project will be carried out as a case study. A case study typically focuses on a small number of research participants, usually language learners (Duff, 2012). As this project focuses on one language learning class this type of design

[1] RQ1: Is the LLAMA a suitable aptitude test battery for a lower secondary school class?; RQ2: How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?; RQ3: Does the LLAMA aptitude test correspond with the teacher's perception of the pupils' aptitude?; RQ4: What pedagogical advantages can aptitude testing have in an upper secondary class in English?; RQ5: What attitude does the teacher have towards aptitude testing and the LLAMA?; RQ6: Will the teacher make pedagogical changes to the English education based on the aptitude results, and if so, what type of changes?

fits the project well. When you have chosen the participants of the case study "[t]he individual's behaviors, performance, knowledge, and/or perspectives are then studied very closely and intensively, often over an extended period of time, to address timely questions regarding language acquisition, attrition, interaction, motivation, identity, or other current topics in applied linguistics" (Duff, 2012, p. 95). The "case" in this case study is represented by the functionality of the LLAMA aptitude testing in the classroom of an upper secondary school first year class. The case study functions as an exploratory research intervention where the goal of the study lies more in studying several aspects with using aptitude as a tool in the Norwegian educational system.

The design of the study is structured by using both quantitative and qualitative methods of retrieving data. The intervention is structured by firstly doing some close studies of the participants and afterwards using interviews with the teacher to assess how the aptitude testing was carried out and what attitude she has towards the LLAMA, as well as her perception of pedagogical advantages with using aptitude testing. The first action that was conducted in the intervention was asking the teacher of the class to familiarize with the LLAMA test and the notion of language aptitude. The teacher was then asked to assess the expected aptitude level of the pupils before the testing. Then, the LLAMA aptitude test was conducted, followed by a questionnaire that was handed out to every participant. These instruments make up the qualitative part of the research design and were all conducted one the same day, a couple of months into the schoolyear. After the classroom intervention, two independent interviews were conducted with the teacher of the class. The first interview included questions that investigated the teacher's perception of the actual testing procedure, and thus occurred immediately after the classroom intervention. The second interview focused more on how the results from the aptitude testing were used by the teacher and was therefore conducted towards the end of the schoolyear. The different steps of the intervention and experiment is illustrated in figure 1 below.

Pilot	Intervention
• Aptitude Prediction • LLAMA Test • Questionnaire	• Aptitude Prediction • LLAMA Test • Questionnaire • Teacher Interview 1 • Teacher Interview 2

Table 1 Illustration of the content of the pilot and the experiment.

3.2 Pilot

The intervention was conducted as a pilot before the actual testing. The pilot consisted of a group of nine participants. This was done in order to ensure the success of the main experiment and to discover possible flaws or issues that might occur when conducting the tests with the main group. The participants for the pilot were 16 or 17 years old and attended an elective English class called “International English” on their second year of upper secondary school. The pilot was conducted with the same aptitude tests and the same procedure for carrying out the study as the main project.

The pilot revealed some major issues with the LLAMA that were important to address before the main testing. The first problem the pilot revealed was that the two subtests that relies on hearing, i.e. LLAMA_D and LLAMA_E, did not function with a wireless hearing device called AirPods. The sounds of the test got obscure and shortened, making it impossible to discriminate the sounds for the participants. As this particular hearing device was popular among the pupils, alternative wired hearing devices had to be used in order to make sure that this problem did not obstruct the main testing. As the wireless AirPods were replaced, the sound recognition tasks were successfully conducted without any further issues regarding sound quality.

Another issue that was revealed during the pilot testing lay in the LLAMA subtests. Both of the errors that were found in the LLAMA subtests occurred in the LLAMA_F, a

grammatical inferencing task. The first of two errors occurred early in the test where the test asked the pupils to discern between what could be translated to either “Green balls” or “Balls Green.” The error in this situation lay in the fact that the test formulated “balls” as “squares,” and thus making it impossible to match the image of the two green balls with a correct sentence. This error was corrected in the main test, where pupils received the correct answer on a PowerPoint slide (see Appendix B). The second error also lay in the LLAMA_F, but as one of the last tasks to be carried out in the subtest. The pupils were asked to choose the grammatically correct sentence that described three red triangles. After a thorough examination of the test, where both myself and my supervising teacher looked several times at this particular task, we concluded that it would be impossible for the pupils to answer this task, as they do not possess enough information about the grammar of the red triangles to choose the correct answer. This error was also corrected by showing the pupils the correct answer on a PowerPoint slide.

3.3 Participants

The participants for the project were 22 upper secondary school pupils. Together they form a first-year class aged 15 or 16. The pupils had all received a traditional 10-year English education, which functions as the standard structure of the English education in the Norwegian Educational System. In light of this, it is reasonable to assume that the participants also possessed the standard level of English proficiency of what you could expect from a first-year Norwegian upper secondary school class. All of the pupils except one had Norwegian as their first language.

The participants were selected on the following criteria: age, language and education. I selected an upper secondary school group because I felt confident that this level of English education would provide participants where age would inflict as little as possible on the test results. This rationale is based on the conclusion made by Rogers et al. (2017) that states that the LLAMA is unsuited for younger learners (see section 2.4). A possibility was also to choose older learners, but the classes would then consist of those who voluntarily had chosen English as an elective subject and the test results could then be affected by the possibility that those who choose to study English voluntarily might have a higher language aptitude than others. As for younger learners, I chose not to include learners younger than 15, because of the uncertainty that the young age of the participants might affect the results. The LLAMA-tests have not been

given a specific age group for which it should be administered to (Meara, 2005). To compare, the MLAT comes with an age recommendation of 14 years and above. To cope with the problem of younger learners, Carroll and Sapon developed an elementary form of the battery, called the EMLAT. This battery is adjusted to fit children between the ages of eight and eleven (Skehan, 1989). Since no similar age reduced version of the LLAMA exist and since the age question is so unclear, I chose to safeguard and choose as old learners as possible, while still ensuring that they were within the scope and goals of this project.

3.4 Aptitude Prediction

Before the initiation of the aptitude testing in the classroom, the teacher was asked to assess the aptitude of the pupils. This was done in order to answer RQ3[2]. The teacher was given an aptitude assessment sheet (see Appendix B) that was used to try to predict each pupil's aptitude level on a scale of 1 – 4. Then, these predicted levels of aptitude were matched with the actual score the participants received on the LLAMA test. For each score of 1 - 4 a description of what that particular aptitude rating entails is provided on the sheet. This description more or less corresponds with how the different scores of the LLAMA are to be interpreted in the LLAMA manual (Meara, 2005).

A participant that gains a score of 0-15 is placed in aptitude level 1, which is described as follows: "pupils that often have a hard time learning English and uses a lot of effort and work to acquire new domains of the language" (Meara, 2005). Level 2 is for participants that gain a score of 20-45. For these pupils, "learning English comes natural and they do not have more difficulties in the learning process than you would expect of a general language learner" (Meara, 2005). Participants that gain a score of 50-70 are placed in level 3 and "easily acquire new knowledge of an unfamiliar language without too much effort" (Meara, 2005). The highest aptitude level was level 4, comprising those who gain a score of 75-100 on the LLAMA tests. These pupils have "exceptional language aptitude and their language talent is far beyond what you would expect from learners at the same stages of a foreign language education" (Meara, 2005). Placing every pupil into one of the four categories was a challenging task for the teacher,

[2] Does the LLAMA aptitude test correspond with the teacher's perception of the pupils' aptitude?

as this is no exact science and knowledge about a language learners' aptitude can be difficult to perceive without conducting an aptitude test.

3.5 Aptitude Tests

After the predictions were done, the actual aptitude testing of the pupils was conducted using the LLAMA aptitude test battery. This testing was done mainly to answer RQ1[3]. By using the results from the testing, accompanied by the results from the questionnaire this research question should be able to answer.

The testing of the participants was conducted by using the four tests of the LLAMA aptitude test battery, namely: LLAMA_B, a vocabulary learning task; LLAMA_D, a sound recognition task; LLAMA_E, a sound-symbol correspondence task and LLAMA_F, a grammatical inferencing task (Meara, 2005). This aptitude test battery was chosen above other test batteries like the MLAT and the PLAB, for several reasons. Firstly, the LLAMA is free to use and does not require any permission or payment to administer and it does not include a commercial side at all (Granena, 2013). The LLAMA is developed by the University of Swansea and does not include any sort of political restrains at all (Meara, 2005). To compare, test batteries like the MLAT and the PLAB are expensive if they are supposed to be administered to large amounts of participators and also functions with a certain political background from the developmental side. The political aspect of these tests is that they are made on order from a governmental institution, to serve a specific purpose. To compare, the LLAMA is developed by the University of Swansea, for language research only. The LLAMA is also easy to administer and has a user-friendly set up that can easily be understood without any problems (Granena, 2013). The fact that the LLAMA is language neutral also makes it easier to administer, as the whole testing can be carried out in Norwegian, rather than English (Rogers et al., 2017). This efficiently rules out possible issues with pupils failing to understand tasks or commands because of a lack of proficiency in English. Because of these reasons, the LLAMA was chosen as the aptitude test battery for this study, instead of other aptitude test

[3] Is the LLAMA a suitable aptitude test battery for a lower secondary school class?

batteries like the MLAT, the PLAB, the HiLAB, the CHANNEL-F or other similar aptitude test batteries.

The LLAMA subtests were conducted with all 22 participants at the same time, and the testing did not continue until every participant was done with each subtest. I carried out the testing by explaining every subtest to the class, while using a PowerPoint presentation (see Appendix C) to illustrate and show the pupils where they were supposed to click and how the tests functioned. After each subtest, the pupils were asked to take a note of their score on a result sheet (see Appendix D). The scores would range between 0-100 for each subtest, describing whether the score was bad, average, good or exceptionally good. The pupils were also handed separate sheets to take notes on, as LLAMA_E and LLAMA_F allows for notes to be taken by participants. The note sheets were also collected after the testing was done.

3.6 Questionnaire

After the actual testing, a survey in the form of a questionnaire (see Appendix E) was handed out to the pupils. Surveys can be important tools for understanding the underlying structure of a language learning process and the attitudes pupils have towards aspects of this process. Dörnyei and Csizèr (2012) concludes that "[i]n sum, surveys can target a wide variety of language-related issues and allow researchers to make inferences about larger L2 learning populations…" (p. 75). The questionnaire is intended to provide data concerning the pupils' overall experience with using the LLAMA. The questionnaire was mainly aimed at answering RQ2[4].

The first question was made to clarify the gender of the participants. The main part of the questionnaire was built up in the form of a Likert scale where pupils were asked to rate five statements from 1 to 5, where 1 on the scale was described in words as "not at all," "very bad" or "very difficult to understand." The highest score on the scale, 5, was described in words like "like it a lot," very good" or "very easy to understand." A typical question from the questionnaire was "Do you think this test, and knowing your language aptitude, is useful for

[4] How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?"

your education?" After the five statements, a rating task was given where the pupils were asked to rank the four subtests of the LLAMA (LLAMA_B, LLAMA_D, LLAMA_E and LLAMA_F) on a scale of 1-4, where a rating of 1 represented the subtest the pupil liked the most, and a rating of 4 represented the subtest the pupil liked the least. Images of the different subtests were also provided to make the recognition of the names of the different subtests easier. This section was added to examine if the experience the pupils had with using the LLAMA was tied to specific subtest or more towards the test battery as a whole and additionally, if some of the subtests were more favourable with the pupils than others. Lastly, the pupils were given a chance to add whatever comments they might have regarding the questions in the questionnaire, the LLAMA or any other comments to the subject of language aptitude in school.

3.7 Teacher interviews

After the testing, two separate interviews with the teacher of the class were conducted to investigate her perception of the test and attitudes towards aptitude testing and also the potential pedagogical implications such testing can have. The questions were audio recorded and later transcribed to this paper. This recording was done with permission from the teacher and the Norwegian Centre for Research Data (NSD). This section will present the methodology applied to conduct the interviews and the questions.

3.7.1 First Teacher Interview

The first teacher interview was conducted mainly in order to answer RQ5[5]. This interview was conducted immediately after the aptitude testing was done. The interview and the testing were done approximately three months into the schoolyear. This was necessary, as by time, the teacher would have had enough time to get to know the class to such an extent that predictions on their aptitude could be made. The 9 questions that were asked in this interview were:

1. How did you like the test?

[5] What attitude does the teacher have towards aptitude testing and the LLAMA?

2. How do you think carrying out the aptitude in class went?

3. Are there any difficulties/problems with using this type of tests?

4. Do you find aptitude testing useful in teaching English?

5. Is this a test you would consider using in the future?

6. Do you think such tests and the knowledge of a pupil's aptitude is useful and interesting to you as an English teacher?

7. How can a test like this inform your English teaching?

8. What other screening tools do you have in English in upper secondary school?

9. Do you think the LLAMA can function together with these tools?

3.7.2 Second Teacher Interview

The second teacher interview was conducted around six months later than the first interview, when the school year was about to come to an end. This interview was conducted mainly in order to answer RQ4 [6]and RQ6[7]. By knowing the aptitude of the pupils early in the schoolyear and afterwards using this as a part of the education, the teacher was supposed to gain insight in the possible pedagogical advantages with aptitude testing in the classroom. The 9 questions that were asked in this interview were:

1. Has the LLAMA test scores been useful to your teaching this year?

2. How do you view this test compared to other mapping tools in English?

3. Have you altered or modified your teaching in any way because of the test results?

---

[6] What pedagogical advantages can aptitude testing have in an upper secondary class in English?

[7] Will the teacher make pedagogical changes to the English education based on the aptitude results, and if so, what type of changes?

4. Regardless of if you have altered your teaching or not based on the test results: What potential pedagogical advantages can you see by using the LLAMA and aptitude testing in general in an instructed language learning situation?

5. Have you seen any development in the aptitude of the pupils during the year?

6. Have you experienced any developmental advantages in learning English with the pupils that scored high on the LLAMA, compared to the pupils that had a lower score?

7. Have you recognised the test scores from the LLAMA in the pupils' development in the subject and English proficiency this year?

8. Again. After using the method of aptitude testing and seeing the effects over a whole school year. Is it something you would consider using in the future or recommend others using?

9. Do you have any further comments, notes or statements?

4. Results

This section will present the results from the class intervention of this study. The results are gathered from the tasks described in the methodology section earlier in this paper. This section will be divided into two separate sections. First, I will present the quantitative data, including the aptitude prediction from the teacher, the results from the LLAMA subtests, and the participant questionnaire collected after the aptitude testing. Second, I will present the qualitative data from the two teacher interviews that were collected on two separate occasions: one was early in the schoolyear and the second one took place towards the end of the schoolyear to track progression. The results are presented in chronological order based on the order of tasks in the intervention.

4.1 Quantitative Results

This section will deal with the quantitative results from the class intervention. Duff (2012) describes quantitative research as "looking for such causal relationships or otherwise quantifying relationships or patterns. They test the significance of findings statistically and may employ an experimental design to test whether, and to what degree of certainty, development or change has occurred and, if so, how it might be characterized and accounted for." (p. 99). A quantitative approach involves the collection and analysis of numerical data in order to describe, explain, or predict phenomena for the purpose of research. The quantitative research methods in this intervention are the aptitude prediction from the teacher of the class, the actual LLAMA aptitude tests and the questionnaire that was handed out to the pupils after the LLAMA tests were conducted to answer the connected research questions[8].

[8] RQ1: Is the LLAMA a suitable aptitude test battery for a lower secondary school class?; RQ2: How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?; RQ3: Does the LLAMA aptitude test correspond with the teacher's perception of the pupils' aptitude?

4.1.1 Aptitude Prediction

An aptitude prediction test (see section 3.4) was conducted in order to investigate RQ3. This prediction was made by familiarizing the teacher with the scoring system of the LLAMA and then using that information to try and predict how well each participant would perform on a scale of 1-4, where 1 is the lowest score and 4 is the highest (range = 4). The teacher used the aptitude assessment sheet (see Appendix B) to familiarize herself with the descriptions for each score (1-4) and the LLAMA aptitude score each aptitude level was connected to. (See section 3.4 for description of scoring system).

First, the results from the aptitude assessment show that an individual teacher assessment of a participant never resulted in a deviation of more than 1 aptitude level. This means that the teacher was never more than 1 aptitude level away from the aptitude prediction of each participant obtained via the LLAMA test. A deviation between the teacher's aptitude assessment of the pupils and the actual score they gained on the LLAMA test occurred in twelve out of 22 cases. This creates an overall mean deviation of 0.55 (MD = 0.55). Figure 5 illustrates the difference between the teacher's prediction and actual aptitude level obtained on the LLAMA test.

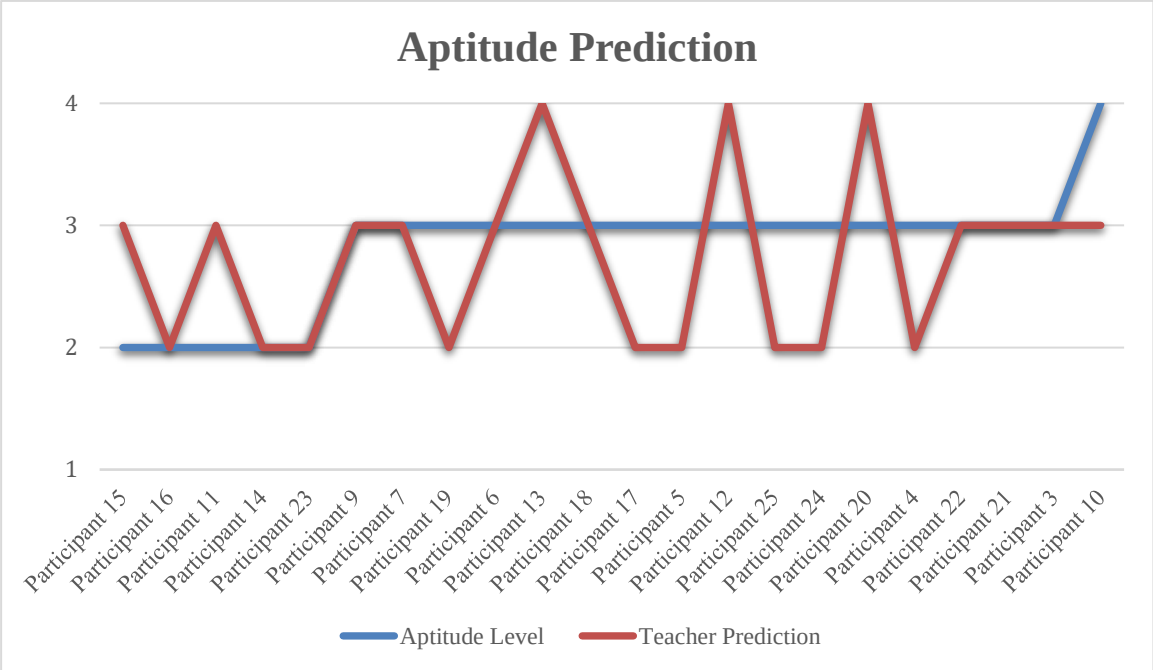


Figure 5 - Comparison of pupil aptitude levels and teacher predictions

The blue line in Figure 2 shows the aptitude level the pupils achieved in all LLAMA tests, while the red line represents the teacher's predictions. As the figure shows, the teacher gave a higher score to 5 out of 22 participants, but the difference is never larger than one (1) aptitude level. At the same time, the teacher gave a lower score to 6 out of 22 participants, but the difference was again never larger than one (1) aptitude level. The number cases when the teacher's prediction corresponded to the pupil's aptitude level obtained on the LLAMA tasks was 11 out of 22.

