



DEPARTMENT of PSYCHOLOGY

**The Influence of Working Memory, Language Learning Experience and
Language Background on LLAMA_F Test Performance**

Maria Tomm

Master Thesis (30 hp)
PSYP01VT22
Summer 2022

Supervisors: Roger Johansson
Elia Psouni
Marianne Gullberg

Abstract

This experimental study investigates how individual variation in cognitive functioning, specifically working memory, previous experience with language learning but also language background, influence language aptitude, in particular grammatical inferencing - the capacity to spot patterns in grammar - as measured by the LLAMA_F test. Firstly, the study explores individual factors hypothesised to influence grammatical inferencing performance. Secondly, it examines whether the latest version of LLAMA_F test can be considered to be language neutral comparing agglutinative and non-agglutinative languages. In a three-part online experiment, native speakers of English, Hungarian and Finnish ($N_{total} = 72$) completed the LEAP-Q questionnaire (language background), n-back task (working memory capacity) and the LLAMA_F test (grammatical inferencing). Working memory capacity was found to correlate positively with grammatical inferencing ($r = .28, p = .018$), while language learning experience did not correlate with grammatical inferencing ($r = .141, p = .119$). After controlling for working memory, there were significant differences in grammatical inferencing between different language groups ($p = .024$); with the Hungarian group outperforming both the English and the Finnish group, and no significant differences between the English and Finnish group. This indicates that language aptitude is influenced by several different factors and that LLAMA_F cannot be regarded as entirely language neutral. Theoretical and practical implications are discussed, such as considering participants' language background when administering LLAMA_F.

Keywords: language aptitude, grammatical inferencing, LLAMA_F, working memory, language learning experience, language background, agglutinative languages

The Influence of Working Memory, Language Learning Experience and Language Background on LLAMA_F Test Performance

Communication and acquiring a first language are an integral part of the human experience. Some even argue that complex language in the way we humans use it and the complex cognitive architecture this process entails are exclusive to humans (Hauser et al., 2002). And while all humans learn some form of first language (L1) and can become proficient users of it, learning a foreign language (L2) after the age of five often does not lead to the same result (DeKeyser, 2000), as individuals vary greatly in their ability to learn foreign languages successfully. Besides general influences from positive factors such as motivation (Dörnyei, 1998) or negative factors such as anxiety (Dikmen, 2021), individuals differ significantly on so-called language aptitude. To put it simply, language aptitude describes a “talent for learning foreign languages” (Dörnyei & Skehan, 2003). Indeed, many studies have indicated language aptitude to be a strong predictor of foreign language proficiency (Li, 2016). In second language acquisition (SLA) research, definitions of language aptitude build on the ideas and work of psychologist John B. Carroll, who first defined and operationalised language aptitude. He saw it as the speed at which a student is able to learn new language material and regarded language aptitude as a relatively stable skill over time (Carroll, 1990).

To this day, there is an ongoing debate in the literature whether language aptitude is a stable skill and, perhaps more importantly, which factors it consists of. Accordingly, over the past six decades different test batteries have been developed to measure language aptitude. One of the most widely used tests is the LLAMA (Meara, 2005), which is based on the assumption that language aptitude consists of four factors, going back to the work of Carroll (1990). The present study focuses on one factor in particular - grammatical inferencing. It is the capacity to spot structural regularities in grammar and is thus an essential part of language learning. The goal of the present study is to take a closer look at the latest version of the

grammatical inferencing task LLAMA_F and to explore the extent to which individual differences in working memory, language experience and language background influence performance in this grammatical inferencing task and acquiring novel grammatical structures in general. In previous studies general language aptitude has been shown to correlate with working memory (Yilmaz, 2013) and language experience (Ma et al., 2018) and a recent study revealed that language background had an influence on grammatical inferencing too (Mikawa & De Jong, 2021). We will first take a look at theories of how grammar learning works and which general cognitive abilities support this process.

Language acquisition and learning

On average, an infant will start babbling at around six months of age (Lang et al., 2019) and utter their first words between nine to twelve months (Simonsen et al., 2014). Once a child begins to put two or more words together, grammar becomes important. At first glance acquiring our first language(s) as infants can come across as an effortless process. However, according to Tomasello (2011), there are two complex mechanisms that mainly contribute to initial language acquisition in infants: firstly, the ability to understand the intention behind an utterance or an action and secondly, the ability to find patterns in the stream of language. This second process is closely related to grammatical inferencing, as it allows the infant to acquire a sense for grammatical structures and their meanings. Understanding these structures is not only important for the infant to become a proficient speaker of their language later in life, but also a way to facilitate the acquisition of new words (Tomasello, 2011).

The process of language learning after a first language has already been established is nevertheless different, as adults who learn another language, have their first language as a starting point. Instead of solely relying on the processes an infant utilises, they can consciously draw parallels between the target language and their own first language; even though grammatical structures cannot always be directly transferred from one language to the other (Meisel, 2011). There is a debate in the literature on how the process of grammar

acquisition differs between infant L1 and adult L2 learners of the same language (Meisel, 2011). One theory with a large empirical support suggests that children rely more on implicit learning processes, such as described above, whereas adults are more likely to rely on explicit learning processes (Service et al., 2014). This is because there seems to be a critical period in which the inductive implicit kind of learning is particularly used and after which acquiring a language to a complete native level becomes increasingly unlikely. This critical period is said to be before the ages six to seven, and the theory and some empirical studies in non-naturalistic classroom settings imply that adults potentially lose these implicit abilities (Granena & Long, 2013). Instead, adults will rely more on explicit learning strategies, a natural process, as DeKeyser (2000) argues, also in part because adult cognitive structures are fully developed as opposed to children's cognitive structures. Another reason as to why the processes of L1 and L2 learning differ, is crosslinguistic influence. Meaning the way in which the experience with one particular language influences our perception and processing of another. These similarities between languages can occur between different aspects of language such as words (lexicon), word structure (morphology) and grammar (McManus, 2021). For instance, in some languages, nouns have an assigned gender (*the day, the school, the car* in English but *der Tag, die Schule, das Auto* in German). In order to learn a L2 our cognitive mechanisms, which are accustomed to L1, need to adapt to L2. One theory of how this works is that the original knowledge is transferred to the new context and only then adapted to suit the new context, if necessary (Sharwood Smith & Truscott, 2006). Another language learning approach based in cognitive psychology and general learning theories is that during the learning process knowledge is not simply copied from one context to the other (Larsen-Freeman, 2013). It is rather that new knowledge is built, while novel mechanisms are generated to select relevant information between each language (McManus, 2021).

Assessment of language aptitude

First attempts at measuring language aptitude were made with the MLAT – Modern Language Aptitude Test which was originally developed in 1959 (Carroll & Sapon). Through several factor analyses, Carroll examined several related cognitive factors, which eventually led to the conclusion that the original construct of language aptitude was made up of four separate contributing factors (Carroll, 1990): phonemic coding ability (ability to differentiate between novel sounds and their visual representation), rote learning ability (ability to create connections between written material and its meaning), inductive language learning ability (ability to spot patterns in novel grammar) and finally grammatical sensitivity (ability to find grammatical meaning of words).

The LLAMA test is in part based on the MLAT and in part on more recent literature. The LLAMA test battery is very popular due to it being freely available, reliable and most importantly claiming to be language neutral (Rogers et al., 2017). It has been used in a variety of language study contexts with language aptitude as a predictor variable, where language aptitude influenced certain aspects of language proficiency such as lexicogrammar complexity and morphological accuracy (Saito, 2017) or the process of language attrition (Bylund et al., 2010).

Like the MLAT, the LLAMA test also consists of four subtests, two of which are similar to two of the original subtests of MLAT, namely word meaning associations and grammatical inferencing. The LLAMA measures vocabulary learning ability (the capacity to relate novel objects to novel words), sound-symbol association (the capacity to relate novel sounds to known symbols), sound recognition (the capacity to differentiate between sounds) and grammatical inferencing (the capacity to spot patterns in a novel grammar).

The present research will focus only on grammatical inferencing, captured by the subtest LLAMA_F. It is a task in which participants are presented with pictures of coloured shapes in different positions and combinations. Each of those pictures has a short caption in

an artificial language. By spotting patterns in the captions and the set-up of the pictures, conclusions can be drawn about the grammar of the artificial language. After a short learning interval, pictures should be matched with the suitable caption. In earlier versions of this task, there were direct translations of the artificial language captions into real language, but to make the test available to more people with diverse language backgrounds, these were eventually excluded. In a more recent version, captions in the testing phase were complete and participants did not have to create the sentences themselves, which is the case with the latest version LLAMA_3. This is more of a challenge but lowers the chance of participants simply guessing the correct captions. The last update to LLAMA_3 was in May of 2021. As the LLAMA keeps getting updated, the latest version of LLAMA_F has not yet been explored as thoroughly as its predecessors. Thus, the present study focused on three important factors of the cognitive architecture involved in the ability to make grammatical inferences, examining the ways in which these factors may be implicated in performance in the current grammatical inferencing task included in LLAMA_F.

Potential influences on grammatical inferencing of LLAMA_F3

Working Memory

A major feature of the cognitive architecture of grammatical inferencing is working memory, which, just like language aptitude, varies greatly between individuals. Working memory as originally defined by Baddeley and Hitch (1974; Baddeley 1986) took a less static approach than previous memory models at the time. Working memory manipulates information temporarily instead of simply storing it and can therefore be seen as a connection between long-term memory, perception and interacting with the environment. It is thus essential to complex learning and reasoning processes (Baddeley, 2003). The model includes the central executive (executive working memory), which is controlling three specific sub-components; the visual sketchpad, the phonological loop and the episodic buffer (Baddeley & Hitch, 1974). Based on this model of working memory capacity, there have been

a number of studies that argued that there are connections between these specific working memory capacity components and language aptitude factors. For instance between the phonological loop and vocabulary learning (Service, 1992) and indeed, also between general executive working memory and grammatical inferencing (Yilmaz, 2013).

While Baddeley's model of working memory capacity still has a large influence in literature to this day, there are some more novel understandings on working memory that do not define it as a strictly hierarchical model (Oberauer, 2009). More recent accounts on working memory capacity suggest that it is a dual-process model, which consists of a declarative and a procedural part, where both contain structures in which new information is integrated, as well as including attention focus capacities and active long-term memory (Oberauer, 2009).

What both working memory models have in common however, is that they try to describe how new information is processed and manipulated. These complex cognitive abilities are also needed when performing typical grammatical inferencing tasks (Martin & Ellis, 2012), for example when memorising novel items and processing their relation to another. Indeed, a recent meta-analysis of 66 language aptitude studies showed that executive working memory was positively related to language aptitude (Li, 2016). Especially performance in grammatical inferencing tasks that are similar to the LLAMA_F test described above has been shown to positively correlate with higher levels of working memory (Sáfár & Kormos, 2008). This correlation might even be stronger with the new version of LLAMA_F3, as it does not involve simple word recognition as much and requires participants to actively create a sentence.

The centrality of working memory for language aptitude is underlined in views that working memory ought to be considered as an integral component of language aptitude rather than a separate concept (Friedman & Miyake, 1998).

Language Experience

A second important feature of the cognitive architecture of grammatical inferencing is that of language experience. Operationalisation of language experience varies in the literature. One approach considers the time an individual has spent with a language, counting from the age when someone began learning a second language (Stafford et al., 2010). According to this approach, the longer the exposure to this language is, the higher the language experience will be. Another approach considers the total number of languages known. Indeed Nayak et al. (1990) explored how monolinguals and multilinguals compare in their ability to detect grammatical rules for the word order patterns of sentences, the syntax, when asked to do so. Multilinguals outperformed their monolingual counterparts, which the authors traced back to their higher language experience. This effect is not unique between monolinguals and multilinguals. A recent study compared language aptitudes of language learners of one L2 with those of language learners with a second L2 and found that learners of more than one L2 scored higher in language aptitude as measured by the LLAMA tests (Ma et al., 2018). Yet another approach defines language experience as the experience with formal language classroom instruction (Lado et al., 2017). According to this approach, a person who has acquired a second language (L2) through formal studies later in life has more language experience than a person who has been raised bilingually.

There are some indications that higher language experience positively influences performance in the grammatical inferencing task of LLAMA_F; possibly because individuals with higher levels of language experience are more meta-linguistically aware and have developed strategies to learn and detect grammar (Rogers et al., 2017). It is not exactly clear, however, how much experience in a L2 is needed to create this effect, for example if the effect grows with the number of languages learned. While there are some well documented cases of polyglots (individuals who use several languages fluently), who excel in language aptitude tests as compared to people who know fewer languages (Hyltenstam, 2016), in terms of

number of languages known, these individuals are the exception rather than the norm. Rogers et al. (2017) found that knowing more than one foreign language already made a significant positive difference to language aptitude, in particular to grammatical inferencing. Another important point to consider is the level of those languages. Lado et al. (2017) found that an intermediate level of Spanish helped participants learn Latin grammar in the short-term, but that only those who had higher, near fluent levels of Spanish were able to retain the Latin grammar rules in the long-term. This implies that some established knowledge is necessary to create this effect and that knowing only the very basics of a language would not suffice.

Language Background

A final important feature of the cognitive architecture of the ability for grammatical inferencing is the nature of one's first language (L1) background. Mikawa and De Jong (2021) argue that the artificial language used in LLAMA_F seems to mainly use a grammatical structure of a particular type of languages, namely "agglutinative languages". Examples for agglutinative languages are Finnish, Hungarian, Estonian, Turkish and Japanese to name but a few. What all these languages have in common is that several morphemes, smallest word units that carry meaning, can be used to form rather long words. And while doing so, these morphemes do not change, as they do in other, "non-agglutinative", languages. For example, in Hungarian (*boldog* = happy; *boldog-abb* = happier; *leg-boldog-abb* = happiest; *leg-bolodog-abb-nak* = the happiest one; *leg-boldog-abb-ak-nak* = the happiest ones). Where non-agglutinative languages such as English would use a sentence, agglutinative languages can sometimes use only one word; for example, in Finnish (*talo* = house; *talo-ssa* = in house; *talo-ssa-ni* = in my house). The artificial language used by LLAMA_F works in a very similar way, as explained in more detail in methods.