Another interesting aspect with the teacher predictions of the pupils' aptitude scores is how the teacher seems to rate pupils differently based on gender. The aptitude prediction was tested on 7 boys (n = 7) and 15 girls (n = 15). This is a relatively low number of participants to draw any conclusions about gender, but as the results for gender seems to stand out, I will include them in this section. As figure 6 shows, the difference between predictions on boys and girls from the teacher was substantial. For the boys, the correct assessment percentage from the teacher was 71%, a quite high accuracy score. On the other hand, for the girls, the teacher only predicted the correct aptitude level for 33%. The most striking result for the genders was still that the teacher did not underestimate the aptitude level of any boys. For the girls, the underestimation percentage was as high as 47%, indicating that the teacher believed almost half of the girls to have a lower aptitude level than what they actually had. The results for gender and aptitude prediction generally show that the teacher had a much greater trouble predicting the aptitude levels of girls than with boys. In addition, girls were underestimated by the teacher when predicting aptitude levels.

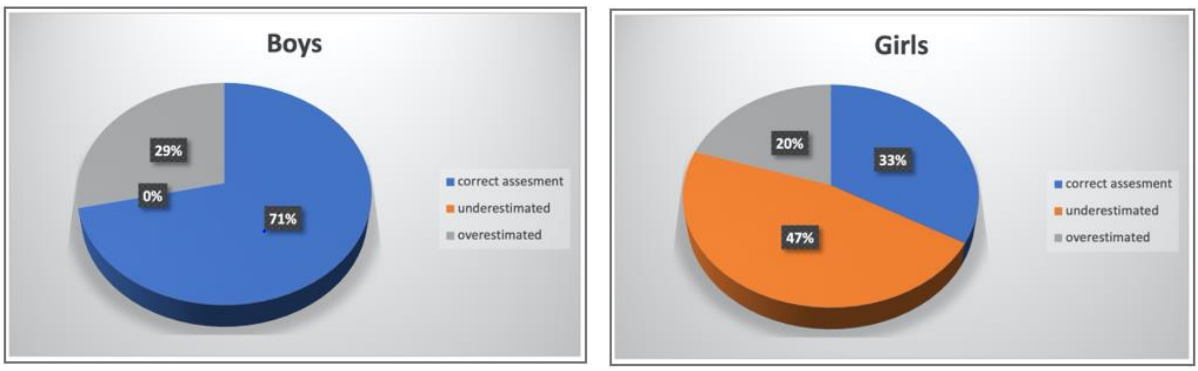


Figure 6 - Gender differentiation in aptitude prediction

4.1.2 Aptitude Tests

The LLAMA test was conducted in order to assess the functionality of the test battery for an upper secondary school class, as well as understanding the potential pedagogical uses of the test. The result from the test is supposed to help answer all of the research questions and in particular RQ1 [9]. The results from the LLAMA language aptitude subtests will be described in this section. The LLAMA aptitude battery tested the participants' aptitude level by using the four subtests called LLAMA_B, a vocabulary learning task; LLAMA_D, a sound recognition task; LLAMA_E, a sound-symbol correspondence task; and LLAMA_F, a grammatical inferencing task (Meara, 2005). I will present the results from the aptitude testing one subtest at the time. A gender distinction for the subtests will also be provided.

4.1.2.1 LLAMA Test Results

The results from the four subtests showed varying results and some subtests were substantially higher in mean scores than others. The results from the first subtest, LLAMA_B a vocabulary learning task, showed a mean score of 49.55 (SD[10] = 17.18, SE[11] = 3.66) and a median score of 47.50, with a 95% confidence interval[12] of 42.37 – 56.73. The range[13] was 80 – 25 = 55, with a minimum score of 25 and a maximum score of 80. A mean score of 49.55 indicates that the class generally fits slightly under the score that ranges from 50-70 which is described as "a good score" in the LLAMA manual (Meara, 2005). The mean score of 49.55 for LLAMA_B represented the second best result by the class, out of the four LLAMA subtests.

The results from the second subtest, LLAMA_D a sound recognition task, showed a mean score of 37.05 (SD = 14.03, SE = 2.99) and a median score of 37.50, with a 95% confidence interval of 31.18 – 42.91. The range was 70 – 10 = 60, with a minimum score of 10

---

[9] Is the LLAMA a suitable aptitude test battery for an upper secondary school class?

[10] SD = Standard Deviation: Measure of the amount of variation or dispersion of a set of values.

[11] SE = Standard Error: The standard deviation of a statistical sample population.

[12] Range of values that are 95% confident to contain the true mean of the population, between a lower and an upper interval.

[13] The difference between the lowest and the highest score of a population.

and a maximum score of 70. A mean score of 37.05 would indicate that the class generally fits slightly above the score that ranges from 15-35 which is described as “an average score; most people score within this range” in the LLAMA manual (Meara, 2005). The mean score for this subtest indicates that this is the subtest the class struggled the most with.

The results from the third subtest, LLAMA_E a sound-symbol correspondence task, showed a mean score of as much as 88.46 (SD = 17.91, SE = 3.82) and a median score of 100, with a 95% confidence interval of 81.15 – 96.12. The range was 100 – 30 = 70, with a minimum score of 30 and a maximum score of 100. A mean score of 88,46 indicates that the class fits into the highest range of scores for this subtest 75-100 which indicates “an outstandingly good score. Few people manage to score in this range. Those who do are mostly trained linguists” in the LLAMA manual (Meara, 2005). This subtest was the highest scoring subtest for the class.

The fourth and final subtest, LLAMA_F a grammatical inferencing task, showed a mean score of 47.27 (SD = 25.08, SE = 5.35) and a median score of 50, with a 95% confidence interval of 36.79 – 84.07. The range was 100 – 0 = 100, with a minimum score of 0 and a maximum score of 100. A mean score of 47.27 fits slightly above the range of 20-45 which indicates that the class achieved “an average score; most people score within this range” (Meara, 2005). In figure 7, the mean scores for the LLAMA subtests are visualized.

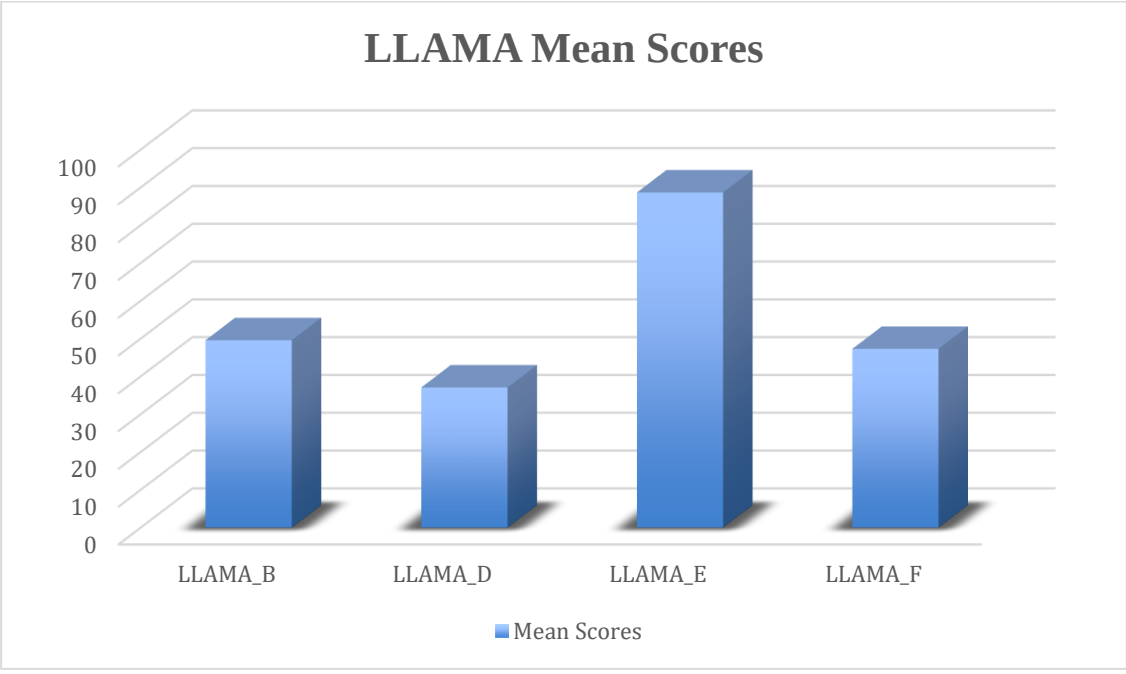


Figure 7 - Mean scores for the LLAMA subtests

An interesting finding from the aptitude testing was connected to gender differences in aptitude subtest scores. Gender differences has been studied in many forms with regards to aptitude and aptitude testing. For instance, Habl (2018) has investigated the correlation between gender and aptitude measured by the LLAMA with little differential results. The current intervention included 15 girls (n=15) and 7 boys (n=7). None of the participants (n=0) affiliated themselves as "other/prefer not to answer" in the question regarding gender in the questionnaire (see Appendix E). The results showed that the difference in mean score based on gender was 6.7 in which the boys were the best scorers and outperformed the girls. Still, the major differences in gender-based results did not lay in the overall mean scores, but in mean scores for the different LLAMA subtests. Figure 8 provides a visual overview of the gender differences in score for each of the four LLAMA subtests.

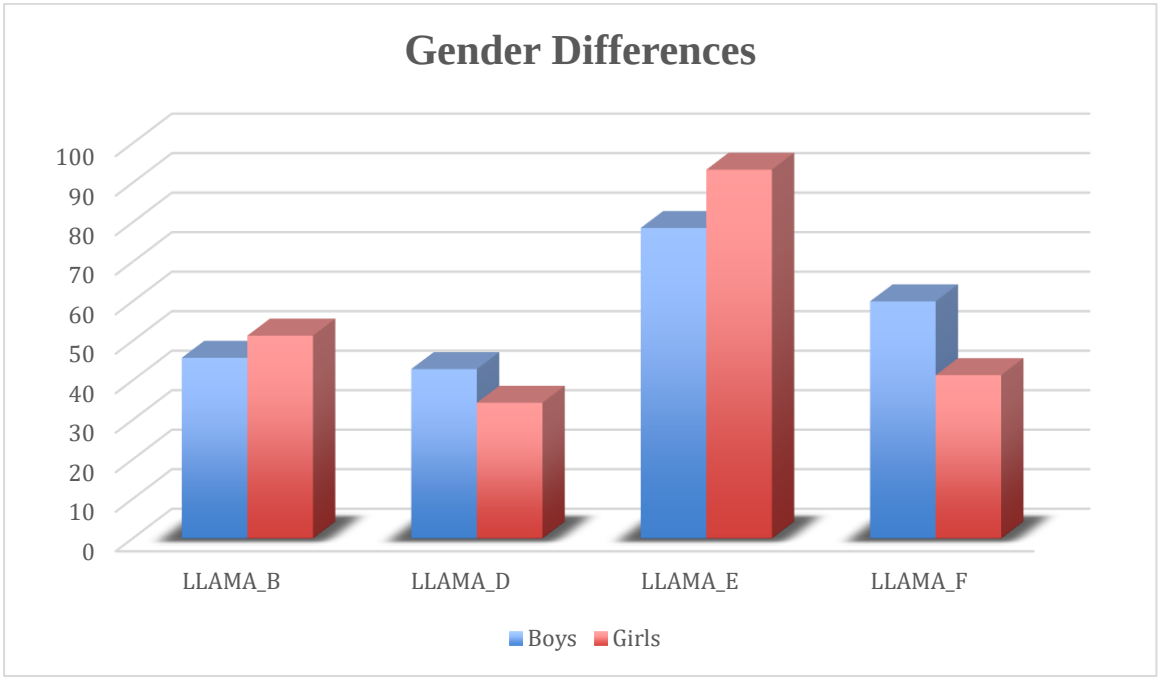


Figure 8 - Gender-Based results for subtests

For LLAMA_B, the difference in mean score was 4.62, where the girls had a mean score of 50.33 and the boys had a score of 45.71, making the difference in gender for this subtest insignificant (5>). The difference between boys and girls in the mean score for LLAMA_D was

8.52 (mean score of 42,86 for the boys, and 34,33 for the girls). The boys seemed to outperform the girls when it comes to sound recognition. The difference between boys and girls in the mean score on LLAMA_E was 14.76, (the girls scored 93.33, and the boys had a mean score of 78.57). The difference in the mean scores on LLAMA_F was also quite high: 18.67, with the boys achieved a mean score of 60, and the girls having a score of 41.33. Finally, the difference between boys and girls in the mean score of all the four subtests summarized together is 1.71 in favor of the boys. The girls had a total mean score of 55.08, whereas the boys performed slightly better with a total mean score of 56.79.

4.1.2.2 Statistical Correlations

Pearson's correlation coefficient (R) was used in order to analyze if the aptitude results from the four subtests were correlated in some sense or if the tests seemed to tap into different areas of language learning. In order to analyze the tests, we used the R statistical software. R is the name of "a statistical analysis and graphics environment and also a programming language." (Stowell, 2014). In order to use the Pearson's correlation test to analyze the data, normally distributed data is needed. Because of this, we had to first assess whether the LLAMA tests were normally distributed or not.

To assess normal distribution, the Shapiro-Wilk's test was used. This is "a hypothesis test that can help to determine whether a sample has been drawn from a normal distribution. The null hypothesis for the test is that the sample is drawn from a normal distribution and the alternative hypothesis is that it is not" (Stowell, 2014). LLAMA_B, LLAMA_D and LLAMA_F passed the Shapiro-Wilk normality test, which means that the Pearson correlation test for normally distributed data could be used to analyze the tests. This is a test that uses measures of association to summarize the relationship between two variables, including the covariance (Stowell, 2014).

The p-value from the Pearson correlation test for LLAMA_B and LLAMA_D was 0.31 ($r = -0.23$), the p-value for LLAMA_B and LLAMA_F was 0.16 ($r = 0.31$), and the p-value for the correlation between LLAMA_D and LLAMA_F was 0.94 ($r = -0.02$). All of these p-values are above 0.05, which means that we cannot reject the null hypothesis that the data are not correlated. As for LLAMA_E, this subtest did not pass the Shapiro-Wilk test, indicating that the results from this subtest are not normally distributed. Thus, another correlation test had to

be used to assess this subtest. Therefore, the Kendall rank correlation coefficient test was applied to LLAMA_E. The results showed that correlations between LLAMA_B and LLAMA_E gave a p-value of 0.14 (r = 0.25), LLAMA_D and LLAMA_E gave a p-value of 0.17 (r = -0.24), and LLAMA_F and LLAMA_E gave a p-value of 0.54 (r = 0.11), which makes it impossible to reject the null hypothesis for any of these subtests. In total, if we look at the p-values for all these subtests there seems not to be any significant correlations between the subtests. The results from the statistical correlation analysis between the individual LLAMA subtests are shown in table 1 below.

LLAMA Subtest	Correlating Subtest	Estimate	p-value	Significance
LLAMA_B	LLAMA_D	-0.23	0.31	≈ not significant
LLAMA_B	LLAMA_F	0.31	0.16	≈ not significant
LLAMA_B	LLAMA_E	0.25	0.14	≈ not significant
LLAMA_D	LLAMA_E	-0.24	0.17	≈ not significant
LLAMA_D	LLAMA_F	-0.02	0.94	≈ not significant
LLAMA_F	LLAMA_E	0.11	0.54	≈ not significant

Table 2 - LLAMA subtest correlation

In addition to investigating statistical correlations between subtests of the LLAMA, an analysis to examine whether there was any correlation between the results from the LLAMA and the other mapping tool that was being used at this school, called *Kartleggeren*, was conducted. This is a mapping tool that measures the pupil's ability in reading and writing, as well as how large the vocabulary of the pupil is. Again, as with the term grade correlations,

LLAMA_E stood out as the only subtest that could show for a correlation to Kartleggeren. With a p-value of 0.09 (r = 0.29), LLAMA_E is evaluated as marginally significant in terms of correlations, meaning that a pupil who scores high on this subtest would probably also gain a high score on Kartleggeren and vice versa for low scorers. The rest of the subtests all showed higher p-values and could not be evaluated as correlated. LLAMA_B gave a p-value of 0.83 (r = 0.04), LLAMA_D gave a p-value of 0.61 (r = -0.12), LLAMA_F gave a p-value of 0.99 (r = -0.0004) and the total mean score of all four LLAMA subtests gave a p-value of 0.57 (r = 13). This would suggest that a high performance on these subtests would not necessarily give a high score on Kartleggeren. In total, the p-values were also higher for the overall correlation between the LLAMA and Kartleggeren than between the LLAMA and their expected term grades. The results from the statistical correlation analysis between the LLAMA subtests and *Kartleggeren* is shown in table 2 below. The marginally significant correlation is bolded.

LLAMA Subtest	Correlating Factor	Estimate	p-value	Significance
LLAMA_B	Kartleggeren	0.04	0.83	≈ not significant
LLAMA_D	Kartleggeren	-0.12	0.61	≈ not significant
LLAMA_E	**Kartleggeren**	**0.29**	**0.09**	**≈marginally significant**
LLAMA_F	Kartleggeren	-0.0004	0.99	≈ not significant
Mean LLAMA	Kartleggeren	0.13	0.57	≈ not significant

Table 3 - Results of the LLAMA tests according to Kartleggeren

The last statistical analysis that was conducted, was an investigation of the correlation between the LLAMA results and the expected term grades for the schoolyear. The only subtest that proved a significant correlation with the expected term grades was LLAMA_E with a p-

value of 0.01 (r = 0.44). This suggests that pupils who gained a high score on LLAMA_E, also received a high English grade from the teacher by the end of the schoolyear and the low LLAMA-E scorers received lower term grades in English. As for the rest of the subtests, none of the others had any evidence for a significant correlation due to higher p-values. LLAMA_B gave a p-value of 0.35 (r = 0.21), LLAMA_D gave a p-value of 0.64 (r = -0.11), LLAMA_F gave a p-value of 0.34 (r = 0.22), and lastly the total mean of all four LLAMA subtests gave a p-value of 0.13 (r = 0.40). This suggests that higher performances in these subtests could not be a strong indication of a high term grade in English for the pupils of the upper secondary class. All of the results from the correlation tests between the LLAMA, 'Kartleggeren' and the expected term grades can be found in table 3 below. Marginally significant correlation results and significant correlation results are bolded.

LLAMA Subtest	Correlating Factor	Estimate	p-value	Significance
LLAMA_B	Expected Grade	0.21	0.35	≈ not significant
LLAMA_D	Expected Grade	-0.11	0.64	≈ not significant
LLAMA_E	**Expected Grade**	**0.44**	**0.01**	**≈ significant**
LLAMA_F	Expected Grade	0.22	0.34	≈ not significant
Mean LLAMA	**Expected Grade**	**0.40**	**0.07**	**≈ marginally significant**

Table 4 – Results of the LLAMA test according to expected grades

4.1.3 Questionnaire

A questionnaire consisting of 7 questions using a 5-point Likert scale was used mainly in order to investigate RQ1 [14]and RQ2[15]. The questionnaire was administered to the pupils, immediately after the aptitude testing. The scale ranged from 1-5, where a score of 1 typically described the least favorable way of assessing a situation, whereas a score of 5 described the most favorable assessment of a statement. Table 2 summarizes the answers to the 7 questions. The first question is omitted from table 2 because it relates to the gender of the participant. Questions 2-6 concerned participant's assessment of their experience on a scale of 1-5 and question 7 was related to a ranking of the LLAMA subtests. Question 7 asked which of the four subtests the participant liked the most and the least. The ranking for every subtest is a summarized score where a rank of 1 gives 4 points and a rank of 4 gives 1 point, with the total amount of points for every subtest reflecting its ranking among the four subtests.

Question	Mean	SD	=n
2. How well do you like English as a school subject from 1-5? (1 = not at all, 5 = like it a lot)	3.68 ≈ like it	.924	22
3. How was your overall experience with using the LLAMA? (1 = very bad, 5 = very good)	3.36 ≈ neutral	.710	22
4. Were the tasks easy to understand? (1 = very difficult to understand, 5 = very easy to understand)	3.40 ≈ neutral	.778	22
5. Do you think this test, and knowing your language aptitude, is useful for your education?	3.55 ≈ useful	.891	22

[14] Is the LLAMA a suitable aptitude test battery for an upper secondary school class?

[15] How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?

(1 = not at all, 5 = very useful)			
6. Did the results you got correspond with your previous assumption of your aptitude? (1 = not at all, 5 = absolutely, yes)	3.09 ≈ neutral	.949	22
7. Look at the four images of the subtests of the LLAMA below. Rank the subtests from 1 – 4 on which subtest you liked the most. 1 is the subtest you liked the most and 4 is the subtest you liked the least. Write the numbers in the green boxes next to the images of the subtests.	LLAMA_B = 60 ≈ liked the most LLAMA_D = 37 ≈ liked the least LLAMA_E = 52 ≈ Neutral LLAMA_F = 57 ≈ Neutral	-	22

Table 5 – Questions and mean results from the questionnaire

As evident from table 5, question 2 asked how well the participants liked English as a subject in school. This means that the pupils generally favored English. The mean score for this question was 3.68 (SD = 0.92) indicating that the pupils generally liked English as a subject. In light of this, it is possible to argue that negative attitudes towards English as a subject can be ruled out as a potential factor for why the participants may have low aptitude scores or give certain scores on the questionnaire. The overall experience the pupils had with the LLAMA showed a mean score of 3.36 (SD = 0.71). This score suggest that the pupils generally rated their experience as ‘neutral’ (neither good, nor bad). The tasks difficulty showed a mean score of 3.40 (SD = 0.78), arguably, indicating that the level of difficulty was suitable: the tasks were neither too easy nor too difficult to understand.

The question regarding the importance of knowing once own aptitude showed a mean score of 3.55 (SD = 0.89) indicating that the pupils generally believed that this information

could be useful for them. The aptitude level they received on the LLAMA tests matched some of the pupils' perception of their own aptitude, while for some of the pupils, the score obtained on the LLAMA tasks did not match their own beliefs regarding their own aptitude. The mean score for question 6 was 3.09 (SD = 0.95) making this a neutral result. None of the 22 participants answered "absolutely yes" on how the results on the LLAMA corresponded with their own perception of their aptitude level. 9 out of 22 ranked question 6 as "yes," meaning that 9 participants felt that the aptitude level corresponded with their previous beliefs. The final question regarding the ranking of the aptitude tests showed that the LLAMA_B was the most popular subtest with a total score of 60 points. The least favorable subtest was the LLAMA_D with only 37 points.

When comparing the LLAMA test scores and the questionnaire results, some interesting findings occurred. The results from the questionnaire and the LLAMA results were compared in order to find out if the participants valued (and subsequently ranked) the subtests mostly based on how well they performed on each subtest. Only 10 out of 22 participants ranked the subtest they received the highest score in as their favorite. Moreover, only 7 out of 22 pupils included the two tests they received the highest score in as their two highest ranked subtests. Interestingly enough, 2 out of 22 participants did not rank any of the two subtests they gained the highest score in among their top two ranked subtests. These results suggest that the ranking of the subtests from the participants perspective was mainly based on the experience they had with the tests and not on the score they got. To support this observation, the results also showed that none of the four lowest scoring participants, that gained an overall aptitude level of 2 out of 4 on the LLAMA tests, assessed their overall experience with the LLAMA (questionnaire Q3) as worse than a score of 3/5 (≈ neutral) on the Likert scale in the questionnaire. Furthermore, the only three participants who assessed their overall experience with the LLAMA (questionnaire Q3) as 2/5 (≈ bad) on the Likert scale all gained an overall aptitude level of 3 out of 4 on the LLAMA tests. Taken together, these results point towards the conclusion that the experience the participants had with the LLAMA did not always reflect the actual results they got during the testing.

4.2 Qualitative Results

This second section of the results will present the qualitative results from the intervention. Duff (2012) describes qualitative research in SLA as "positivist, testing hypotheses or looking for cause–effect relationships and seeking an objective reality or "truth" about the nature of SLA under scrutiny; they may be interpretive, trying to understand the experiences, abilities, and performance of learners, or their perceptions of those experiences and reconstructions of them through narratives or interviews, for example; or they may be critical, examining learners in terms of larger social issues related to power, oppression, and discrimination." (p. 98). This is generally a type of data that can be observed and recorded. The quantitative dimension of this research project consists of two semi-structured teacher interviews. These interviews are analyzed by transcribing the audio recording from the interviews and then later drawing out the relevant information in order to investigate the connected research questions. The information is divided into thematically section based on the type of information that was retrieved from the interviews.

4.2.1 Teacher Interviews

Two separate interviews were conducted with the teacher during the course of this research project in order to catch the immediate reactions from the teacher after the testing session and a later interview to see how the LLAMA test results had been used. The teacher was asked to answer a total of 18 questions regarding her perception, attitude and opinion on the LLAMA aptitude testing. Some of the answers are presented in direct quotations and some are paraphrased. Also, some of the information from the interview has been omitted as it has not been deemed valuable to the study in my opinion. The teacher has later read through this section to ensure that the paraphrasing has not caused wrong assumptions or uttered incorrect statements from herself. She was also allowed to add comments or answers to the questions later. This section is structured by presenting the information that emerged from the interviews into 6 main themes that were relevant to answer the research questions. These being *teacher's attitude towards aptitude testing and the LLAMA, challenges with aptitude testing and the LLAMA, ethical considerations, other mapping tools, pedagogical implications* and lastly *pedagogical choices.* Each of these are connected to separate RQs from this project.

4.2.1.1 Teacher's Attitude Towards Aptitude Testing and the LLAMA

The results regarding the teacher's attitude towards aptitude testing is mainly related to RQ5[16]. The response from the teacher was mainly that she was surprised and intrigued by this way of mapping pupils and showed interest into the further development for the attitude battery. She stated that: "I liked it more than I would have thought." She elaborated by saying that: "I liked that it was easy to use and understand and that it was not that time consuming" and that "[i]t was surprising how much new information about the pupils you could find by using this test." The teacher added that she did not see any issues with conducting the test in class and that it seemed to go quite unproblematic for the pupils. The teacher also said: "It was a good PowerPoint and I think they needed that," referring to the PowerPoint (see Appendix C) I made to guide the pupils through the test. The pupils were also described to be understanding the concepts of aptitude and the reasons for aptitude testing, as this is also presented in the introducing PowerPoint. She concluded by stating that she did not see any major issues with testing procedures such as the LLAMA, as the pupils are generally familiar with these types of tests and also the fact that there are no grades involved in the testing. She also added that "I think doing the test in Norwegian is smart" and explained the value she saw in the test being language neutral.

She continued by explaining that she found it useful to know how the different pupils learn. A concept that was also stressed by the teacher is the notion that aptitude testing can show the pupils that language is a universal concept and that the same underlying abilities enables learners to learn new languages regardless of what language is being learned. "I think it quickly gives you a picture of their language aptitude." She continued by elaborating that testing the pupils' aptitude quickly gives a valuable insight into the learning abilities of the pupils. Furthermore, she argued that using the aptitude results actively could be beneficial in some teaching situations where tasks could be adapted to the different aptitude levels of the pupils in the four components of aptitude the LLAMA is testing. Her final note about the test and its pedagogical advantages was: "I was quite intrigued by this test and it gave me a few moments of realisation."

---

[16] What attitude does the teacher have towards aptitude testing and the LLAMA?

4.2.1.2 Challenges with Aptitude Testing and the LLAMA

The results regarding the challenges the teacher found in using the LLAMA and using aptitude testing as a mapping tool in general is connected to RQ1[17] and RQ2[18]. The only negative point she made towards the carrying out of the tests were the flaws in the test regarding the wireless hearing devices that did not function with the test and the errors that occurred in the LLAMA. She also explained that there could be difficulties in understanding how to use the results properly. On the other hand, she also pointed out that "the difficulty is that we have full classes." She underlined that this makes adapting the teaching based on aptitude level more difficult, but that this is something teachers should still strive for. She explained that this difficulty mainly derives from a limited amount of time in class, accompanied by a large curriculum.

4.2.1.3 Ethical Considerations

The ethical considerations found in the teacher interviews are connected mainly to RQ1. Firstly, teacher said that: "There are always problems with regards to grouping pupils." She explained that this kind of grouping can sometimes make pupils feel stigmatized and hurt by low aptitude results, but that this is something that can be avoided by good facilitation and smooth groupings from the teacher. She continued by underlining that "I think it is important that the whole school uses this type of test procedures and not just one single teacher." Another concept to take into account that was pointed out was the pupils' personal considerations of such testing. "It is important to take the privacy of the pupils into consideration so that only the teacher gain access to the results." Knowing too much of their aptitude could also create a problem, as pupils may be put into boxes, based on their aptitude level.

---

[17] Is the LLAMA a suitable aptitude test battery for an upper secondary school class?

[18] How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?

4.2.1.4 Other Screening Tools

The information about other screening tools acquired in the interviews is mainly retrieved to help discuss RQ4[19]. On question about other screening tools, the teacher replied that "we only use *Kartleggeren*." She furthermore explained that "this test gives you an idea of the pupils' abilities in the areas of reading, writing and understanding in English." This is also pointed out as a point of improvement from the teacher, as she admits that the results are not used as efficiently as they should be, due to limitations in time. She also explained that the main reason for using this mapping tool is to locate the weakest pupils and offer them additional support. She also discussed the LLAMA in contrast to *Kartleggeren* where she stated: "Yes I do, I think it [the LLAMA] could give us a wider picture of each student and help us understand what areas of language learning they are struggling the most" and "[w]e do not get the same detailed picture of the pupils when we use mapping tools as Kartleggeren." She viewed Kartleggeren as a more general mapping tool, whereas she felt the LLAMA was more of a detailed tool for mapping pupils. On a follow up question about the possible uses of the results from Kartleggeren and the LLAMA, she agreed that the results from Kartleggeren are more aimed at finding the pupils that are struggling with the subject, whereas the LLAMA could potentially function as a mapping tool to give more answers about the pupils' language learning abilities. She also noted that: "Since I have not used the LLAMA that much, I am not quite sure, but I think the LLAMA could provide a broader picture of the language learning abilities of the pupils."

4.2.1.5 Pedagogical Implications

Pedagogical implications for using aptitude testing in teaching is a major topic in this project and the results found in the interview that relates to these implications are mainly included to answer RQ4. The first point the teacher discussed was the advantage the LLAMA could provide towards adapting the teaching to foreign pupils and immigrants. She said that "we have a larger variation of pupils in the classes now than before, with for example immigrants, and students with English as their first language." She also pointed out that immigrants from other countries are often the lowest scoring pupils in English, due to a lack of English education from their

---

[19] What pedagogical advantages can aptitude testing have in an upper secondary class in English?

mother country. The teacher explained that the LLAMA could function as a tool to create a better understanding of what areas of language learning these foreign students master and what they do not master. Since some foreign students are strong language learners and others are not, it is important to gain an understanding of why they are struggling with learning English. If it is because they are struggling with language in general or if it is because of the lack of English education from their mother country.

She continued by saying that the LLAMA aptitude test could provide a stronger starting point for adapting the English education to the abilities of these individuals in their language learning process. She explained that when knowing more specifically what the problem with the language learning of the pupil lies in, the school as a whole can, create better systems to adapt the education of the pupils towards more quality and more precise feedback on the language learning. She reflected that: "I could become better at adapting my teaching to each student. For instance, if I know that 15 of my students are struggling with the same area of language learning, I could focus more on the type of tasks that help learners more based on their aptitude profile."

She also noted that "I think it [the LLAMA] could be used across several language classes (both English and other foreign languages in school) and also to inform both students and language teachers about the underlying structures of language learning." Regarding how to facilitate teaching for high- and low scorers, the teacher answered: "I am not sure." She explained that given how the Norwegian School System I built up, there is not much time to facilitate for the best possible learning for the high scoring pupils. Still, she added: "I think it could help us understand how we should give the pupils that obtain a high score new tasks and challenges adapted to their language level." She also said that the results can be used to give more effective feedback, as the high scorers on the LLAMA seems to understand the feedback faster.

With regards to the pedagogical potential of the LLAMA, the teacher immediately answered: "Yes, I think the potential [with using the LLAMA] is good because you get a broader picture of each pupil." She was surprised about some of the results from the test and underlined that her feedback to pupils on assignments improved. She also agreed that by using the LLAMA, it gets easier to differentiate between if the results of a pupil are mainly caused by individual differences such as effort, motivation and attitudes towards language learning, or if the results come from the pupil's language learning abilities.

4.2.1.6 Pedagogical Choices

In order not only to discuss what potential the LLAMA and aptitude testing has, questiona about the actual pedagogical changes the teacher made through the schoolyear was added. These results are mainly related to RQ6. The teacher begun by stating: “Yes, I think so. It gave me a broader and more specific understanding of the pupils’ abilities in English” She also noted that the LLAMA gave more thorough information and insight into the abilities and underlying language learning skills of the pupils, than of what a other available assessment situations would provide.

On questions about the effect of the LLAMA, the teacher answered: “Yes, I have been affected in some ways. I have become more aware of the different components of language learning.” She explained that the LLAMA gave insights into why pupils were struggling with some areas of the language learning process. “I have been paying more attention to working with vocabulary and pointing pupils in the correct direction of how to work with this. I have also adapted new ways of working with grammar tasks, especially with use of digital tools.” She used many of these language learning techniques because of the test results and the LLAMA Functionality Frame (see Appendix F), provided by me to help the teacher use the results in a practical way.

During a discussion about whether aptitude is a stable or dynamic trait, the teacher answered that “I want to say yes, I think aptitude could be seen as a dynamic trait.” She underlined that this is a difficult question to answer, but that she noticed some development in their aptitude, based on the speed and the amount of effort used for the pupils to acquire new knowledge of English. “I think they have developed in the way they answer tasks and more of them understand new and unfamiliar tasks.”

5. Discussion

This chapter will discuss the results of this master's thesis. The discussion section will be presented one research question at the time for structural matters. A section that reflects on the ethical considerations of aptitude testing and a section that suggests limitations and further research will also be provided. The discussion in this paper will aim to examine and discuss the six research questions examining the results of this intervention in light of previous research. The research will mainly coincide with the research presented in section 2 of this paper (Theoretical Background). The results that will be discussed are the results found in the intervention of this project, presented in section 4 of this paper (Results).

5.1 Suitability for Upper Secondary School

RQ1 [20]of this paper was asked to investigate whether it would be suitable to carry out an aptitude testing like the LLAMA in an upper secondary school classroom. In order to find pedagogical advantages with using the LLAMA and explore different attitudes towards aptitude testing, I will have to rely on a suitable and usable aptitude test battery, and thus the LLAMA must be tested for its suitability in the current educational setting. For this research question, several aspects of the previous research done on this field of study has been important. Many of the actions in the intervention has also been relevant to answer this research question.