In a recent study, (Mikawa & De Jong, 2021) found that, indeed, the subtest concerning grammatical inferencing, LLAMA_F, might depend on the participant's first language background. This would not be desirable, as that would potentially give native

speakers of certain languages an advantage in solving the grammatical inferencing task and make results of different native speaker groups less comparable. This is especially important if a study uses LLAMA as an instrument to measure language aptitude. The study by Mikawa and De Jong (2021) compared language aptitudes of agglutinative language native speakers (i.e., Japanese, and Hungarian) with non-agglutinative language native speakers (Dutch) and found that Japanese L1 speakers outperformed the other language groups in LLAMA_F performance. This can partly be explained by the nature of the grammatical inferencing task, as the artificial language used here is grammatically more similar to agglutinative languages. However, it does not explain why Japanese L1 speakers performed better than Hungarian L1 speakers.

Aims of this study

Taking into account previous and recent findings on factors in the cognitive architecture of the ability of grammatical inferencing, the aim of this study is to assess the impact of these cognitive, crosslinguistic and individual factors for individual performance on the LLAMA_F. Firstly, the study assesses and compares the extent to which working memory, language experience (here defined as the level and number of languages learned), and L1 background impact the performance in LLAMA_F3.

Secondly, the study explores whether the latest version of the grammatical inferencing test LLAMA_F is in fact language neutral. By including participants with varying native language background (English, Hungarian, Finnish and Japanese language groups), we explore whether Mikawa and De Jong (2021) results are replicated with the new version LLAMA_F3 and whether this pattern continues in another agglutinative language, Finnish. Exploring this would help determine whether language background does in fact have an impact on grammatical inferencing. The following hypotheses will be tested:

1. Higher performance in the working memory task correlates positively with grammatical inferencing.

2. Higher language experience correlates positively with grammatical inferencing.
3. Agglutinative language speakers outperform non-agglutinative language speakers in the LLAMA_F grammatical inferencing task.

Methods

Participants

Adult native speaking participants were recruited for three different language groups – speakers of two agglutinative languages, Hungarian ($N = 30$) and Finnish ($N = 16$), and speakers of English ($N = 26$) as a non-agglutinative control group. Participants in the English-speaking control group had not learned an agglutinative language as a L2 previously. There were no other participant-specific exclusion criteria. To make the samples comparable, we also assessed gender and educational background. See Table 1 for demographic information.

Despite a high initial participation rate, a total of 233 participants needed to be excluded from the analysis eventually. This was largely due to incomplete datasets, as not all participants finished the experiment, and the lack of an outcome variable score made their data unusable. Only participants with a completed LEAP-Q, LLAMA_F and n-back task over at least 50% were included. Further six native English speakers were excluded as they reported an agglutinative language as their L2. Furthermore, there were only four complete datasets from native Japanese speakers, which was not sufficient for analysis, so the entire language group was omitted from further analysis and discussion.

Table 1*Demographic Information*

	Hungarian	Finnish	English
<i>N</i> (male/female/other)	30 (17/13)	16 (9/6/1)	26 (18/7/1)
<i>Age</i> (<i>M</i>, <i>SD</i>)	28 (7)	33 (9)	28 (12)
<i>N</i>_{Academic Degree} (%)^a	24 (80%)	12 (75%)	22 (84,61%)
<i>N</i>_{Foreign Languages} (Range)	2,5 (1 – 4)	2,62 (1 – 4)	1,23 (0 – 3)
Other Languages	English, German, French, Spanish, Swedish, Italian, Russian, Danish, Romanian, Latin, Hebrew	English, Swedish, German, Spanish, French, Italian, Dutch, Japanese	French, Spanish, German, Irish, Welsh, Portuguese, Italian, Russian, Latin, Korean, Hindi, Gujarati, Swedish, Dutch, Norwegian, Tagalog

Note. Other languages are reported in order of how often they occurred in the sample.

^aEducational background of participants. Graduate school, college, master's degrees, or a PhD counted as an academic degree.

Instruments

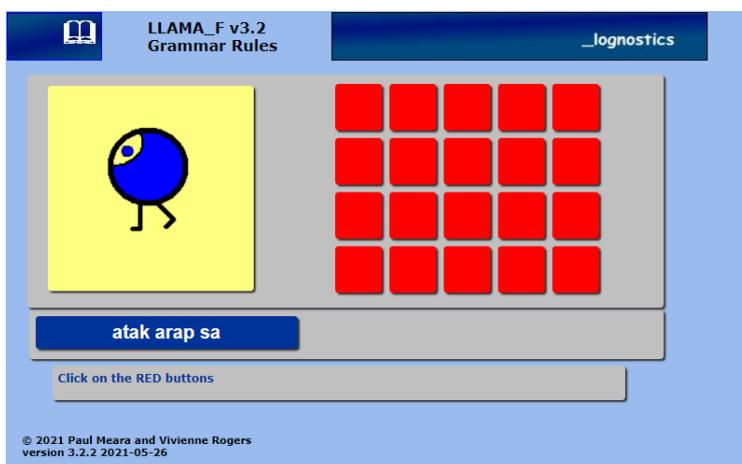
To assess participants' language background, parts of the LEAP-Q – the Language Experience and Proficiency Questionnaire (Marian et al., 2007) were used. The LEAP-Q includes detailed self-reported information about the experience and proficiency of each language a participant speaks, as well as educational background and age. It is a reliable instrument that has been translated into over 20 languages, is freely available and widely used (Kaushanskaya et al., 2020). As not all questions from this instrument were relevant for this

research, we only used the items with the strongest predictive power of language ability. They were self-estimation of the age a participant began acquiring/became fluent in a language, when they began reading/became fluent in reading and how much time they spent in each language environment; and finally, estimates of their proficiency in speaking and reading in a language on an eleven-point Likert scale from none to perfect (0 - 10). Cronbach's alpha is rather high for the self-reported proficiency factor (L1 $\alpha = .92$, L2 $\alpha = .88$), which also has a high external validity (Marian et al., 2007).

The language aptitude test battery LLAMA (Meara, 2005) was used, specifically the subtest LLAMA_F, to assess grammatical inferencing. The latest version (LLAMA_F3) is free to use on the official LLAMA website (Meara & Rogers, 2019). In the learning phase of the test participants initially saw twenty red buttons and a yellow square. Whenever they clicked one of the buttons, a picture was displayed in the yellow square with a corresponding sentence in the artificial language under it. The artificial language was presented in Latin letters. See Figure 1 for a screenshot of the LLAMA_F3 learning screen. Participants were allowed to take notes if they wished.