To begin with, questions have been raised by many researchers regarding age and the use of the LLAMA test battery. The study by Rogers et al. (2017) concluded that "[t]he current LLAMA tests are not suitable for use with younger learners." In this study, 'older learners' were two groups of participants ages 20-21 and 30-70, whereas 'younger learners' were one group of participants aged 10-11 (see section 2.4). This age range is interesting since the current study investigates participants ranging from age 15-16 who will then fall into the middle of the two youngest age groups. The results from the intervention show that the participants from the current group scored much closer to the older groups of learners than the younger ones. In the study of Rogers et al. (2017) the youngest learners (aged 10-11) obtained a mean score of 28,67 (SD = 14.920) for the LLAMA_B and a mean score of 18,50 (SD = 13.528) for LLAMA_D. The older learners (aged 20-21) obtained a mean score of 45,68 (SD = 21.592) for LLAMA_B

[20] Is the LLAMA a suitable aptitude test battery for an upper secondary school class?

and a mean score of 29,32 for LLAMA_D. To compare, the participants of the current study obtained a mean score of 49,55 (SD = 17.183) for LLAMA_B and 37,05 (SD = 14.034) for LLAMA_D (all of the results are shown in table 4 below). This means that the participants of the current study scored better than the older learners from Rogers et al. (2017), making them far closer in scores to the older, than the younger group of learners. This supports the notion that the age group of 15-16, placed in the first year of an upper secondary school, is a suitable age group and setting for using the LLAMA tests.

Age Group	M / s.d.	LLAMA_B	LLAMA_D
Group 1: 10-11 (n = 30)	M	28.67	18.50
	s.d.	14.920	13.528
Group 2: 20-21 (n = 44)	M	45.68	29.32
	s.d.	21.592	17.206
Group 3: 30-70 (n = 30)	M	44.33	24.50
	s.d.	24.380	17.536
Group from current study: 15-16 (n = 22)	M	49.55	37.05
	s.d.	17.183	14.034

Table 6 - LLAMA Mean results from Rogers et al. (2017) and the current study compared to age

In order to assess the suitability of the LAMA test battery for upper secondary L2 English learners, one important aspect of the test battery that had to be investigated was whether the tasks were comprehensible for the pupils. Robinson (2001, p. 377) stated that "[w]hen the cognitive demands of tasks are increased along resource-dispersing factors (e.g., planning time), learners' memory and attentional resources will be dispersed, which will affect language production and uptake of focus on form in negative ways." This means that if the LLAMA tasks were too difficult, the aptitude scores the participants gained would not be representable for their actual aptitude level. The results from the questionnaire suggest that the participants found the task difficulty quite appropriate. With a neutral result (3,40/5,00) on a question about task

difficulty, the tasks in the aptitude battery seems not to be too difficult, nor too easy. In addition to this, the teacher also stated that the pupils seemed to understand the test with some facilitation provided by myself, indicating that the test was comprehensible and understandable for the pupils. If the PowerPoint presentation and visualization of the test were removed from the facilitation and preparation process, these learners might have had a slightly more difficult time in mastering and understanding the tasks, especially with emphasis on LLAMA_F, which requires some explanation. This thorough facilitation might prove even more important if the test should be administered to younger learners than the participants of this study.

Another factor that arguable played a role in the success of the administration of the test and a wider understanding of the term aptitude from the pupils, was the fact that the LLAMA test is language neutral. The studies from Rogers et al. (2017) and Granena (2013) showed that the LLAMA could be administered to different L1 groups without affecting the results (see section 2.2). This is important for the use in the Norwegian classroom. The fact that the test can be used on pupils with different L1s gives the school a strong mapping[21] tool for pupils who have other L1s than Norwegian. This was also pointed out by the teacher of the class as one of the stronger functional sides with the LLAMA: teachers, can map foreign pupils and immigrants to find their language learning potential without the results being inflicted by the L1 of these pupils by using the LLAMA. On many occasions, immigrants and foreign English pupils are dismissed as merely 'bad English learners' because of a lack of performance on tests such as *Kartleggeren* which only provides information about current skills in English and not the potential for learning a language, as the LLLAMA does. This lack of performance on regular mapping tests could often derive from factors such as previous English education or lack of exposure to English. By using the LLLAMA as a mapping tool the potential of these pupils can be revealed and enhanced.

The functionality of the test is well supported by the fact that the L1 of the participants does not inflict the results. In addition to this, the test can also be administered to pupils without having to use English as language of instruction. If you take the MLAT as an example, this is an aptitude test that is required to be administered in English for the pupils to be able to understand the tasks (Carroll & Sapon, 1959). The same goes for the PLAB, which is also an

[21] *Map/mapping* is chosen as the English translation for *kartlegge/kartlegging,* as Udir claims this is the most appropriate translation. See https://www.udir.no/verktoy/ordbok/ for more information.

aptitude test founded on English as the language of instruction (Pimsleur, 1966). These two tests would have to be modified and restructured in order to be administered to other participants than L1 English speakers. In contrast, the language neutrality of the LLAMA ensures that administration and tasks does not affected the results because of miscommunication due to a lack of English proficiency or any sorts of anxiety for asking questions in English. The teacher of the class in the present study was also a teacher of French and viewed the language neutrality in the test administration process as a huge advantage in terms of adapting the test to other language classes than English. This makes the test applicable for use in all of the foreign language classes in the Norwegian educational system (mainly French, Spanish and German).

In terms of functionality, validity and reliability is also an important factor to ensure that the test does measure different components of aptitude. It is clear from both the creator and other researchers that the LLAMA is not yet properly validated and standardized and that the test should not be administered in high stake situations (Granena, 2013; Meara, 2005). A statistical analysis of the aptitude tests was done in this intervention to see if the tests correlated or not (see section 4.1.2.2). This was done to see if the subtests of the LLAMA tapped into different areas of language learning or not. The results showed that there was no statistical evidence for a correlation, providing evidence that the subtests of the LLAMA do in fact measure different areas of the language learning process. This result provides evidence that support results from studies conducted by Granena (2013) and Rogers et al. (2016, 2017) and, where both studies suggested an acceptable validity for the subtests. The base of empirical data is still small from this intervention and taking the criticism of Bokander and Bylund (2019) and Bokander (2019) (see section 2.2) into consideration one cannot completely trust the LLAMA until further standardization and validation is done.

5.2 Pupils' Perception of LLAMA Functionality

For the aptitude testing to be useful and meaningful in a classroom and to the teacher, an important key to success is also that the procedure of aptitude testing and the test itself has a general approval with the pupils. If the pupils have a negative experience with the LLAMA test it would probably mean that teachers would in general be more reluctant to using it in the future.

This is why RQ2 [22]is important to discuss. The discussion is mainly based on the results from the questionnaire the pupils were given after the LLAMA testing, the first teacher interview and general observations from the teacher and me.

The questionnaire (see section 4.1.3) showed a neutral attitude towards how well the pupils liked the LLAMA test. This result can suggest that the experience the pupils had by using the test was not negative, and therefore this is not an attitude that should hinder the use of the LLAMA for the teacher in any sense. I did not expect the pupils to like the test in any significant fashion, in the sense that it challenges their mental capacity to a high extent and provides many of them with lower aptitude results than they might have expected. Given that question 6 in the questionnaire showed that the correspondence between the pupils' assumption and their actual test scores was neutral, many of the pupils might have felt a more negative attitude towards the test because their aptitude was not as high as they had anticipated. Some of the errors that we experienced with the LLAMA when testing could also have affected the attitudes of the pupils towards the LLAMA. It should still be noted that my personal observations suggests that most of the pupils found the test as a tolerable English activity. Some of the pupils even uttered that they liked the testing session more than they liked their regular English sessions. Another indication of the approval of the LLAMA from the pupils was that question 2 regarding how well they liked English as a subject (3,68) only had a mean score that was 0.32 higher than how well they liked the LLAMA test (3,36). This indicates that the test should not be an unpleasant change from the regular English teaching, making the LLAMA testing something the teacher can conduct without any concerns with regards to it being a disruptive or negative English experience in this context, with this specific learner group.

The teacher also had opinions regarding how well the pupils seemed to perceive the functionality of the LLAMA aptitude test. She was positive towards how conducting the test in class went and could not see any issues except from the technical errors we experienced. These positive observations from the teacher also suggest that using the LLAMA in class was an unproblematic experience for the pupils. In addition to this, the pupils also found the results useful for their own English education. This is something I would not have expected to see in the results from the questionnaire. Still, this indicated that the pupils also saw the potential

[22] How do pupils in an upper secondary class perceive the functionality and use of the LLAMA aptitude test battery?

value of knowing their own language aptitude level. In total, there seems to be positive results regarding the experience the pupils had with the LLAMA. This makes it easier for teachers to use such mapping tools actively in the English teaching and as a part of the individually adapted teaching.

5.3 Teacher Predictions and Pupil Aptitude Correspondence

To understand the relationship between the perception the teacher has about the concept of aptitude and the aptitude levels of the pupils, an aptitude prediction of all pupils was conducted by the teacher. This was done to answer research question RQ3[23]. The aim of this research question is to investigate the difference between the teacher's perception of the pupils' aptitude levels and the actual scores they gain on the LLAMA aptitude battery. This can be interesting for many reasons. In terms of how a language aptitude test can be validated and also to investigate whether there is a need for an aptitude test, it might be of relevance to look at how the teacher perceives the aptitude of the pupils, without using an aptitude test battery to gain results. The predictions of the teacher can also reveal other interesting deductions for discussion.

The teacher tried to predict the aptitude of the pupils, for the results to be compared afterwards with the actual LLAMA scores the pupils obtained (see section 3.4). The resuls from this prediction showed a mean deviation of 0,55, which describes the difference in predictions and actual obtained LLAMA scores (see section 4.1.1). This suggests that there is a need for an aptitude test battery in terms of knowing the correct aptitude level of the pupils. Since the teacher was unsuccessful in predicting over half of the pupils, these findings indicate that using the LLAMA would inform the teacher in a useful sense. These results can help the teacher to better understand the aptitude profile of the pupils and thus enhance the and further guide the teacher in offering a more individually adapted teaching to each pupil. If the mean deviation had been too low, it would have indicated that there was no need for an aptitude mapping of the pupils, as the teacher was intuitively able to predict this.

Notably, from reading the results from the aptitude prediction, the teacher had a tendency to assess girls quite different from boys. The teacher underestimated 47% of the girls,

---

[23] Does the LLAMA aptitude test correspond with the teacher's perception of the pupils' aptitude?

whereas none of the boys were underestimated (see section 4.1.1). The reasons for this noticeable difference in predictions, based on gender is quite difficult to determine. The teacher herself was surprised to see this and could not clearly account for why this difference in prediction occurred. Research from Habl (2018) suggest that there is little difference between genders on LLAMA_B and LLAMA_F, whereas both Poschner (2018) and Hörder (2018) found that females outperformed males significantly of the LLAMA_B. In the current study females did obtain a higher score for LLAMA_B and LLAMA_D (see section 4.1.2), but were still underestimated by the teacher. This underestimation of girls contrasted what I expected from this aptitude prediction task. I believed the teacher would overestimate the girls, in light of the commonly held folk wisdom that girls are better language learners than boys (Richter, 2018). It is interesting that the teacher chose to move away from this common perception of gender and language learning. Her explanation for her choices was that she viewed many of the subtests as more 'mechanical' meaning that the tasks were to be solved with simple deductions and straight forward answers. She contrasted this to the girls, who she viewed as more capable of tasks which involve more thorough problem solving.

5.4 Pedagogical Advantages

In order to assess the impact a test like the LLAMA can have on English teaching. To discuss the functionality and possible use of the LLAMA and aptitude testing in general in the Norwegian school system, the aspect of how the test results can be of pedagogical use in the classroom. The goal of asking RQ4[24] is to discuss how the LLAMA can be used actively in teaching English to optimize the education every pupil receives, based on the individual differences and language learning abilities they possess. The results connected to this research question mainly relies on the second teacher interview that was conducted a full semester after the actual testing to track progress and teacher perceptions as well as pedagogical choices.

First, the LLAMA showed great pedagogical potential in the form of mapping pupils in more detail than other mapping tools currently used in school. The school of the intervention currently used the mapping tool *Kartleggeren* in the English classes for screening pupils in the beginning of the schoolyear. Still, the teacher stated that the problem with this mapping tool

[24] What pedagogical advantages can aptitude testing have in an upper secondary class in English?

has been that it does not provide the information needed about the different pupils. This has been a problem for mapping tools used in Norway for many years and many of the mapping tools have only given information to identify pupils who struggle the most in English. This is because the tasks in the mapping tests have been built up in a way that guarantee correct answers for most of the pupils, making the test identify those who need extra support in English (Brevik & Helness, 2018). However, current research on mapping Norwegian pupils, in line with the new subject curriculum, has begun to move more in the direction of mapping for formative assessment. This refers to all assessments that with the aim of improving students' learning processes and/or the teacher's teaching procedures (Burner, 2018). To map pupils on the base of formative assessment during a schoolyear, new mapping tests are designed to measure all levels of English competence and not just the lowest scorers (Utdanningsdirektoratet, 2021). The new tests make use of a similar scoring scale as the Common European Framework of Reference for Languages to ensure a more descriptive and detailed language profile. This is because the new mapping tools are made to support development and individually adapted teaching to each pupil, based on their language profile and abilities.

What is central to the discussion of this paper is that the way of understanding the concept of aptitude has developed during the last couple of years. As mentioned, Wen, Biedroń and Skehan (2017) argued that aptitude could now be viewed as a dynamic trait rather than the previous way of viewing aptitude as a stable and latent trait, that was unable to develop over time (see section 2.2). By using the LLAMA aptitude test pupils can be mapped, not only to predict language learning success in a language (as was the initial idea with the MLAT from Carroll and Sapon), but to enhance individual teaching by using the results to understand the aptitude profile for every pupil and how it might develop through time. The teacher supported this notion by stating that the LLAMA was much more detailed and that it should be used to train teachers into understanding the underlying structures of language learning. This is because an aptitude test like the LLAMA gives information about the pupils' performance by using the underlying concepts of language learning as a base, instead of just testing the general abilities of say reading and writing. Thus, the language potential is revealed and not only current abilities. This means that the teacher would be able to see the overall potential the pupil has got in abilities that create strong and rapid language learners, instead of just knowing to what degree the pupil masters the abilities of reading, writing and how large their vocabulary is, as *Kartleggeren* measures.

With the tools to understand what makes a good language learner, teachers would be better equipped to understand why a pupil has a problem with improving language skills. If a teacher has administered the LLAMA to the pupils, more detailed information of the underlying language learning abilities of the pupils will be available. This could help the teacher understand why a pupil scores high or low in the subject. This can help the teacher differentiate what other individual differences affects the language learning process of the pupil. E.g. if a pupil receives a low grade on a written task or an oral presentation, the teacher could look to the LLAMA results to better understand if these low performances are connected to language learning skills or other individual differences such as motivation. Then, the suitable measures can be taken to deal with the issue, rather than just asserting that the pupil is not a strong language learner as the cause of the weak performances. If the motivation were the issue with the pupil in question, there are other ways of dealing with this issue than if the issue lies in weaker language learning abilities.

The LLAMA serves well as a mapping tool to provide an insight into the pupils' abilities and language learning profiles in second language learning and can function as a tool to adapt the teaching individually to each pupil. The problem for a teacher at this point is that it is impossible to acquire any sort of framework for how to use the results from the LLAMA testing. This is why I created the *LLAMA Functionality Framework* (see Appendix F), so that the teacher could effectively match the results from the LLAMA testing to a framework that describes how the results can be used to better adapt the teaching individually to each pupil. The guide applies recent research to develop central functionality tips, both based on individual LLAMA subtests, but also general tips for how to adapt teaching to low- or high scorers of the test. The manual (Meara, 2005) does not provide this information and there currently exist no such official guide, which in my opinion would be much needed in order to gain more from the testing. This was also a point the teacher made, and she clearly stated the need for a clear and structured guide for practical uses of the LLAMA (see section 4.2.2). However, current research systematically focuses more on matching individual differences in abilities to the information processing demands of different L2 tasks, indicating that such frameworks will be further developed in the near future of aptitude research (Robinson, 2005).

As mentioned previously (see section 5.2) there are many advantages connected to the LLAMA in terms of using the potential of foreign pupils and immigrants. These pupils are often affected by mastering more than two languages, making them multilinguals (Burner & Carlsen, 2019). If the LLAMA can be used to see beyond the affect previous English education or a lack

of exposure to the language has on the English proficiency, there exist several opportunities to use this multilingualism to enhance the proficiency of the foreign pupil, as well as other bilingual pupils.

Research has indicated that multilingual language learners are better at seeing the connections in the structure of languages, more creative and use more appropriate language learning strategies. Overall, most studies have shown that bilingual language learners are better language learners than monolingual learners (ac cited in Burner & Carlsen, 2019). If this language learning potential can be discovered early in the language process (preferably much earlier than upper secondary school), the language potential of these pupils can be fostered and nurtured in a more language enhancing way. By viewing these pupils in a different way, teachers can adapt the education by using learning strategies that makes use of the pupils' L1 in the language learning process, instead of providing easier tasks, way below their language potential, because they have a lower proficiency in English (and often also Norwegian) at the time they arrive for enrollment in the Norwegian educational system (see Burner & Carlsen, 2019). In addition to this, the first language of these foreign pupils can also be actively used to enhance language learning for other pupils by creating what can be referred to as a *multilingual classroom*. This entails using the possible advantages that lie in the experiences and knowledge the foreign pupils possess about their first language. This concept, along with the new subject curriculum, has been implanted to change the perception of multilingualism in the classroom from being an obstacle to learning into becoming a tool for motivation, individually adapted teaching and new knowledge about other languages around the world (Brøyn, 2019).

5.5 Teacher Attitudes

The overall goal of this project is to investigate whether the LLAMA test is functional for an upper secondary school class, and to discuss potential pedagogical advantages with using the test. RQ5[25] is asked because if the LLAMA test should have any chance of being used actively in English teaching, the attitude the teacher has towards aptitude testing and the LLAMA is imperative. The discussion of this research question will mainly be based on the two teacher interviews that were conducted throughout the project.

[25] What attitude does the teacher have towards aptitude testing and the LLAMA?

General attitudes towards aptitude and aptitude testing have been somewhat negative both in the Norwegian context as well as the international SLA contexts over the past decades. Several researchers point towards the development of instructional language learning, which has embraced a more communicative approach, making aptitude less irrelevant to some researchers in contrast to the relevance it had for the previously dominating audiolingual language learning methods (Skehan, 2002). In addition to this, aptitude research and testing has been an area of little research in the Norwegian SLA context, compared to research on other individual differences such as motivation or learner attitudes. Many teachers have also viewed ideas of aptitude testing as something that puts pupils in predisposed categories and blocks before they receive their education. Thus, it was my concern that the teacher of this intervention and Norwegian teachers in general would share some of these more dismissive attitudes.