Figure 1

Example of Learning Phase Screen of LLAMA_F

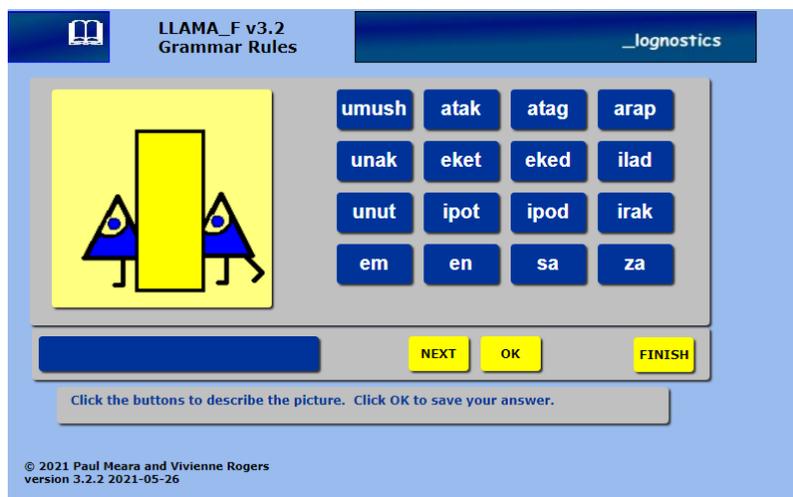


Note. The sentence “atak arap sa” in the artificial language “Patsi” refers to the picture displayed above. “Atak” stands for blue, “arap” for round and “sa” indicates a singular object.

The picture-sentence combinations could be explored until the program closed the learning phase after five minutes and moved on to the testing phase. Here, the yellow square was displayed again as well as a new set of sixteen blue buttons, containing words/fragments from the artificial language. The same twenty pictures that had been seen previously were displayed one by one in the yellow square. Participants were required to select the appropriate word combinations by clicking the blue buttons. There was no time limit in this test. A maximum of twenty points could be achieved, which was displayed to participants after finishing the test. See Figure 2 for a screenshot of the LLAMA_F3 testing screen. LLAMA_F has been shown to have an internal consistency of $\alpha = .60$ and a test-retest reliability of $r = .56$ (Granena, 2013).

Figure 2

Example of testing Phase Screen of LLAMA_F



Note. Participants should select the correct sentence and then click the “OK” button.

To test participants' working memory capacity, we used a visual n-back task.

Originally developed by Kirchner (1958), it is a task that predicts working memory processes (Engle & Kane, 2003). It requires the participant to follow an on-screen succession of shapes and to remember their order, as they will be asked to report if the current shape matches the one they saw n-numbers back. For the purpose of this study, this version of an n-back task is

particularly helpful, as it only uses shapes as opposed to letters and is therefore relatively language neutral apart from the instructions. The n-back task is a widely used instrument to measure working memory capacity, since it requires the participant to store and process new information at the same time. It has been criticised for its poor construct validity, as it only weakly correlates with other working memory measures such as the reading span task (Jaeggi et al., 2010). However, there are some studies that report good reliability scores for visual 2-back tasks ($r = .91$) (Friedman et al., 2006), especially test-retest reliability ($r = .82$) (Soveri et al., 2018). Furthermore, the language-neutral nature of the task as well as the compatibility with an online experiment made the 2-back task an appropriate choice for this study. The task was presented using the online platform Pavlovia, while the n-back task was created in PsychoPy3. It used seven different stimuli (see Figure 3 for an example) and 150 total trials out of which there were forty-five possible hits. Each stimulus was presented for the duration of 1 second. This task was a 2-back task.

Figure 3

Example of stimuli in the 2-back task



Note. Four of the seven possible items are displayed. This is not a screenshot, as items appeared one at a time.

Procedure

A computer-based experiment was carried out online for practical reasons such as accessibility of participants with different L1 backgrounds. To recruit participants, the study was advertised online through various social media platforms and through word of mouth. First, participants filled in a short questionnaire about their language backgrounds using the platform Qualtrics and they were assigned a participant number. At the end of the

questionnaire, two links were provided that each opened another window. These were for two further tests, one for grammatical inferencing and one for working memory. All participants took part on a device with a physical keyboard, as this was a requirement for the working memory task. Participants were asked to enter their participant number for each of the two experiments, so that results could be matched anonymously. On average, the entire experiment took 25 - 30 minutes in total.

Ethics

Before taking part in the study participants gave their informed consent. They were aware that participation was voluntary and could be withdrawn at any time without justification and that they were not identifiable from their data. Name and contact details of the researchers were provided if participants had further questions or wanted to withdraw their participation. There was no physical or mental risk in participating in the study. Only adults above the age of eighteen were able to take part in the study, in line with Swedish ethical research guidelines.

Data analysis

Data was analysed using IBM SPSS Statistics (Version 21). All variables were operationalised in different ways. For working memory *d-prime* was used, the z-transformed PR-Score (Hit rate – False alarm rate) of the n-back task. Hit rate was the total number of correct responses, when a stimulus was correctly identified as repeated, while false alarm rate stood for the total number of wrong responses, when a stimulus was falsely identified as a repeated stimulus. This technique was used in order to get a more accurate estimate of working memory capacity than simply calculating the number of correct responses in proportion to the total number of stimuli (Snodgrass & Corwin, 1988).

Language experience was operationalised in two ways. Firstly, as the total number of L2 reported and secondly as the number of L2 spoken above or at “adequate” level, ranked from zero to four. “Adequate” level was number five on the self-report proficiency scale from

zero to ten (none - perfect). For both these variables separate correlations were calculated with the LLAMA score to address hypothesis one and two. For hypothesis one a one-tailed Pearson correlation was used and for hypothesis two, two one-tailed Spearman correlations.

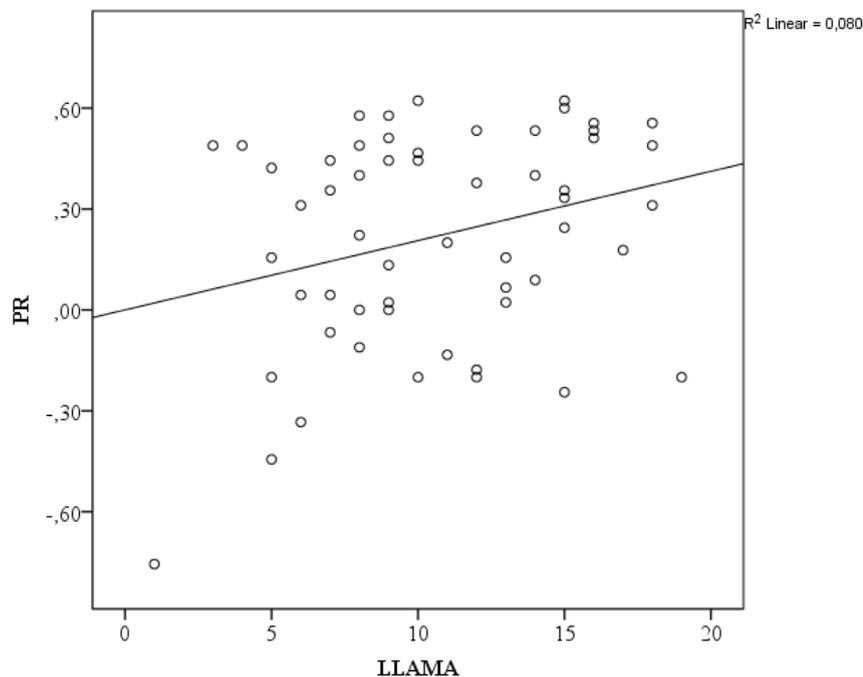
Language background was operationalised as the reported L1 of participants: English, Finnish or Hungarian. For hypothesis three, firstly, an ANOVA was conducted with the LLAMA scores (0 - 20) of three language groups to see if grammatical inferencing differed between groups. Afterwards, an ANCOVA with language group (independent variable) grammatical inferencing (dependent variable) and working memory (covariate) was introduced to further explore the difference between the language groups English, Finnish and Hungarian. All assumptions for ANCOVA were met. Homogeneity of regression slopes was given for the dependent variable as there was no significant interaction ($p = .941$). A Shapiro-Wilk test confirmed that residuals were normally distributed ($p = .740$). And finally, error variance was equal across language groups, as a Levene test revealed ($p = .931$).