Apart from the above concerns, the teacher generally showed a great interest in the concept of aptitude and appeared to have what was to me a surprisingly positive attitude towards aptitude testing and the LLAMA. The teacher was in general positive towards aptitude and aptitude testing and viewed it as a large asset to the teaching process. She stated that the test was very interesting and that this was a concept she believed could be useful to her as a teacher. She was actually surprised by how much she liked the test and very eager to learn about how the results could have pedagogical implications for her teaching. This attitude reflects some of the more recent attitudes the idea of aptitude has had in the research field of SLA, where the interest has risen in the last couple of years. Peter Robinsons (2002) anthology is often seen as a turning point in aptitude interest with his re-conceptualization of the construct of FL-aptitude (as cited in Wen, Biedron & Skehan, 2016). Because of this, research on language aptitude has re-emerged as a field of interest, and is now on its way to become one of the major points of interest in the area of second language research. This gives aptitude and the idea and functionality of aptitude testing new interest and new and extensive research is currently being done on this particular field of study (see Aetieda & Munoz, 2016; Granena, 2014; Granena & Long, 2012; Kourtali & Révész, 2019; Yalçin, Çeçen & Erçetin, 2016).

The goal for researchers must now be to enhance the interest in aptitude research so that more teachers can share the same opinion of language aptitude and aptitude testing as the teacher from this project. By finding a renewed interest in this concept from school administrators and teachers around Norway, aptitude testing could become a tool for mapping pupils to improve the individually adapted teaching in Norwegian language classes. The teacher from the current project was not entirely sure if the reasons behind her interest and positive

attitudes towards aptitude testing lay in the fact that she was very interested in language and linguistics or if it was because she felt that it was so useful to her teaching. In any case, if this test becomes more standardized and validated and more familiar to teachers and school administrators across Norway with the proper training, many more might share the views of myself and the teacher from this project. That language aptitude is something that has the potential of becoming a useful tool for language teaching.

5.6 Pedagogical Choices Based on Aptitude

Much of the focus of this project lies in investigating the potential advantages by using the LLAMA aptitude tests to map upper secondary school pupils. In order to research these advantages RQ6 [26]was added to see if the teacher of the class made pedagogical changes to her teaching between the time the test was conducted and the end of the schoolyear, when the second teacher interview was conducted. This question is of course difficult to answer for the teacher as a complete framework for how to use the LLAMA results was not provided immediately. She has still used the 'LLAMA Functionality Framework' (see Appendix F) which was made by me and based on current SLA research, to inform the teacher of potential pedagogical uses from the aptitude results. The second teacher interview was used to collect information about this research question.

On questions about whether the teacher had changed her teaching based on the LLAMA results or not, she confirmed that some changes had been made. To begin with, she had in general become more aware of the language learning profile of the pupils and she was able to familiarize herself with the underlying structures of language learning. This improved her teaching and enabled her to provide better feedback and understand why some pupils struggled with certain tasks, while others did not. She was also able to adapt some of the teaching based on what she could expect certain pupils to master. This is an important step into gaining benefits from mapping by using the LLAMA. Research done by Sternberg, Grigorenko and Zhang (2008) on adaptions based on language abilities showed that matching pupils with teaching methods that fit the way they think facilitates for a stronger learning outcome (see section 2.5). This relates to the importance that teachers who aim to use the LLAMA as a mapping tool,

[26] Will the teacher make pedagogical changes to the English education based on the aptitude results, and if so, what type of changes?

understand the underlying components of aptitude. If they acquire this knowledge about language learning, they might be able to, as the teacher of this intervention has experienced, be more capable of understanding what affects the language learning process of the pupils, and then be able to adapt the teaching in suitable ways based on aptitude scores from aptitude batteries such as the LLAMA.

More specifically, the teacher also outlined some specific changes she had made because of the LLAMA results and the Functionality Frame I provided. The first change she wanted to present was the alteration of how she worked with vocabulary learning. The part of the Functionality Frame that focuses on vocabulary learning is the section that describes the potential pedagogical advantages connected to the results of LLAMA_B, which is a vocabulary learning test (Meara, 2005). This section is based on the research by Poschner (2018) and his findings regarding aptitude and vocabulary learning (see section 2.5). The teacher had used this research to focus more on vocabulary learning strategies with a special focus on the low scoring pupils that, according to Poschner (2018), had the most benefit by using these strategies. She also noted that she was now more able to point pupils in the direction they needed in order to acquire more vocabulary, whether it was high- or low aptitude scorers. This had helped the teacher and she felt that her teaching could now enhance the vocabulary acquisition of the pupils. By combining the LLAMA results and the Functionality Frame, teachers can now be able to gain more from teaching vocabulary more explicit than before. If teachers become more aware of what strategies are useful and what strategies are not results from teaching might be enhanced. The teacher can adapt these strategies to the language learning profiles of the pupils, and thus enhance the gains of traditional vocabulary enhancing activities, like reading.

Another pedagogical change the teacher had made was altering the way she worked with grammar. She changed the way she taught grammar by using more digital tools in order to better adapt the teaching to each pupil, based on their aptitude score. This would be especially important for the scores pupils gain on LLAMA_F, a grammar inferencing task (Meara, 2005), as this test is the aptitude test from the LLAMA test battery that measures success in grammar acquisition with most accuracy. Research done by Erlam (2003, 2005) showed that pupil with a high analytic ability (high LLAMA_F-scorers) benefitted from an inductive approach with a structured input method (see section 2.5). Also, research done by Hwu and Sun (2012) supported this view and concluded that the inductive approach was more suitable for the high-aptitude scores. Research by Hauptmann (1971) as early as the beginning of the 70s even provided evidence towards a beneficial inductive approach for high-aptitude learners. Because

of this the teacher can benefit by teaching grammar to these high-aptitude scorers by exposing them to examples of the target language and then asking the pupils to figure out the rules that governed the examples that were given (see Lee & Van Patten, 2003). The teacher was able to take these points into account and help improve the teaching of grammar. Still, she could possibly expand her grammar teaching even more by differentiating the teaching of these domains even more.

When it comes to differentiating grammar-teaching based on aptitude scores there is also a more suitable method of instruction for teaching pupils that are low aptitude scorers. This especially counts for pupils that are low scorers of tests related to grammatical sensitivity such as the LLAMA_F (Meara, 2005). The results from a study by Hwu, Pan and Sun (2013) (see section 2.5) showed that low aptitude scorers benefitted more from deductive teaching approaches to grammar. This means that pupils should be exposed to rules that govern the language first, and then be given the chance to use these rules in examples from the target language. These findings are important for teaching approaches based on the aptitude treatment interaction method (see section 2.5). The teacher was able to take this point into consideration when she taught grammar and pupils were challenged in different ways, based on their aptitude profile. The teacher also experimented by using digital tools to teach grammar, but she found that adapting the grammar teaching in this situation could often prove more challenging, even though she was positive towards the way digital grammar learning is structured.

The previous two paragraphs propose the idea that inductive learning strategies benefit high-aptitude scorers, whereas a deductive approach benefit low-aptitude scorers. This deduction can arguably be opposed by the idea that one of the main reasons for the gap in aptitude lies in the fact that the high scoring group has attained better inductive learning strategies, which benefit this type of aptitude abilities, than the low scoring group. Since the discussion revolves around language learning strategies, Poschner's (2018) study of how high- and low- aptitude scorers make use of language learning strategies becomes relevant. The study shows that high-aptitude scorers do not use language learning strategies more frequently than low-aptitude scorers (Poschner, 2018). This notion thus points toward the idea that it is not the knowledge and use of language learning strategies that govern which teaching should be administered to each aptitude group, but rather the aptitude level of the pupils. On the other hand, evidence still suggest, that the low-aptitude scorers benefit more from using these inductive strategies than high-scorers would do. This can suggest that even though low-aptitude scorers do not benefit from inductive language learning strategies, compared to high scorers,

they might develop more in terms of the abilities connected to grammatical sensitivity by making use of inductive leaning methods. The decision the teacher must make, is whether the issues related to comprehension of the inductive tasks and methods for low-aptitude pupils outweigh the possible learning outcome these pupils might have in the abilities connected to grammatical sensitivity by using this teaching method.

5.8 Ethical Considerations

Measuring aptitude and the idea of mapping pupils can seem like a very direct and problematic way of assessing pupils. Therefore, it important that ethical considerations of such testing are thoroughly reflected on and discussed. I believe that aptitude testing has a great pedagogical potential in terms of facilitating adapted teaching for each pupil. Still, when you make use of a tool that describes so fundamental components of what makes a good language learner, there is a risk that the test will be used as more of a sorting tool. One must also be aware of other individual differences when measuring aptitude and using it in the education of the pupils. If other factors such as motivation or intelligence has a large impact on the language learning process of the pupil, this might affect the performances. If the teacher has then already evaluated this pupil as a bad language learner, this might have an undesired effect on the grades the pupil receives. In addition to this, when newer conceptualizations of the term are taken into consideration, aptitude can even be seen as a dynamic trait, that is able to develop and change over time. This can also create issues, at least when the teacher bases the teaching on aptitude results from several years earlier. I we think about Carroll and Sapon's (1959) MLAT, this aptitude battery was created to sort language learners into the correct language classes, based on how well they were predicted to learn the language, and not for enhancing the language learning.

The teacher of the class also had some concerns towards the idea of aptitude testing and the LLAMA. She explained that she had a general aversion towards grouping pupils from the same class. This has to do with the fear that some pupils will be viewed as uncapable and others as more capable. This did not seem to be a major concern for the teacher, given that the results and the following individually adapted teaching can be offered each pupil without revealing the language aptitude of every pupil for the rest of the class to avoid unpleasant situations. As long as the teacher is able to see beyond this grouping issue and facilitate so there will be no stigmatizing situations based on the results, this is a problem that can be tackled. Other negative

attitudes from the teacher pointed more towards the actual use of the LLAMA and the results that came from it, and not aptitude testing or the LLAMA itself. Still, I would argue that if aptitude testing is done correctly and with a functional aptitude test battery, the results could help teachers in their language teaching, based on the individual learning profiles of the pupils.

5.9 Limitations and Suggestions for Further Research

Since aptitude is viewed more as a dynamic trait now than before, I think investigating whether the aptitude of the pupils in this intervention changed during the schoolyear would be an interesting addition to the project. In the current intervention, only one test was conducted, and no follow-up test was administered. This was because pupils would recognize the LLAMA test and be familiar with the tasks. This in turn would probably affect the results to such an extent that there would be no point in running the test a second time. Still, the teacher offered a notion that her perspective on this and stated that she believed aptitude to be a dynamic trait, capable of changing. She explained that the task of spotting and assessing this is very difficult, but that there is a possibility in investigating how rapid the pupils take up new skills and abilities connected to language learning and how fast they understand tasks that are new to them. In this sense, aptitude can be seen as a dynamic trait, even though it was statistically and analytically impossible to prove in this intervention. Further research can therefore aim at using different aptitude batteries to measure if this change in aptitude is a reality. Granena (2019) has for instance discovered links between the LLAMA and the Hi-LAB, which can be used as a starting point for this type of testing.

For teachers and pupils to experience pedagogical advantages with the LLAMA, the school has to use a mapping tool like the LLAMA actively together so that every teacher receives training in how and when to administer the test and how to use the test results. This is also a general notion *Utdanningsdirektoratet* has stressed when informing about how to use mapping tests in the Norwegian educational system. That the mapping tools should be used as a foundation for individually adapted teaching and that the School administration has to take part in the work surrounding the testing to make best use of the results (Utdanningsdirektoratet, 2021). In addition to this, it is important that every teacher is aware of the current flaws concerning the LLAMA, (as mentioned in section 3.2) until the errors can be detected by the creators and corrected so that this does not become a barrier for using the LLAMA actively in teaching English. These errors are distinct weaknesses with this intervention and if these

barriers are lifted, I believe the LLAMA can function as a tool to help individually adapted teaching.

When a teacher uses the LLAMA, it is important that there exists a guide or framework that quite clearly, and in an understandable language, describes the pedagogical advantages of the LLAMA in terms of adapting the education. This LLAMA Functionality Framework has functioned well in this project, but a new guide with more accuracy and more elaborate explanations for how to use the results should be created. For this to happen I believe that a research project should be started for the purpose of developing a pedagogical guide to how the results of the LLAMA should be used. A theoretical investigation must be conducted, where research based on the LLAMA is gathered and compared in order to find pedagogical implications that has correlation to the test results. If this information is synthesised and made understandable in a concise functionality frame, there might be a possibility for actively using the LLAMA in Norwegian upper secondary schools.

Another limitation with this study is that the attitudes towards aptitude testing and the LLAMA is of a highly qualitative format. Research should be done in order to investigate how other teachers in the Norwegian school system views the idea of aptitude testing. Even though the teacher in this intervention was very positive, other teachers might be more restrained towards the idea. Since teacher attitudes might be the most important key to realise the goals of this master's thesis, I believe it to be an important area of investigation for further research.

The last limitation from this study I will discuss has to do with the evidence from this study showing that the LLAMA is a suitable aptitude test battery for the age group of 15-16. Younger age groups should be tested and evaluated on the same terms as in this study. Rogers et al. (2017) investigated learners aged 10-11, whereas this study investigated learners aged 15-16. Research should be done to see how learners between these two age groups perform on the LLAMA test and also if the reason for the high performances for the current learner group lies in a general high language learning aptitude, or if it is connected mainly to age. This could be done by testing both learners in the same age group as the current study and later test learners in the ages down to 12 years old.

6. Conclusion

The conclusion will aim at using the discussion of all the RQs of this thesis to make some final assertions about whether teachers should strive to use aptitude testing and the LLAMA in Norwegian classrooms or not. The evidence from the project suggests several advantages by actively using aptitude testing as a mapping tool to enhance individually adapted teaching, but this testing and especially using the results actively in teaching English, are time consuming tasks and a challenge that needs a lot of effort from the teacher.

Based on the results from the interventions made in this project and the following discussion that enlighten some of the results found, I am personally not in doubt when I say that we should use aptitude testing for mapping in upper secondary school. The results from the intervention suggest that the pupils had a decent experience with using the LLAMA and even saw some use in doing it. The teacher was intrigued and had almost only positive feedback to give with regards to the test and the discussion showed that there are definitive pedagogical advantages with using the results from the LLAMA. With much research pointing towards there being a clear relations hip between aptitude and cognitive styles, adapting teaching becomes essential (see section 2.5). Current research on the field has also proven that there are clear advantages by using these aptitude profiles, by using an aptitude-treatment interaction method where instructional techniques and methods are matched with pupil abilities to facilitate individually adapted teaching (Hwu & Sun, 2012; Robinson, 2007; Wen, Biedron & Skehan, 2016). There is also sufficient evidence that the validity and reliability of the LLAMA should be sufficient for these test situations as this test is used to help the learning process of the pupil, which cannot be said to be a high-stake situation (see Granena, 2013; Meara, 2005). So, all in all, from the results of this intervention, supplied with current research, there is little doubt that conducting an aptitude test to map pupils in upper secondary school is a good idea.

When it comes to the test itself, the intervention has shown that running the test and the actual testing procedure is quite unproblematic in this context, with this learner group. The test is easy to administer and use, free of charge, politically neutral and has a modern computer-based set up. In addition to this, the test only takes about 25 minutes to administer and the only thing that is required to run it is a computer, pen and paper to take notes, and hearing devices (not AirPods) and you are all set to use the test. The test has also, even though it was not used in this intervention because of practical and ethical reasons, an automatic scoring system the

teacher can read after the test is done (Meara, 2005). The importance for future use of this test is still that the creators of the test make sure that as many potential errors as possible are corrected so that teachers are not bothered with trying to figure out how to solve this when testing. These errors occur because the LLAMA lacks the proper standardization other older aptitude tests have already acquired (Granena, 2013). With these factors in mind and a correction of errors, the LLAMA test will presumably contribute to the willingness to use aptitude testing as an integrated part of the English education of Norwegian upper secondary school pupils.

Even though the LLAMA itself is easy to administer, there is a general lack of a proper and clear way to use the results from the testing in teaching English. Since there is no clear guide to how the results of the LLAMA should be used in teaching English, I have had to create my own Functionality Frame (see Appendix F) for this project. I think that if teachers are to use aptitude testing in classroom, there has to be a more thorough and well-developed guide that can help teachers to use the results in a logical way. Haukås (2012) showed that 67% of teachers agreed or partially agreed that they needed to learn more about language learning strategies before they would use them in teaching. Therefore, without this clear guide and teacher training in how to use the results, I fear that teachers will not see the benefit in using the LLAMA for other purposes than finding a langue learning potential with the pupils. There has been done a lot of research on how to match aptitude scores with instructional methods (Erlam, 2005; Granena, 2013; Hwu & Sun, 2012; Poshner, 2018; Robinson 2002; Sternberg, Grigorenko & Zhang, 2008; Wen, Biedron & Skehan, 2016; Yilmaz, 2013). The problem is just that there are no studies that have used the LLAMA as a starting point for investigating how to use the results produced from this particular aptitude battery.

Given that there some time would be produced a sufficient LLAMA Functionality Frame for teachers to use, there is still little doubt that a functional and proper use of the results in teaching English is a time consuming and effortful task for teachers. “Ay, think of it; wish it done; will it, - but to do it!” These are words from Ibsen’s Peer Gynt, (Act 3, Scene 1) which expresses Peer’s thoughts about doing something, but his cowardice when it comes to actually doing something (Haukås, 2018). The point here is not that teachers are cowards, but that teachers often have ideas about what to do and how to do things in instructional setting, but do not see it done. An example which comes quite close to the content of this project is the research done by Haukås (2012) on Norwegian teachers’ attitudes to language learning strategies. The study showed that 90,5% of the teachers agreed or partially agreed that “Language learning

strategies should be a natural part of teaching through the whole of the school year." Still, 62% of the teachers also answered that their students seldom or never tried out different vocabulary learning strategies in class (Hukås, 2012). This show that there often is a gap between what teachers think they should do and what they actually do. Time pressure and a lack of knowledge about aptitude could be factors that makes teachers skip using aptitude testing, even though many of them might think it is a good idea. That is why teachers needs to be safe and understand what they are doing when using the LLAMA and also be granted enough time, for instance in meetings between teachers of English, to work with the concept of aptitude testing and the LLAMA. With enough training, time and commitment, I think the LLAMA could be a valuable asset for the Norwegian educational system if the teachers can be able to change the way they think into the way they act.

References

Artieda, G., & Muñoz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. In P. Cirino (Ed.). *Learning and Individual Differences, 50,* 42-48. https://doi.org/10.1016/j.lindif.2016.06.023

Ameringer, V., Green, L., Leisser, D., & Turker, S. (2018). Introduction: Towards an Interdisciplinary Understanding of Language Aptitude. In Reiterer, S. M. (Ed.), Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience (pp. 1-18). Springer.

Bokander, L. (2019). Predictive validity of the LLAMA language aptitude tests in a group of mixed L1 beginner learners of Swedish. *EuroSLA 29: The 29th Conference of the European Second Language Association* (pp. 202-203). Lund University.

Bokander, L, & Bylund, E. (2019). Probing the Internal Validity of the LLAMA Language Aptitude Tests. *Language Learning: A Journal of Research in Language Studies, 70*(1), 11-47. https://doi-org.mime.uit.no/10.1111/lang.12368

Brevik, L. M., & Helness, H. L. (2018). Engelsk VG1: Nye læringsstøttende prøver. *Bedre Skole.* Retrieved May 1, 2021 from https://www.utdanningsnytt.no/files/2019/06/27/Bedre%20Skole%202%202018.pdf

Brøyn, T. (2019). Flerspråklighet – et fruktbart kaos. *Bedre Skole.* Retrieved May 5, 2021 from https://www.utdanningsnytt.no/files/2020/07/08/UTD-BedreSkole-0419-WEB.pdf

Burner, T. (2018). Formative Assessment in English. In H. Bøhn, M. Dypedahl, & G-A Myklevold (Eds.), *Teaching and Learning English* (pp. 248-265). Cappelen Damm Akademisk.

Burner, T., & Carlsen, C. (2019). "I mix all the spark": om engelskopplæring i flerspråklige klasserom. *Bedre Skole.* Retrieved may 5, 2021 from https://www.utdanningsnytt.no/files/2020/07/08/UTD-BedreSkole-0419-WEB.pdf

Carroll, J. B. (1981). Twenty-five years of research in foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude.* Newbury House.

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

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition.* Taylor & Francis Group.

Dörnyei, Z., & Csizér, K. (2012). How to Design and Analyze Surveys in Second Language Acquisition Research. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide*. Wiley-Blackwell.

Dörnyei, Z., & Skehan, P. (2003). Individual Differences in Second Language Learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 589-630). Blackwell Publishing.

Duff, P. A. (2012). How to Carry Out Case Study Research. In A. Mackey & S. M. Gass (Eds.). *Research Methods in Second Language Acquisition: A Practical Guide.* Wiley-Blackwell.

Engelsen, B. U. (2012). *Kan læring planlegges?: arbeid med læreplaner - hva, hvordan, hvorfor.* Gyldendal Akademisk.

Epstein S. (1990). Cognitive-experiential self-theory. In L. Pervin (Ed.), *Handbook of personality theory and research* (pp. 165 -192). Guilford Press.

Erlam, R. (2003). The effects of deductive and inductive instruction on the acquisition of direct object pronouns in French as a second language. *Modern Language Journal, 87*(2), 242.260. https://doi.org/10.1111/1540-4781.00188

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. In *Language Teaching Research, 9*(2), 147-172. DOI: 10.1191/1362168805lr161oa

Granena, G. (2013). Cognitive Aptitudes for Second Language Learning and the LLAMA Language Aptitude Test. In G. Granena & M. Long (Eds.), *Sensitive Periods,*

Language Aptitude, and Ultimate L2 Attainment (pp. 105-130). John Benjamins Publishing Company.

Granena, G. (2014). Language Aptitude and Long-Term Achievement in Early Childhood L2 Learners. *Applied Linguistics, 35*(4), 483-503. https://doi-org.mime.uit.no/10.1093/applin/amu013

Granena, G. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics, 37*, 577–600. doi:10.1017/S0142716415000120

Granena, G. (2019). Cognitive Aptitudes and L2 Speaking Proficiency: Links Between LLAMA and Hi-LAB. *Studies in Second Language Acquisition, 41*, 313-336. DOI: 10.1017/S0272263118000256.

Granena, G. (2020). *Implicit Language Aptitude.* Cambridge University Press. DOI: 10.1017/9781108625616.

Granena, G., & Long, M. H. (2012). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311–343. DOI: 10.1177/0267658312461497.

Habl, C. (2018). Language Aptitude and Gender. In M. Reiterer (Ed.), Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience (pp. 229-242). Springer.Housen, A., & Pierrard, M. (2005). Investigating Instructed Second Language Acquisition. In A. Housen & M. Pierrard (Eds.), *Investigations in Instructed Second Language Acquisition* (pp. 1-30). Mouton de Gruyter.

Hauptmann, P. C. (1971). A structural approach vs. a situational approach to foreign-language teaching. *Language Learning, 21*, 235-244. https://doi.org/10.1111/j.1467-1770.1971.tb00062.x

Haukås, Å. (2012). Lærarhaldningar til språklæringsstrategiar. *Norsk Pedagogisk Tidsskrift, 96*(2), 114-128.

Haukås, Å. (2018). Teacher's Beliefs about Language Instruction. In H. Bhøn, M. Dypedahl & G-A. Myklevold (Eds.). *Teaching and Learning English* (pp. 343-357). Cappelen Damm Akademisk.

Hörder, S. (2018). The Correlation of Early Multilingualism and Language Aptitude. In M. Reiterer (Ed.), *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (pp. 277-304). Springer.

Hummel, K. M. (2013). *Introducing Second Language Acquisition: Perspectives and Practices.* Wiley Blackwell.

Hwu, F., & Sun, S. (2012). The aptitude-treatment interaction effects on the learning of grammar rules. *System, 40*(4), 505-521. https://doi.org/10.1016/j.system.2012.10.009

Hwu, F., Pan, W., & Sun, S. (2013). Aptitude-treatment interaction effects on explicit rule learning: A latent growth curve analysis. *Language Teaching Research, 18*(3), 294-319. https://doi-org.mime.uit.no/10.1177/1362168813510381

Kazuya, S. (2017). Effects of Sound, Vocabulary, and Grammar Learning Aptitude on Adult Second Language Speech Attainment in Foreign Language Classrooms. *Language Learning: A Journal of Research in Language Studies, 67*(3), 665-693. https://doi-org.mime.uit.no/10.1111/lang.12244

Kourtali, N-E., & Révész, A. (2019). The Roles of Recasts, Task Complexity, and Aptitude in Child Second Language Development. *Language Learning: A Journal of Research in Language Studies, 70*(1), 179-218. https://doi-org.mime.uit.no/10.1111/lang.12374

Lee, J. F., & Van Patten, B. (2003). *Making Communicative Language Teaching Happen*. (2nd Ed.). McGraw-Hill.

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

Lightbown, P. M., & Spada, N. (2013). *How Languages are Learned* (4th ed.). Oxford University Press.

Linck, J., Hughes, M., Campbell, S., Silbert, N., Tare, M., Jackson, S., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level

language proficiency. *Language Learning: A Journal of Research in Language Studies*, *63*(3), *530*–566. https://doi-org.mime.uit.no/10.1111/lang.12011

Maddah, Z. G., & Reiterer, S. M. (2018). Language Transfer vs. Language Talent? Individual Differences and Aptitude in L2 Phonology of Persian-Speaking Learners of English. In M. Reiterer (Ed.), *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (pp. 363-388). Springer.

Meara, P. (2005). *LLAMA Language Aptitude Tests: The Manual.* University of Wales Swansea.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*(6), 972–987. https://doi.org/10.1037/0022-3514.76.6.972

Pimsleur, P. (1966). *The Pimsleur Language Aptitude Battery.* Harcourt, Brace, Jovanovic.

Poschner, J. (2018). Vocabulary Acquisition Strategies and Language Aptitude. In M. Reiterer (Ed.), *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (pp. 245-260). Springer.

Rizvanovic, N. (2018). Motivation and Personality in Language Aptitude. In S. M. Reiterer (Ed.). *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (pp. 101-116).

Robinson, P. (2001) Individual differences, cognitive abilities, aptitude complexes and learning conditions in second language acquisition. *Second Language Research, 17*(4), *368-392.* DOI: 10.1177/026765830101700405

Robinson, P. (2002). Learning conditions, aptitude complexes and SLA: A framework for research and pedagogy. In Robinson, P. (Ed.), *Individual Differences and Instructed Langauge Learning* (pp. 113–133). John Benjamins.

Robinson, P. (2005). Aptitude and Second Language Acquisition. In *Annual Review of Applied Linguistics*, *25*, 46-73. Cambridge University Press. https://doi.org/10.1017/S0267190505000036

Robinson, P. (2007). Aptitudes, abilities, contexts, and practice. In R. M. DeKeyser. (Ed.), *Practice in second language* (pp. 256–286). Cambridge University Press.

Rogers, V. E., Meara, P., Aspinall, R., Fallon, L., Goss, T., Keey, E., & Thomas, R. (2016). Testing Aptitude: Investigating Meara's (2005) LLAMA Tests. *EUROSLA Yearbook, 16*(1), 179-210. https://doi.org/10.1075/eurosla.16.07rog

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

Sawyer, M. & Ranta. L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 319-353). Cambridge University Press.

Skehan, P. (1989). *Individual Differences and Second-Language Learning.* Edward Arnold.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford University Press.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual Differences and Instructed Language Learning* (pp. 69-93). John Benjamins Publishing Company.

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

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

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language. *Applied Psycholinguistics, 25*(2), 293-321. doi:10.1017/S0142716404001146

Utdanningsdirektoratet, (2021). *Kartleggingsprøver*. Retrieved 1 May from, https://www.hivolda.no/biblioteket/skrive-og-referere/apa-eksempelsamling/apa-offentlege-kjelder

Wen, Z. (2011). Foreign Language Aptitude. *ELT Journal, 66*(2), 233-235. https://doi.org/10.1093/elt/ccr068

Wen, Z., Biedroń, A., & Skehan, P. (2017). Foreign Language Aptitude Theory: Yesterday, Today and Tomorrow. *Language Teaching, 50*(1), 1-31. 10.1017/S0261444816000276

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

Yalçın, Ş., Çeçen, S. & Erçetin, G. (2016). The realtionship between aptitude and working memory: an instructed SLA context. *Language Awareness, 25*(1-2), 144-158.

Yalcin, S., & Spada, N. (2016). Language Aptitude and Grammatical Difficulty: An EFL Classroom-Based Study. *Studies in Second Language Acquisition,* 38, 239-263. doi:10.1017/S0272263115000509.

Yilmaz, Y. (2013). Relative Effects of Explicit and Implicit Feedback: The Role of Working Memory Capacity and Language Analytic Ability. *Applied Linguistics, 34*(3), 344-368. https://doi-org.mime.uit.no/10.1093/applin/ams044

Appendix

Appendix A - Letter of consent

Vil du delta i forskningsprosjektet

Aptitude in the Classroom: an empirical study of the pedagogical functionality of the LLAMA test battery in an upper secondary school?

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å undersøke hvordan man kan teste elevers språkøre i videregående skole og hvordan det kan hjelpe å legge til rette for undervisningen. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål

Dette forskningsprosjektet er en del av min masteroppgave på lektorutdanningen 8.-13. trinn i faget engelsk. I min oppgave skal jeg gjøre en undersøkelse av hvordan man best kan teste elevens språkøre og om dette er hensiktsmessig for en lærer å benytte seg av i engelskundervisningen i videregående skole, samt hvilke fordeler dette kan ha for elever og lærer.

Dataene som samles inn i dette prosjektet vil kun bli brukt av meg selv til min masteroppgave og av faglærer, for å legge til rette for undervisningen i faget og vurdere nytten elever og lærer kan ha av en slik kartlegging. Alle personopplysninger vil såklart bli anonymisert

Hvem er ansvarlig for forskningsprosjektet?

UiT – Norges Arktiske Universitet er ansvarlig for prosjektet.

Hvorfor får du spørsmål om å delta?

Grunnen til at akkurat du blir spurt om å delta er fordi jeg har blitt tildelt din faglærer som praksisveileder for min 5.-årspraksis på lektorutdanningen. Klassen er valgt ut som passende siden jeg anser at en klasse på videregående skole i engelsk gir de best mulige resultatene for en studie slik som denne. Det er kun din klasse i engelsk som inkluderes i dette prosjektet, samt din faglærer.

Hva innebærer det for deg å delta?

Dersom du velger å delta vil du bli bedt om å gjennomføre en test som kalles *LLAMA*. Dette er en test som gir deg ulike oppgaver for å prøve ut ditt språkøre. Denne testen gjennomføres

på nett og tar ca. 30 minutter. Du vil også motta et spørreskjema der du får noen spørsmål med valgalternativer som omhandler din opplevelse av LLAMA-testen. Resultatene registrerer du på et ark du vil få utdelt i forkant av testen. De samlede resultatene av testen og spørreskjemaet vil lagres elektronisk på en sikker harddisk, der navn og andre opplysninger om deg ikke vil komme frem. Det kan også bli aktuelt å innhente informasjon om dine resultater i faget engelsk fra din faglærer på videregående skole. Disse vil heller ikke komme frem i oppgaven eller i testen på noen måte.

Dersom dine foreldre ønsker å se spørreskjema eller testen i seg selv er det bare å ta kontakt med meg.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg. De som ikke ønsker å delta vil få et alternativt opplegg den aktuelle skoletimen og må allikevel møte opp på skolen.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger
Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket. De som vil ha tilgang til dine opplysninger ved UiT er meg selv og mine to masterveiledere. Jeg vil også lage en kode for ditt navn som bare jeg og din faglærer vet om slik at ingen vil kunne spore opplysningene til deg. Alle data vil være låst inne på en egen låst mappe, adskilt fra annet materiale jeg besitter. Kun resultatene dine på testen og spørreskjemaet vil publiserer sammen med alle de andre i klassen sine resultater. Det vil ikke komme frem noe navn og man vil derfor ikke kunne kjenne deg igjen.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?
Opplysningene anonymiseres når prosjektet avsluttes/oppgaven er godkjent, noe som etter planen er i mai 2021. Alle data vil slettes etter denne datoen og ingen vil kunne finne de anonymiserte dataene.

Dine rettigheter
Så lenge du kan identifiseres i datamaterialet, har du rett til:
- innsyn i hvilke personopplysninger som er registrert om deg, og å få utlevert en kopi av opplysningene,
- å få rettet personopplysninger om deg,
- å få slettet personopplysninger om deg, og
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra UiT – Norges Arktiske Universitet har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?
Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med:

- UiT – Norges Arktiske Universitet ved:
  Christopher Loe Olsen (veileder)
  Christopher.l.olsen@uit.no

  Natalia Mitrofanova (veileder)
  natalia.mitrofanova@uit.no

  Morten Skillingstad Larsen (student)
  Mla158@post.uit.no

- Vårt personvernombud: Joakim Bakkevold
  personvernombud@uit.no
  777 46 322 og 976 915 78

Hvis du har spørsmål knyttet til NSD sin vurdering av prosjektet, kan du ta kontakt med:

- NSD – Norsk senter for forskningsdata AS på epost (personverntjenester@nsd.no) eller på telefon: 55 58 21 17.

Med vennlig hilsen

Christopher Loe Olsen og Natalia Mitrofanova
(Forskere/veiledere)

Morten Skillingstad Larsen
(Student)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet *Aptitude in the Classroom* og har fått anledning til å stille spørsmål. Jeg samtykker til:

- ☐ å delta i LLAMA test-batteriet
- ☐ å delta i spørreskjema om din opplevelse av test-batteriet
- ☐ at faglæreren din i engelsk kan gi opplysninger om meg til prosjektet – hvis aktuelt

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

---

(Signert av prosjektdeltaker, dato)

Appendix B - Teachers' aptitude assessment of pupils

Description of the levels:

1 = Pupil might often have a hard time learning English and uses a lot of effort and work to acquire new domains of the language. This score would approximately reflect a result in the area of 0-15 on the LLAMA test battery.

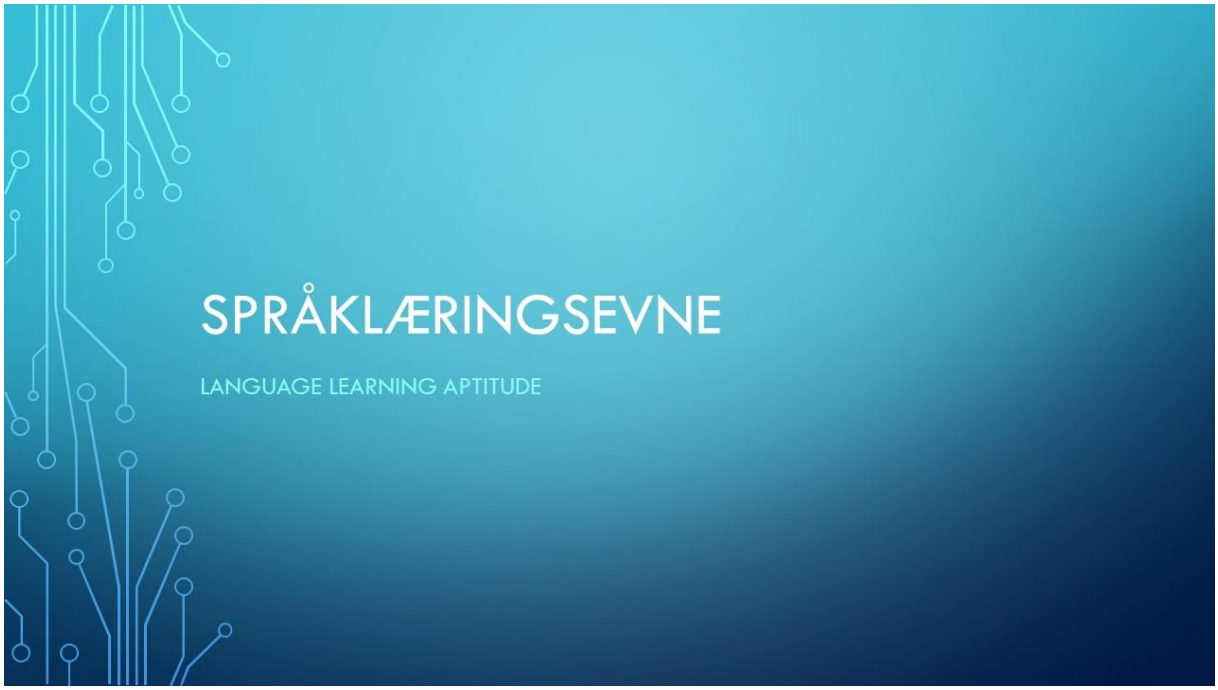
2 = Learning English comes quite natural to this pupil and the pupil does not seem to have more difficulties in the development of the language than what you would expect in this level of English education. This score would approximately reflect a result in the area of 20-45 on the LLAMA test battery.