Results

To test hypothesis one, a one-way Pearson correlation was conducted. There were twelve missing data points and an additional four outliers in the working memory variable. All four outliers were excluded from analysis, as they had a *d-prime-score* of < -1.0 , which indicates that their answers were due to chance. The analysis showed a significant positive correlation between LLAMA_F scores and *d-prime*, $r(54) = .28$, $p = .018$. Performance in the working memory task correlated positively with grammatical inferencing; thus, hypothesis one could be confirmed. See Figure 4 for a scatterplot of the Pearson correlation. On the contrary, the two one-way Spearman correlations for the second hypothesis showed that there were no significant correlations between LLAMA_F and the total number of languages $r(70) = .16$, $p = .085$ nor between LLAMA_F and the number of languages above adequate level $r(70) = .14$, $p = .119$. Hypothesis two could therefore not be confirmed.

Figure 4

Correlation of PR-Score and LLAMA_F Score



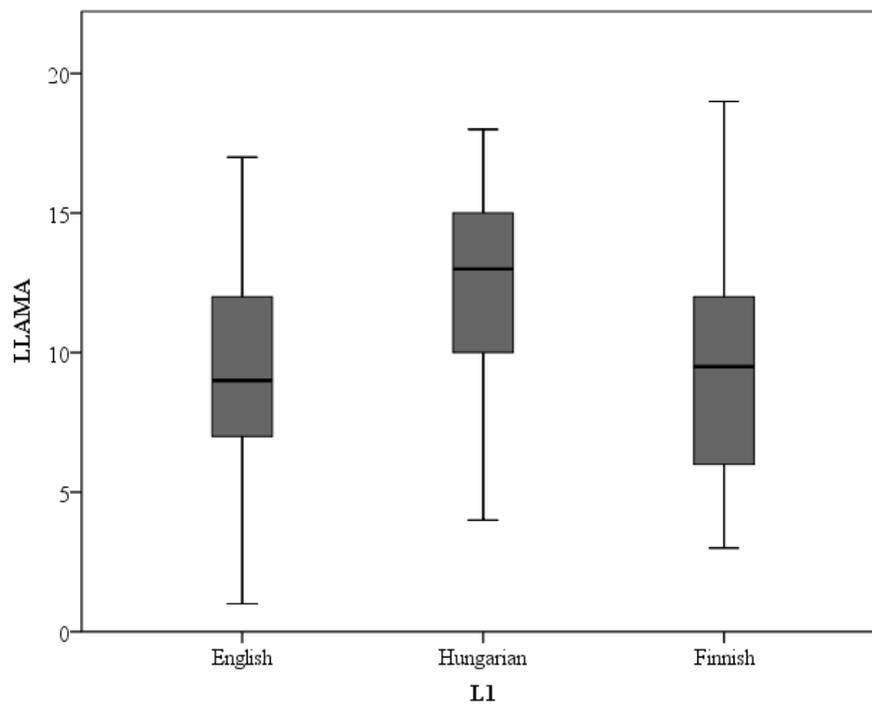
Note. PR-Score measuring working memory and LLAMA_F-Score measuring grammatical inferencing. Regression line is included. The four outliers are already excluded here.

For hypothesis three, to find out whether language background influenced performance in the grammatical inferencing task, while controlling for working memory, an ANCOVA was conducted. Before introducing the covariate working memory, LLAMA_F-Score means were compared between language groups in an ANOVA. There, the lowest mean score on LLAMA_F of the language groups was found in the English group ($M = 8.85$; $SD = 3.47$), followed by Finnish ($M = 9.56$; $SD = 4.29$) and Hungarian with the highest mean ($M = 12.46$; $SD = 4.31$). See Figure 5 for a boxplot. After introducing working memory as a covariate, the English group still had the lowest mean ($M = 9.08$; $SE = 0.90$) followed by Finnish ($M = 9.56$; $SE = 0.99$) and Hungarian ($M = 12.27$; $SE = 0.82$). The scores still differed significantly on LLAMA_F between the groups ($F(2,56) = 3.97$, $p = .024$, partial $\eta^2 = .124$). The Bonferroni post-hoc test revealed a significant difference between the English and Hungarian language group ($p = .012$, $M_{\text{Diff}} = 3.20$, 95% CI [0.73, 5.66]), as well as the Finnish and Hungarian

language group ($p = .040$, $M_{\text{Diff}} = 2.72$, 95% CI [0.14, 5.30]). There was no statistically significant difference between the English and Finnish language group ($p = .723$; $M_{\text{Diff}} = .478$, 95% CI [-2.21, 3.17]). Therefore, hypothesis three could partly be confirmed. See Figure 6 for a bar chart.

Figure 5

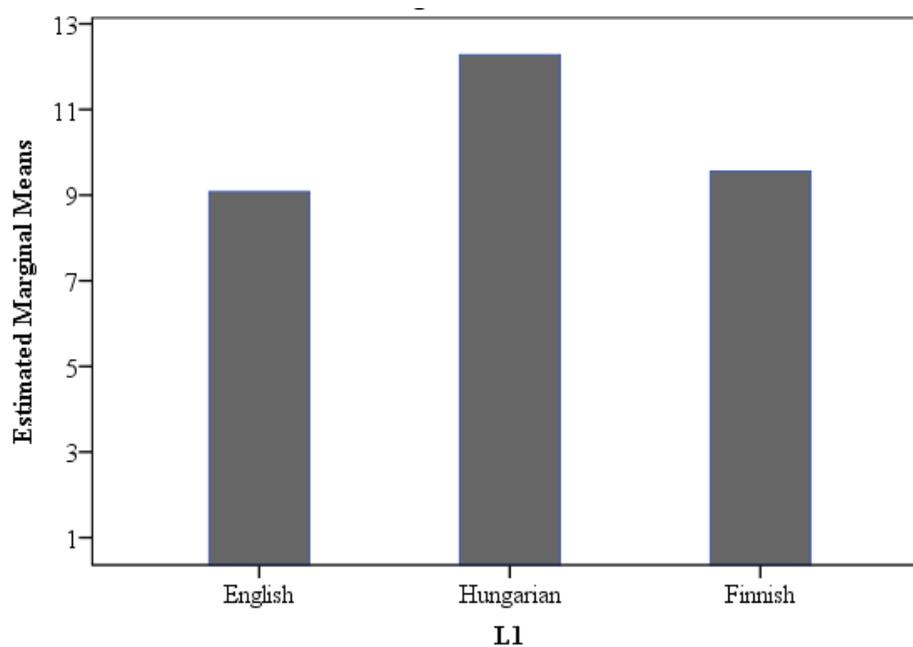
Boxplot for LLAMA_F score means according to language group



Note. Language group differences in ANOVA, before working memory was included as a covariate.