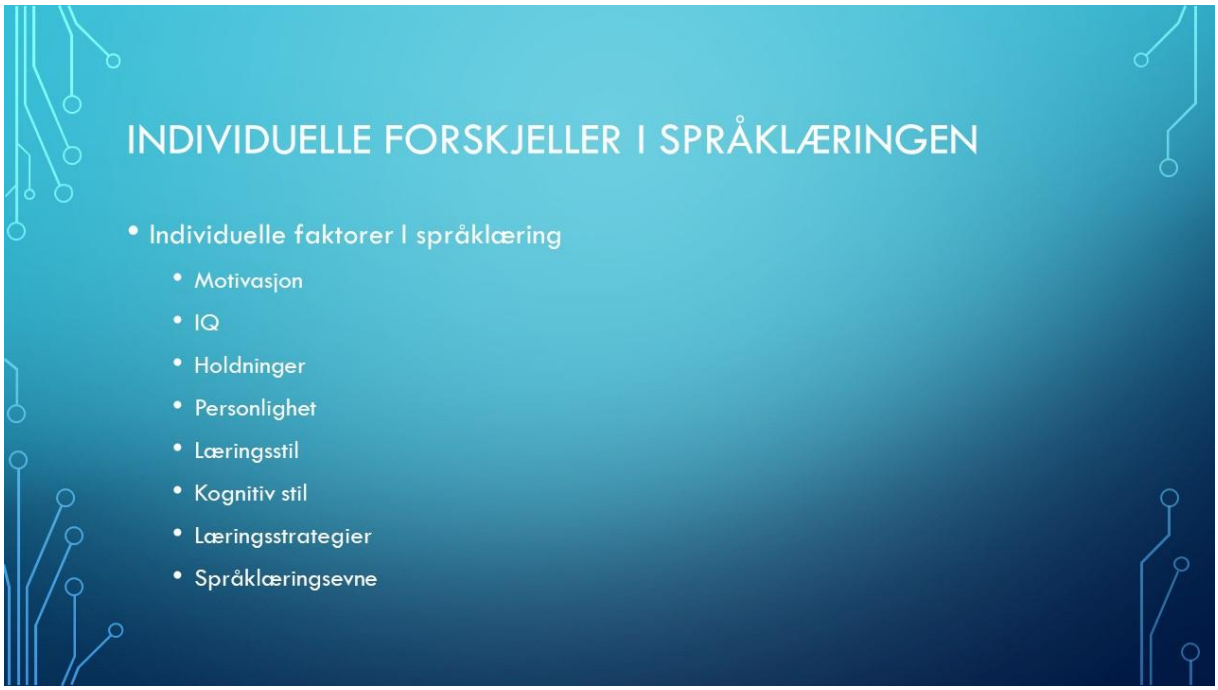
3 = Pupil easily acquires new knowledge of unfamiliar domains of English, without too much effort and the language seems to come naturally and at a faster pace than expected in this level of English education. This score would approximately reflect a result in the area of 50-70 on the LLAMA test battery.

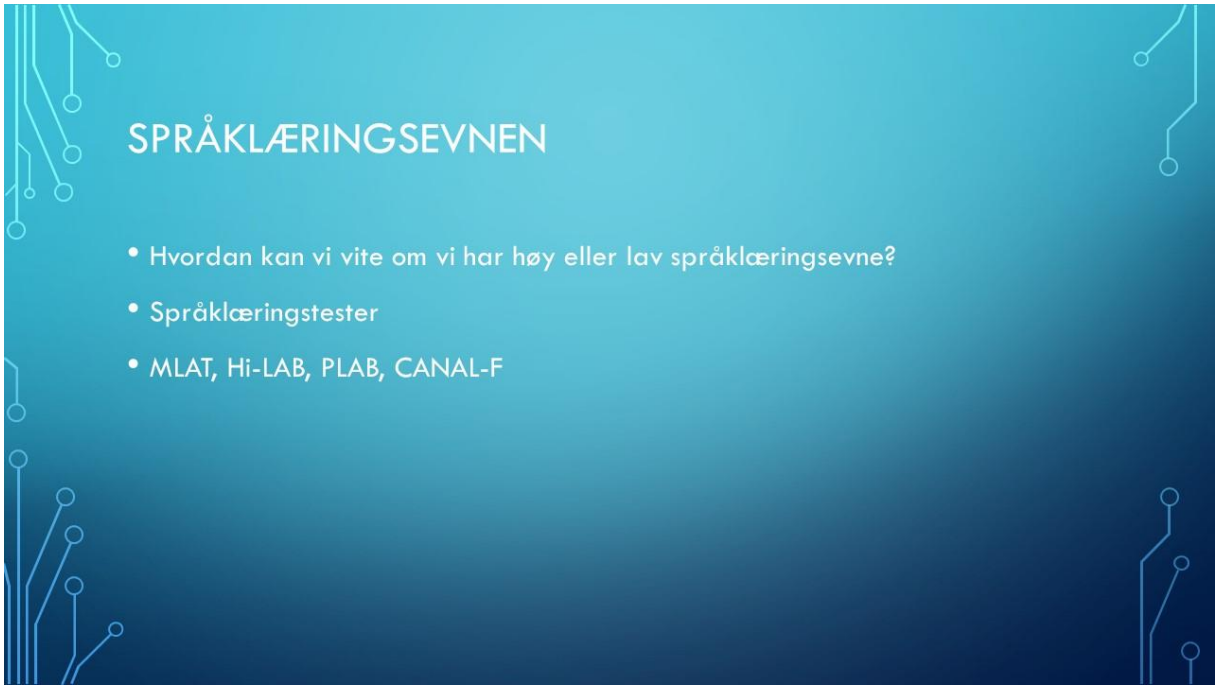
4 = Pupil has exceptional language aptitude and one can easily spot that this is a language talent far beyond what is expected at this level of English education. This score would approximately reflect a result in the area of 75-100 on the LLAMA test battery.

Pupil nr.	**Anticipated aptitude level (1-4)**	**Pupil nr.**	**Anticipated aptitude level (1-4)**	**Pupil nr.**	**Anticipated aptitude level (1-4)**
1		10		19	
2		11		20	
3		12		21	
4		13		22	
5		14		23	
6		15		24	
7		16		25	
8		17		26	
9		18		27	

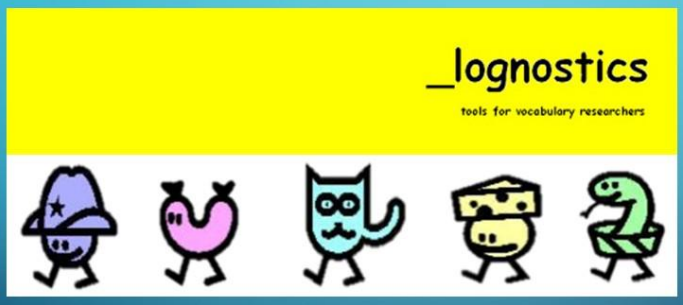
Appendix C - PowerPoint with LLAMA Instructions




INDIVIDUELLE FORSKJELLER I SPRÅKLÆRINGEN
Individuelle faktorer I språklæring
Motivasjon
IQ
Holdninger
Personlighet
Læringsstil
Kognitiv stil
Læringsstrategier
Språklæringsevne


SPRÅKLÆRINGSEVNEN
Hvordan kan vi vite om vi har høy eller lav språklæringsevne?
Språklæringstester
MLAT, Hi-LAB, PLAB, CANAL-F

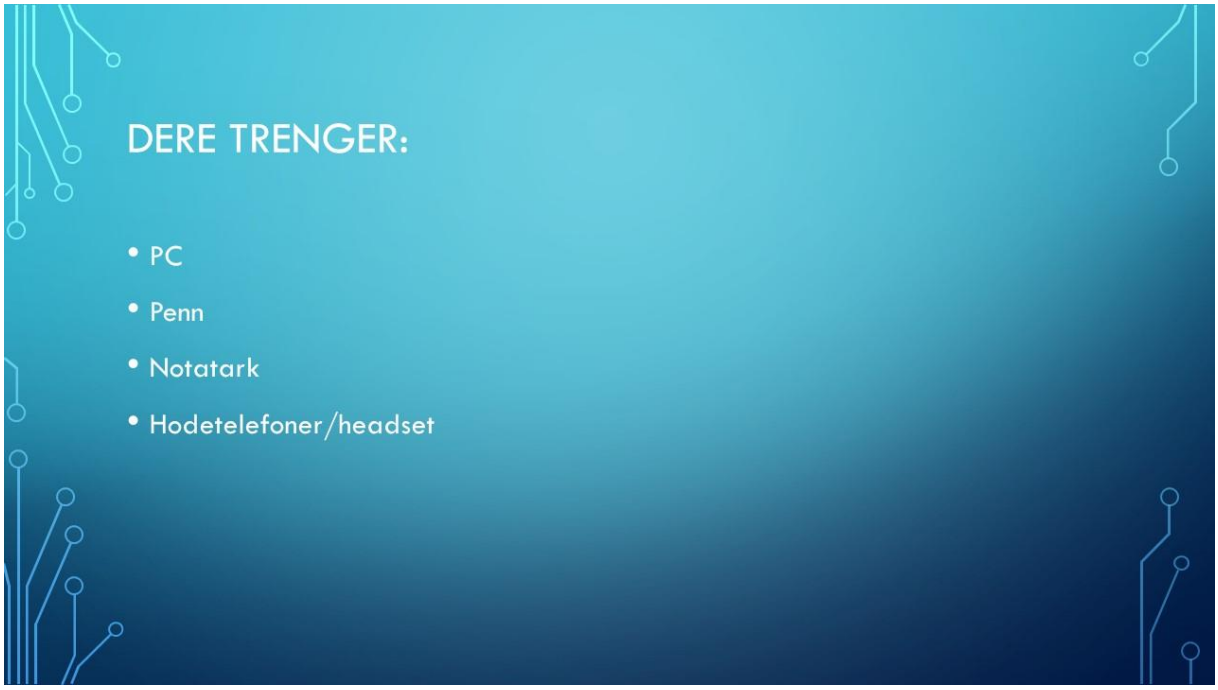
LLAMA



MIN MASTEROPPGAVE

«Aptitude in the Classroom: an empirical study of the pedagogical functionality of the LLAMA test battery in an upper secondary school»


MIN MASTEROPPGAVE
• Passer LLAMA som en språklæringstest for en VG1-klasse?
• Hvilke fordeler kan språklæringstesten ha for engelskundervisningen?
• Hvordan synes elevene i Videregående skole at LLAMA fungerte?


DERE TRENGER:
• PC
• Penn
• Notatark
• Hodetelefoner/headset

LA OSS TESTE!

- http://www.lognostics.co.uk/tools/llama/

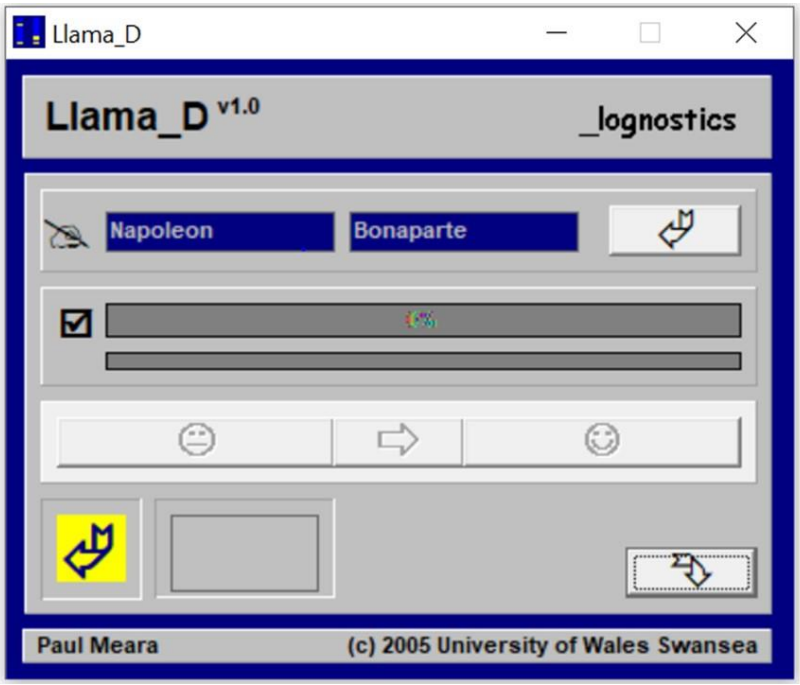
Free software from _lognostics
The LLAMA Language Aptitude Tests

The Llama programs replace the Lat03 suite of programs.
The Llama tests are an ongoing demonstrator project that we use to teach research skills to undergraduates. The result is a set of innovative, and slightly-off-the-wall tests which we think might be used to assess language aptitude. In terms of theory, the current suite is largely based on the MLAT tests, but the formats have been radically adapted to a more snazzy presentation style.
The Llama tests are very experimental, and should be handled with extreme caution. Although they are user-friendly, and quick to use, they have not been extensively standardised, and should NOT be considered a replacement for MLAT in high-stakes situations.
The tests have now been redesigned so that you can download individual tests, instead of the entire suite. This makes the programs more manageable: all the files you need are included in a single download for each subtest.
Four sub-tests are currently available:
Llama_B a vocabulary learning task
Llama_D a test of phonetic memory
Llama_E a test of sound-symbol correspondence
Llama_F a test of grammatical inferencing
The manual file includes instructions for downloading the tests, and and directioons for how to use each individual subtest.

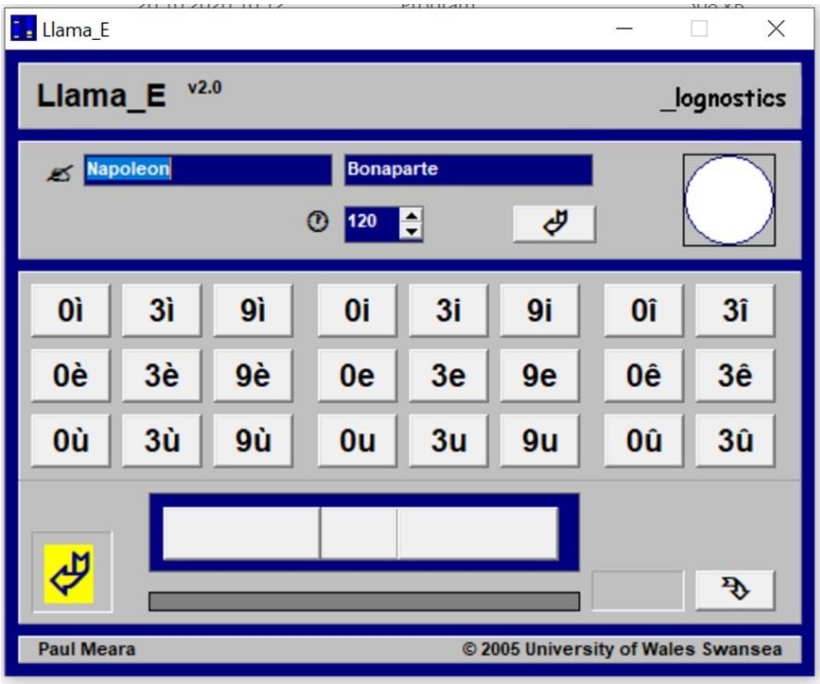
- download the Llama Manual
- download the Llama_B executables
- download the Llama_D executables

- download the Llama Manual
- download the Llama_B executables
- download the Llama_D executables
- download the Llama_E executables
- download the Llama_F executables
- download the Llama_Data_Reader

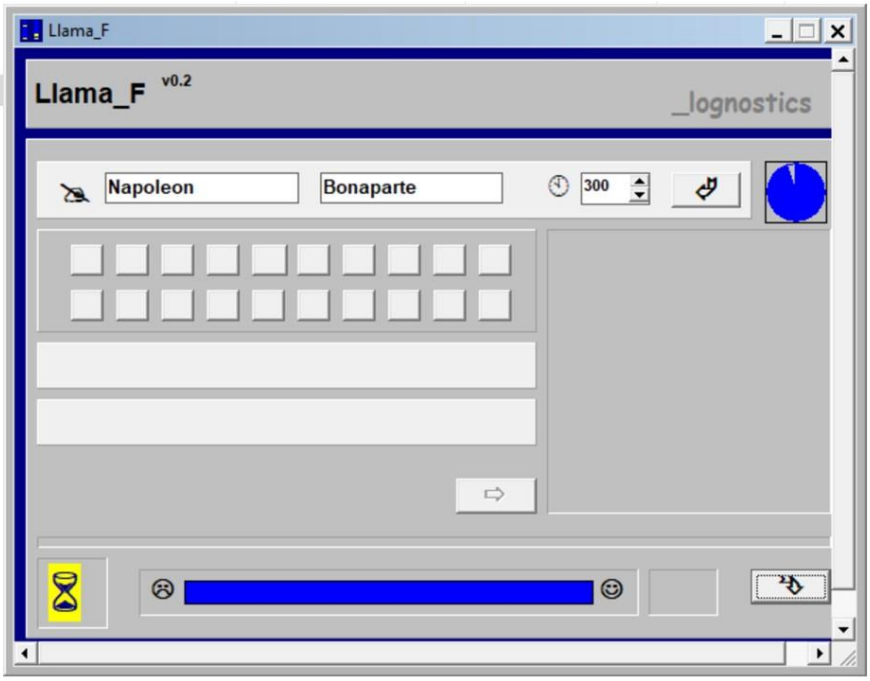
- download the Llama Manual
- download the Llama_B executables
- download the Llama_D executables
- download the Llama_E executables
- download the Llama_F executables
- download the Llama_Data_Reader