Figure 6

Bar chart for LLAMA_F estimated marginal means in ANCOVA



Note. Language group differences in ANCOVA, after working memory was included as a covariate.

Discussion

The aim of the present study was to assess the extent to which grammatical inferencing, as measured by the LLAMA_F test, is dependent on major factors of the cognitive architecture involved in this ability, namely working memory and previous language experience. Further, we aimed to assess the extent to which this performance may also be dependent on L1 background. To our knowledge, this is the first study to scrutinise these cognitive, individual and crosslinguistic factors together, which explain individual variations in performance in grammatical inferencing. Overall, the present results are in line with previous studies highlighting the centrality of working memory in grammatical inferencing (Sáfár & Kormos, 2008) and the influence of L1 background (Mikawa & De Jong, 2021). They also underline the potential of language experience being an important individual factor to take into account when measuring grammatical inferencing.

The centrality of working memory

The first hypothesis, that higher performance in the working memory task would correlate positively with grammatical inferencing could be confirmed, in line with previous findings on the links between working memory and grammatical inferencing (Sáfár & Kormos, 2008).

While a correlation does not necessarily mean that there is a causal relationship between variables, considering the previous findings on working memory and language aptitude (Sáfár & Kormos, 2008; Yalçın et al., 2016), causality can be inferred.

Furthermore, as working memory in the present study was measured by a non-verbal working memory capacity task, the n-back task, working memory here can be taken to capture a broader concept than language aptitude, as there was no language required to solve the n-back task. Whether working memory capacity is in this context practically inseparable from language aptitude, as some authors suggest (Friedman & Miyake, 1998), is still a matter of debate, but the present results would suggest that it is not. Assuming working memory were the sole factor in the cognitive architecture for language aptitude, the overlap would have been much higher and thus the correlation would have been even stronger. Since that was not the case, it seems more reasonable that working memory capacity contributes to language aptitude in terms of grammatical inferencing but is not the sole cognitive component involved. Additional support that working memory capacity is a different concept comes from studies (Yoshimura, 2001; Granena, 2013), that have shown that working memory and language aptitude correlate as overall concepts, but when divided into its subfactors, only word association and grammatical inferencing can be predicted by working memory performance. A principal component analysis showed that language aptitude subcomponents loaded on different factors as working memory, which also indicates that working memory capacity is a separate concept (Granena, 2013).

Thus, our findings overall confirm the centrality of working memory for language aptitude and highlight that this may be in part due to a sharpened ability of grammatical inferencing.

The role of language experience

Higher language experience did not correlate with grammatical inferencing; neither as the total number of foreign languages reported nor as the number of languages spoken above a self-reported adequate level. However, even though there was no significant statistical support, the relationship of the two variables was still positive, which is in line with previous research that has found a positive connection between language aptitude and language experience (Nayak et al., 1990; Rogers et al., 2017; Ma et al., 2018). Thus, rather than assuming that language experience does not influence grammatical inferencing performance, we suspect that a larger sample size, with a higher power, would likely have returned significant results. Another reason for the null result in the present study could be that only ten participants in the current sample reported not knowing any language at or above adequate level, and out of which eight were monolinguals. Moreover, results might have been different had language experience explicitly been operationalised as the exposure to formal language instruction. For instance, a longitudinal study found that there were significant effects of formal foreign language classroom instruction on the overall language aptitude as measured by the LLAMA tests over the course of an academic year, where instruction was monitored for the duration of the study (Sáfár & Kormos, 2008). This was not possible in our case, so we assumed that people who acquired a language after the age of five had had formal language instruction. However, we do not know for certain how participants gained their L2 knowledge, with much or little formal instruction, which might be confounding the current data. As language experience did not correlate with the performance of grammatical inferencing, it was not included as a covariate in the ANCOVA that was conducted for hypothesis three.

Does language background confound grammatical inferencing performance as measured by LLAMA_F?

Hypothesis three, that agglutinative language speakers would outperform non-agglutinative language speakers, could partly be confirmed. Interestingly, while the Hungarian native speakers did score significantly higher than the English native speakers, the former also scored higher than Finnish native speakers and there was no significant difference between English and Finnish native speaker scores. As the group differences persisted after controlling for working memory, we conclude that these differences are due to the language groups and not due to working memory as a confounding variable. This is also in line with skill acquisition theory, in which working memory, and other individual cognitive abilities, are less important than previous domain knowledge in adults when learning new skills (Ackerman, 2007). The domain knowledge in this case is the grammatical closeness of the artificial language from LLAMA_F to Hungarian rather than to English, which would have given the Hungarian language group an advantage. That the Hungarian language group outperformed the English language group was hypothesised and could be explained by the nature of these languages, agglutinative vs. non-agglutinative.

When looking at all three languages in this study, they form an interesting pattern, since it partly replicates the findings of Mikawa and De Jong (2021), where agglutinative language speakers (Hungarian and Japanese) also outperformed non-agglutinative language speakers (Dutch). The pattern could not be fully replicated due to the lack of Japanese speakers in the current sample, however Mikawa and De Jong (2021) found that one agglutinative language group (Japanese) outperformed the other agglutinative language group (Hungarian), which is similar to the current finding. That Hungarian and Finnish native speakers' performance differed significantly is a somewhat surprising result, which could be explained by taking a closer look at the grammar of both languages. While both Hungarian and Finnish belong to the same language family, Uralic languages, and are both agglutinative languages, there are

some significant grammatical differences between the languages. Finnish differs from other agglutinative languages for instance, through its attributive adjectives. These are adjectives that are directly attached to a noun such as “the happy child” instead of “the child is happy”. In Finnish the endings of attributive adjectives change together with the noun they are attached to, but, with very few exceptions, attributive adjectives do not change with the noun in Hungarian (Anhava, 2010). In the artificial language LLAMA_F, using the words that describe colour could be seen as attributive adjectives, as they always describe one of the figures (triangle, circle or square). Their endings change according to the noun they are describing (*ipot arap* = red circle; *ipod ilad* = red square), while the nouns themselves do not change endings. Based on the current results it is not possible to say whether differences between the Hungarian and Finnish group can be traced back to this, therefore it would be interesting to analyse the mistakes each group made and to see if there are patterns that could explain the differing performance between language groups.

Even though these results seem rather ambiguous, it is likely that even this new version of LLAMA_F is not fully language neutral when it comes to agglutinative languages.

Considering the results of this study as well as the similar results of Mikawa and De Jong (2021), we conclude that the new LLAMA_F test should be administered with caution when testing participants with different language backgrounds.

Taken together, the results of the present study demonstrate that the complex cognitive architecture of language aptitude and in particular grammatical inferencing include the cognitive factor working memory as well as the crosslinguistic factor L1 background. They indicate that higher working memory capacity and a particular L1 background in certain contexts can be of an advantage when it comes to learning new grammar. Interestingly, previous studies have shown that working memory capacity (Sawyer & Ranta, 2001) and crosslinguistic similarities (Bokander, 2020) are the most beneficial in the early processes of

language learning. This makes sense when considering that in order to learn a L2, cognitive mechanisms, which are accustomed to L1 need to adapt to L2 (McManus, 2021).

Limitations and suggestions for future research

This study was conducted without funding and therefore no monetary compensation for the participants. Non-monetary compensation in the form of evaluative feedback was not possible either, because of the experiment's set-up on three different platforms. As the experiment was divided into three parts, it required participants to follow two different links and enter their participant codes there, which most participants were possibly not accustomed to. Because of these reasons the drop-out rate was very high and affected the final sample size. It would have been desirable to have a higher number of participants, more equal group sizes as well as more Japanese speakers, so that a comparison of all four language groups would have been possible as originally planned. This would have painted a richer picture of the factors and their relationship with grammatical inferencing. Furthermore, these results should only be generalised to different populations with much caution, as a majority of participants were rather young and well-educated. There were also more male than female participants, particularly in the English group. Gender is an additional individual factor that future studies could be taking into account when exploring grammatical inferencing. Although some studies have shown some relation of grammatical inferencing and gender, it is not clear exactly how gender plays a role, as there is some conflicting evidence as to the direction of the relation: Wucherer et al. (2018) found that female participants outperformed male participants in the MLAT IV, measuring grammatical sensitivity (Wucherer & Reiterer, 2018); while there are also counter indications, where male participants outperformed female participants as measured by the LLAMA_F (Rizvanović, 2018). It would be highly valuable to gain a perspective on this factor, as it helps interpret studies such as this with a gender uneven sample.

As previously alluded to, language experience could have been operationalised more explicitly as formal language instruction. Future studies would benefit from exploring this more closely as the time spent on studying the language in formal settings or perhaps a self-estimation on how much time students spend on studying grammar in particular.

It should be noted that the n-back task was the last part in the online experiment and participants motivation and concentration were not at peak performance as indicated by the very high drop-out rate especially around the n-back task. Working memory capacity was assessed using a single n-back task, which does give a good indication of working memory capacity but does not replace more elaborate testing with a combination of other working memory capacity measures, such as the reading span task (Daneman & Carpenter, 1980), which is reported to have high test-retest reliability scores of $r = .93 - .95$ and good construct validity (Waters & Caplan, 2003). Using different working memory capacity measures in future research could also be insightful to understand the interplay of working memory and language aptitude more thoroughly. For instance, how different aspects of working memory such as focus attention or long-term memory access (Oberauer, 2009) relate to language aptitude.

Data analysis of the present study partly used correlations to answer hypotheses and as alluded to earlier, significant correlations cannot be regarded as causal relationships between variables. However, this approach is not unique in the field of language aptitude research (Granena & Long, 2013; Yilmaz, 2013). Correlational results give a valuable first insight into the relation of working memory and language aptitude, on which further research can build.

In this study, in order to access the LLAMA_F data generated by our participants, the authors of the LLAMA_F needed to be contacted and direct access to the data sets was never possible. Because of this, it was never possible to see how long participants took to complete the LLAMA_F, as there is no time limit, and to see what kind of mistakes they made to explore if there were certain patterns for different language groups. This process data could

have potentially given some indication as to why language groups differed in performance. For future studies, it would thus be valuable to integrate the individual completion time of LLAMA_F and the kind of errors in the analysis to explore if there are certain patterns for different language groups.

Conclusions

Firstly, this study has shown that our ability to learn foreign languages, particularly novel grammar, is not only influenced by language aptitude but also by cognitive factors such as working memory capacity and other individual differences such as L1 background. These factors should be taken into consideration when investigating language aptitude in the future. On a larger scale, these results serve as an indication that language aptitude as a measurable concept should potentially be reconsidered and broadened in the future.

Secondly, the inconsistencies in LLAMA_F test performance between different language groups that were first reported by Mikawa and De Jong (2021) still persisted in the new version. Therefore, even the latest version of the LLAMA_F test should be used with caution, especially when comparing different language groups and when grouping a sample of participants with mixed language backgrounds.

References

- Ackerman, P. L. (2007). New Developments in Understanding Skilled Performance. *Current Directions in Psychological Science*, 16(5), 235–239.
<https://doi.org/10.1111/j.1467-8721.2007.00511.x>
- Anhava, J. (2010). Criteria for case forms in Finnish and Hungarian grammars. In K. Karttunen (Ed.), *Anantam Śāstram. Indological and Linguistic Studies in Honour of Bertil Tikkanen* (pp. 239–244). Helsinki: Finnish Oriental Society.
- Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press.

- Baddeley, A. D. & G. J. Hitch (1974). Working memory. In G. A. Bower (Eds.), *Recent Advances in Learning and Motivation Vol. 8* (pp. 47–90). New York: Academic Press.
- Ackerman, P. L. (2007). New Developments in Understanding Skilled Performance. *Current Directions in Psychological Science*, *16*(5), 235–239.
<https://doi.org/10.1111/j.1467-8721.2007.00511.x>
- Baddeley, A. (2003). Working Memory: Looking Back and Looking Forward. *Nature Reviews*, *4*, 829–839. <https://doi.org/10.1038/nrn1201>
- Bokander, L. (2020). Language Aptitude and Crosslinguistic Influence in Initial L2 Learning. *Journal of the European Second Language Association*, *4*(1), 35.
<https://doi.org/10.22599/jesla.69>
- Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2010). The Role of Language Aptitude in First Language Attrition: The Case of Pre-pubescent Attriters. *Applied Linguistics*, *31*(3), 443–464. <https://doi.org/10.1093/applin/amp059>
- Carroll, J. B., & Sapon, S. (1959). *The modern languages aptitude test*. San Antonio, TX: Psychological Measurement.
- Carroll, J. B. (1990) Cognitive Abilities in Foreign Language Aptitude. In T.S. Parry, C.W. Stansfield (Eds.) *Language Aptitude reconsidered*. Englewood Cliffs: New Jersey.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466.
[https://doi.org/10.1016/S0022-5371\(80\)90312-6](https://doi.org/10.1016/S0022-5371(80)90312-6)
- DeKeyser, R. M. (2000). The Robustness of Critical Period Effects in Second Language Acquisition. *Studies in Second Language Acquisition*, *22*(4), 499–533.
<https://doi.org/10.1017/S0272263100004022>