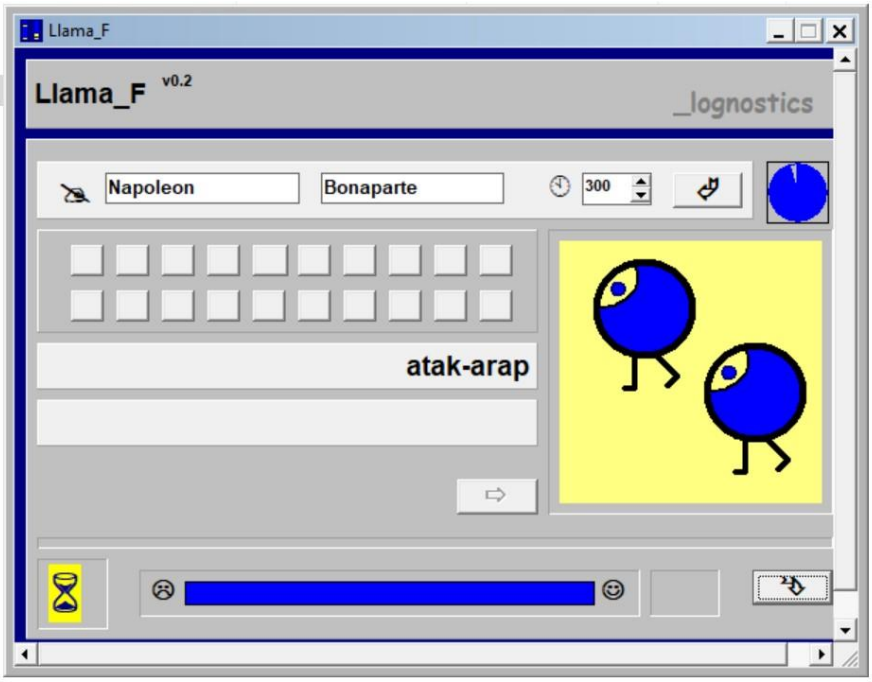


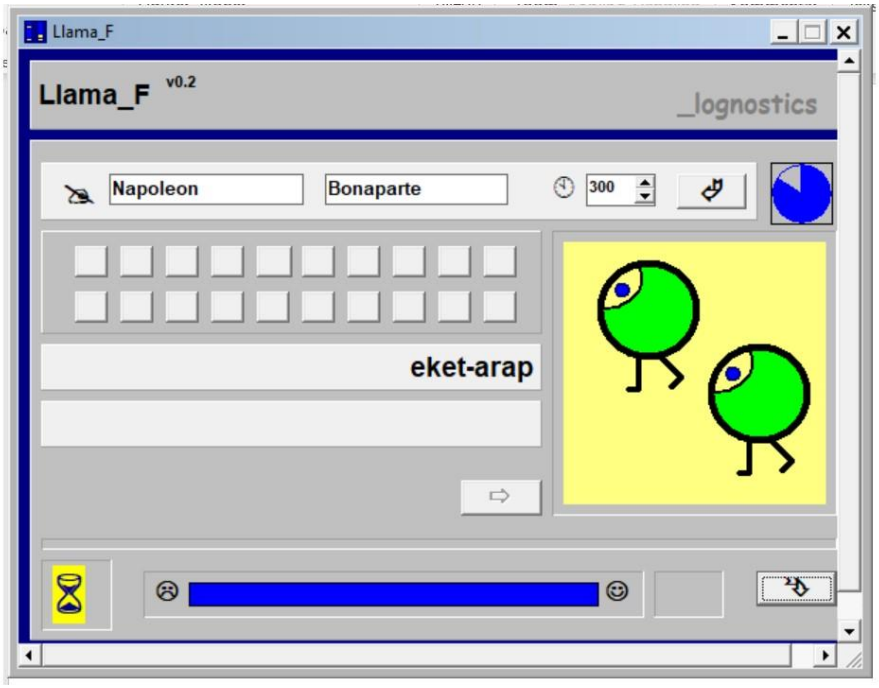
- download the Llama Manual
- download the Llama_B executables
- download the Llama_D executables
- download the Llama_E executables
- download the Llama_F executables
- download the Llama_Data_Reader



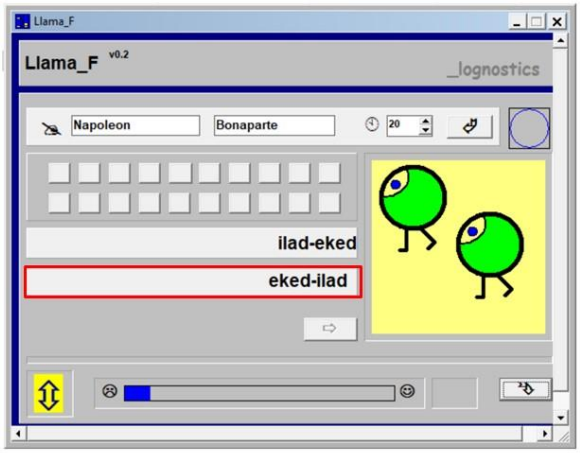
- download the Llama Manual
- download the Llama_B executables
- download the Llama_D executables
- download the Llama_E executables
- download the Llama_F executables
- download the Llama_Data_Reader

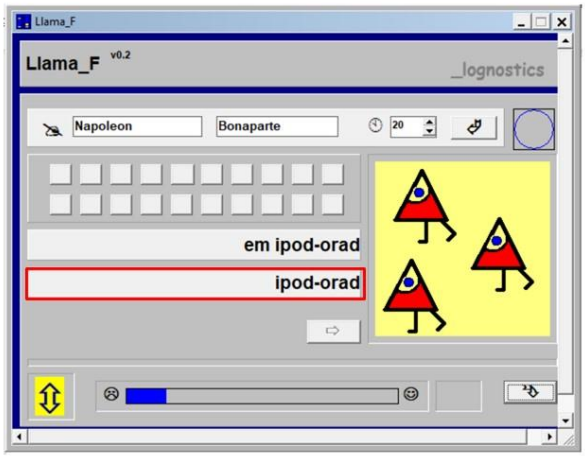

Llama_F
Llama_F v0.2
_lognostics
Napoleon
Bonaparte
300


Llama_F
Llama_F v0.2
_lognostics
Napoleon
Bonaparte
300
atak-arap



Feil i LLAMA_F





Til slutt: Spørreskjema

Appendix D - LLAMA Test Battery Result Sheet

Participant nr: _____

Subtest	Score
LLAMA_B	
LLAMA_D	
LLAMA_E	
LLAMA_F	

Appendix E - Questionnaire for pupils

Questionnaire for pupils

Participant nr _____

1. What is your gender?

Boy - Girl - Other/prefer not to answer

2. How well do you like English as a school subject from 1-5?

(1 = not at all, 5 = like it a lot)

1 2 3 4 5

3. How was your overall experience with using the LLAMA?

(1 = very bad, 5 = very good)

1 2 3 4 5

4. Were the tasks easy to understand?

(1 = very difficult to understand, 5 = very easy to understand)

1 2 3 4 5

5. Do you think this test, and knowing your language aptitude, is useful for your education?

(1 = not at all, 5 = very useful)

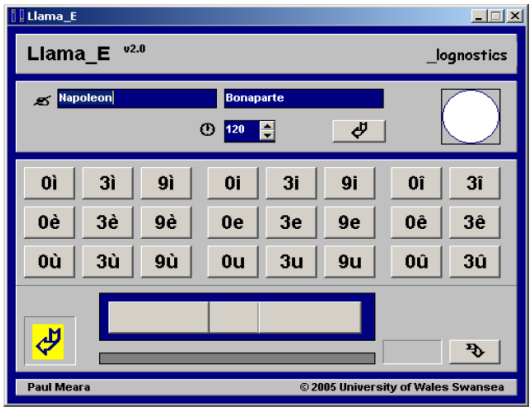
1 2 3 4 5

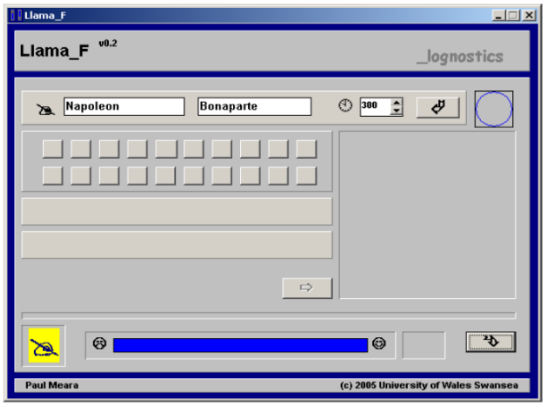
6. Did the results you got correspond with your previous assumption of your aptitude?

(1 = not at all, 5 = absolutely, yes)

1 2 3 4 5

7. Look at the four images of the subtests of the LLAMA below. Rate the subtests from 1 – 4 on which subtest you liked the most. 1 is the subtest you liked the most and 4 is the subtest you liked the least. Write the numbers in the green boxes next to the images of the subtests.





8. If you have any comments to the questions, the test or any additional thoughts on this subject you would like to add, please elaborate below:

_____

_____

_____

_____

Appendix F - LLAMA Functionality Frame

The Guide

This paper is a guide to how the LLAMA test results can be used in practice. The guide breaks down every LLAMA subtest and explains how a certain score can be used to understand the learning ability of the participant and how these results in turn can be used in a language learning situation. The guide explains several learning strategies and learning methods which, according to current research on the area, could prove to be useful for the language learner in question. This guide will not offer any explanation of the test, its background or theoretical framework, nor will it describe how the subtests works. For further information about this see; Meara (2005). The guide will break down every subtest and firstly explain the research that lies behind the suggested use of the LLAMA-test results, and secondly summarize the main points of functionality that can be applied to the teaching of the different aptitude areas. In the end of the guide a section called “General Functionality Tips” will be presented. This section will offer some general tips in how to use the test results, without going specifically into one specific LLAMA subtest, but rather describe how the results can be used in a more general way.

LLAMA_B

The research

The LLAMA_B is a vocabulary learning test, which measures the participants ability to learn a relatively large amount of vocabulary in a short time span (Meara, 2005). This test is graded from 0-100 where the different scores are described as follows: 0-20: a very poor score, 25-45: an average score, 50-70: a good score, 75-100: an outstandingly good score.

Poschner’s recent study (2018) on vocabulary learning skills and its connection to language learning strategies is closely related to answering the question of how to apply the results from the LLAMA_B to practical teaching and pedagogy. The main findings of Poschner was that there is no difference between the use of cognitive vocabulary strategies between high- and low scorers of the LLAMA_B, i.e. high- or low vocabulary learning students. His study

shows that the low scorers might experience great benefits by using these strategies. Still, they do not use the strategies and therefore it is important to focus on these strategies when working on vocabulary acquisition for low scorers of the LLAMA_B. The high scorers are not more aware of these strategies than the low scorers, but they do not seem to have as much use and as high a learning benefit by using these strategies as the low scorers do.

The cognitive strategies in question here are mnemonic strategies, learning with pictorial representations, the use of synonyms and antonyms, grouping words together in meaningful groups, and using no specific technique or strategy (Poschner, 2018).

Summarized functionality tips

- Low and high scorers of the LLAMA_B are not aware of- and do not use cognitive language learning strategies.
- Low scorers benefit the most by using these strategies and should therefore use and learn these explicitly.
- High scorers do not benefit as much as low scorers and the focus on cognitive learning strategies should therefore not be as strong for this group.
- Use strategies like: mnemonic strategies, pictures, synonyms and grouping
- Reading is a recommended learning strategy as high scoring pupils gain a large amount of new vocabulary with little effort

LLAMA_D

The research

The LLAMA_D is a sound recognition test that measures if you are able to recognize short stretches of spoken language that you were exposed to a short while previously (Meara, 2005). This test is graded from 0-100 where the different scores are described as follows: 0-10: a very poor score, 15-35: an average score, 40-60: a good score, 75-100: an outstandingly good score.

Research done by Speciale, Ellis and Bywater (2004) and Service and Kohonen (1995) supports the idea that students who are capable of learning pseudowords by listening to them are also capable of recognizing patterns and small variations in language. This makes these students capable of discerning important key features by listening to a target

language, without necessarily reading or writing the word in question. Individuals who rely on intuition and a more holistic approach to information processing may be better at learning complex patterns or hidden covariations in the environment implicitly (Granena, 2016).

The pedagogical frame for using the LLAMA_D results in a classroom situation is here presented by the importance of form and meaning as well as oral and written learning of vocabulary. I would suggest that pupils who gain a low score on the LLAMA_D are in need of more repetition of new words and could possibly also be in need to write/read the words they are expected to learn, making them highly dependent on written learning of vocabulary. These students rely more on explicit learning of vocabulary (Granena, 2013). Pupils who perform high on this test could be said to be less dependent on written learning of new vocabulary and would benefit more on oral communication and listening to words in order to enhance their vocabulary. High input of English oral communication would be beneficial for these pupils. These pupils would benefit more on implicit learning of English (Granena, 2013). The study of Maddah & Reiterer (2018) showed that scores in the LLAMA_D test revealed a significant, positive relationship with the subjects' English pronunciation score ($r = .66$) which proves that subjects with better short and long-term memories could achieve a higher native-like attainment in the pronunciation of a second language.

The understanding of collocations has also proven to correlate with the performance on the LLAMA_D. Learners who gain high scores also have a greater understanding of new, wrong or awkward collocations than low scorers and the understanding of the entire concept also seems more natural for the high scoring learners (Lundell & Sandgren, 2013).

Summarized functionality tips

- High scorers of the LLAMA_D do not have the same need to read and write a word in order to learn it and benefit from high input flow through communication
- Low scorers of the LLAMA_D has a greater need to write and read a word in order to learn them and can easily miss out on new words if they are only presented trough audio and communication

- Tasks like role-play, listening to texts, watching movies without texting and other communicative tasks are examples of more beneficial tasks for high scorers of the LLAMA_D
- Tasks like writing down new words, using a dictionary, reading texts, specific vocabulary learning tasks, watching a movie with English subtitles and using scripted communicative tasks are examples of more beneficial tasks for low scorers of the LLAMA_D
- Collocations are more understandable for high scorers
- High scorers will benefit from oral tasks and might easier achieve native-like proficiency in English

LLAMA_E

The research

The LLAMA_E is a sound-symbol correspondence task which test the participants in their ability to work out the relationship between sounds and the writing system presented on the screen (Meara, 2005). This test is graded from 0-100 where the different scores are described as follows: 0-15: a very poor score, 20-45: an average score, 50-65: a good score, 75-100: an outstandingly good score.

Research done by Granena (2013) shows that high scorers on the LLAMA_E test would indicate that a pupil has a strong sense of understanding when it comes to analyzing the correct pronunciation of a word, based on how it is written. Pupils that gain a low score on the LLAMA_E will typically have difficulties in pronouncing the words correctly and reading could also be challenging as the words and their structure and pronunciation do not fall naturally and intuitively to the pupil.

Meara (2005) concludes that LLAMA_E is especially good at picking out participants that are able to dissociate sounds from the way they are normally written in English. This means that these learners will be able to connect words and sounds faster and with more precision than other learners. This can be an advantage in incidental language learning and this supports the notion that a large amount of input could be beneficial for this group of learners.

Summarized functionality tips

- Input through communication is rewarding in terms of language acquisition for high scorers on the LLAMA_E
- High learners of the LLAMA_E will more easily be able to associate sound with meaning and form
- Low scoring pupils on the LLAMA_E might have difficulties with pronunciation and reading out loud
- Listening tasks or oral communication, accompanied with text might help low scorers to understand the connection between sound and words

LLAMA_F

The research

The LLAMA_F subtest is a grammatical inferencing test that asks the participant to work out the grammatical rules of an unknown language (Meara, 2005). This test is graded from 0-100 where the different scores are described as follows: 0-15: a very poor score, 20-45: an average score, 50-65: a good score, 75-100: an outstandingly good score.

Research done by Erlam (2005) shows that there is a strong connection between aptitude scores and pedagogical choices when teaching. She suggests that pupils with a high analytic ability should be taught with an inductive approach and a structured-input method. The pupils with this skill would be those who typically score high on the LLAMA_F. This means that they would benefit from being exposed to examples of the target language and then asked to figure out the rules that govern it. It is a kind of induction that lets the pupils explore the language themselves before it is structured for them. The structured-input method means that you present input that is manipulated in ways that push learners to become dependent on form and structure to get meaning (Lee & Van Patten, 2003). Activities that support the structured-input method are supplying information, matching, binary options, ordering/ranking and selective alternatives.

Pupils that score high on LLAMA_F would typically be what Scovel calls “grammarians” (as cited in Granena, 2016). These high-analysis learners will typically thrive in language learning when they are allowed to search for rules, develop rule-based representations of the language and they will always strive for accuracy (Granena, 2016). Research done by Yilmaz

(2013) has shown that high scorers on the LLAMA_F benefit more from explicit feedback (explicit correction) than they do from implicit feedback (recasts).

Summarized functionality tips

- An inductive approach to teaching will benefit high scorers on the LLAMA_F
- Structured-input methods benefit high scorers on the LLAMA_F
- Low scorers on the LLAMA_F will typically struggle to understand the connection between language and grammatical domains
- Methods for high LLAMA_F scorers can be searching for rules, selecting alternatives and matching, binary options.
- High scorers benefit more from explicit correction than from recasts

GENERAL FUNCTIONALITY TIPS

- If you have a class with a highly mixed level of language aptitude the deductive approach should be used for teaching (Erlam, 2005). That means you should explain the concepts and rules of the language firstly and then later introduce examples and relevant situations in which the previously learned skills can be used.
- LLAMA_D seems to support learners who are generally good implicit learners, whereas the other subtests seem to focus more on explicit language learning (Granena, 2013).
- Teaching speaking strategies is very important for low aptitude scorers and especially in EFL situations where they are not exposed as much to the target language
- There seem to be little or no connection between LLAMA scores and scores on oral performance tasks (Yalcin, 2012). Still, one would expect learners who gain a high score on LLAMA_D and LLAMA_E to be more precise in their oral production.
- LLAMA test scores can be used as an additional tool to mapping tools such as “kartleggeren” and would be able to give more accurate insight into the underlying language learning strengths and weaknesses of the pupils.
- Aptitude testing can be used to gain insight into the reasons for why a low-preforming pupil is not doing well in the language learning process, especially when there seems to be an issue that does not concern other common individual differences such as motivation or intelligence.
- Aptitude testing can be used to reveal learning difficulties in certain areas. E.g. if the oral and incidental tests in the LLAMA (LLAMA_D) is a high score and the other tests are very low scoring.
- Low aptitude may cause classroom anxiety