- Dikmen, M. (2021). EFL Learners' Foreign Language Learning Anxiety and Language Performance: A Meta-Analysis Study. *International Journal of Contemporary Educational Research*. <https://doi.org/10.33200/ijcer.908048>
- Dörnyei, Z. (1998). *Motivation in second and foreign language learning*. 21. <https://doi.org/10.1017/S026144480001315X>
- Dörnyei, Z., & Skehan, P. (2003). Individual Differences in Second Language Learning. In C. J. Doughty & M. H. Long (Ed.), *The Handbook of Second Language Acquisition* (p. 589–630). Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470756492.ch18>
- Engle, R. W., & Kane, M. J. (2003). Executive Attention, Working Memory Capacity, And A Two-Factor Theory Of Cognitive Control. *Working Memory Capacity*, 44, 55. [https://doi.org/10.1016/S0079-7421\(03\)44005-X](https://doi.org/10.1016/S0079-7421(03)44005-X).
- Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not All Executive Functions Are Related to Intelligence. *Psychological Science*, 17(2), 172–179. <https://doi.org/10.1111/j.1467-9280.2006.01681.x>
- Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test*. In G. Granena & M. Long (Hrsg.), *Language Learning & Language Teaching* (Bd. 35, S. 105–130). John Benjamins Publishing Company. <https://doi.org/10.1075/lllt.35.04gra>
- Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, 29(3), 311–343. <https://doi.org/10.1177/0267658312461497>
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298(5598), 1569–1579. <https://doi.org/10.1126/science.298.5598.1569>
- Hyltenstam, K. (2016). *Advanced Proficiency and Exceptional Ability in Second Languages*. De Gruyter. <https://doi.org/10.1515/9781614515173>

Jaeggi, S. M., Buschkuhl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the *N*-back task as a working memory measure. *Memory*, *18*(4), 394–412.

<https://doi.org/10.1080/09658211003702171>

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism: Language and Cognition*, *23*(5), 945–950. <https://doi.org/10.1017/S1366728919000038>

Kirchner, W. K. (1958). Age Differences In Short-Term Retention Of Rapidly Changing Information. *Journal of Experimental Psychology*, *55*, 7.

Lado, B., Bowden, H. W., Stafford, C., & Sanz, C. (2017). *Two Birds, One Stone, or How Learning a Foreign Language Makes You a Better Language Learner*. 19.

Lang, S., Bartl-Pokorny, K. D., Pokorny, F. B., Garrido, D., Mani, N., Fox-Boyer, A. V., Zhang, D., & Marschik, P. B. (2019). Canonical Babbling: A Marker for Earlier Identification of Late Detected Developmental Disorders? *Current Developmental Disorders Reports*, *6*(3), 111–118. <https://doi.org/10.1007/s40474-019-00166-w>

Larsen-Freeman, D. (2013). Transfer of Learning Transformed: Transfer Transformed. *Language Learning*, *63*, 107–129. <https://doi.org/10.1111/j.1467-9922.2012.00740.x>

Li, S. (2016). The Construct Validity Of Language Aptitude: A Meta-Analysis. *Studies in Second Language Acquisition*, *38*(4), 801–842.

<https://doi.org/10.1017/S027226311500042X>

Ma, D., Yao, T., & Zhang, H. (2018). The effect of third language learning on language aptitude among English-major students in China. *Journal of Multilingual and Multicultural Development*, *39*(7), 590–601.

<https://doi.org/10.1080/01434632.2017.1410162>

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and

- Multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4), 940–967.
[https://doi.org/10.1044/1092-4388\(2007/067\)](https://doi.org/10.1044/1092-4388(2007/067))
- Martin, K. I., & Ellis, N. C. (2012). The Roles of Phonological Short-term Memory and Working Memory in L2 Grammar and Vocabulary Learning. *Studies in Second Language Acquisition*, 34(3), 379–413. <https://doi.org/10.1017/S0272263112000125>
- McManus, K. (2021). *Crosslinguistic Influence and Second Language Learning* (1. Ed.). Routledge. <https://doi.org/10.4324/9780429341663>
- Meara, P. (2005). *LLAMA Language Aptitude Tests The Manual*. 22.
- Meisel, J. M. (2011). *First and Second Language Acquisition: Parallels and Differences* (1. Ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511862694>
- Mikawa, M., & De Jong, N. H. (2021). Language neutrality of the LLAMA test explored: The case of agglutinative languages and multiple writing systems. *Journal of the European Second Language Association*, 5(1), 87–100. <https://doi.org/10.22599/jesla.71>
- Nayak, N., Hansen, N., Krueger, N., & McLaughlin, B. (1990). Language-Learning Strategies in Monolingual and Multilingual Adults. *Language Learning*, 40(2), 221–244.
<https://doi.org/10.1111/j.1467-1770.1990.tb01334.x>
- Oberauer, K. (2009). Chapter 2 Design for a Working Memory. In *Psychology of Learning and Motivation* (Bd. 51, S. 45–100). Elsevier.
[https://doi.org/10.1016/S0079-7421\(09\)51002-X](https://doi.org/10.1016/S0079-7421(09)51002-X)
- Rizvanović, N. (2018). Motivation and Personality in Language Aptitude. In S. M. Reiterer (Hrsg.), *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (Bd. 16, S. 101–116). Springer International Publishing. https://doi.org/10.1007/978-3-319-91917-1_6
- Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, 1(1), 49–60. <https://doi.org/10.22599/jesla.24>

- Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *IRAL - International Review of Applied Linguistics in Language Teaching*, 46(2).
<https://doi.org/10.1515/IRAL.2008.005>
- Saito, K. (2017). Effects of Sound, Vocabulary, and Grammar Learning Aptitude on Adult Second Language Speech Attainment in Foreign Language Classrooms: Role of Aptitude in Second Language Speech. *Language Learning*, 67(3), 665–693.
<https://doi.org/10.1111/lang.12244>
- Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319 - 353). Cambridge: Cambridge University Press.
- Service, E. (1992). Phonology, Working Memory, and Foreign-language Learning. *The Quarterly Journal of Experimental Psychology Section A*, 45(1), 21–50.
<https://doi.org/10.1080/14640749208401314>
- Service, E., Yli-Kaitala, H., Maury, S., & Kim, J.-Y. (2014). Adults' and 8-Year-Olds' Learning in a Foreign Word Repetition Task: Similar and Different: Adults' and Children's Foreign Word Repetition. *Language Learning*, 64(2), 215–246.
<https://doi.org/10.1111/lang.12051>
- Sharwood Smith, M., & Truscott, J. (2006). Full transfer full access: A processing-oriented interpretation. In S. Unsworth, T. Parodi, A. Sorace, & M. Young-Scholten (Hrsg.), *Language Acquisition and Language Disorders* (Bd. 39, S. 201–216). John Benjamins Publishing Company. <https://doi.org/10.1075/lald.39.10sha>
- Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. *First Language*, 34(1), 3–23.
<https://doi.org/10.1177/0142723713510997>

- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. <https://doi.org/10.1037/0096-3445.117.1.34>
- Soveri, A., Lehtonen, M., Karlsson, L. C., Lukasik, K., Antfolk, J., & Laine, M. (2018). Test–retest reliability of five frequently used executive tasks in healthy adults. *Applied Neuropsychology: Adult*, *25*(2), 155–165. <https://doi.org/10.1080/23279095.2016.1263795>
- Stafford, C. A., Sanz, C., & Bowden, H. W. (2010). An experimental study of early L3 development: Age, bilingualism and classroom exposure. *International Journal of Multilingualism*, *7*(2), 162–183. <https://doi.org/10.1080/14790710903528122>
- Tomasello, M. (2011). Language Development. In U. Goswami (Ed), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (2nd ed., pp. 239-257). Wiley-Blackwell.
- Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550–564. <https://doi.org/10.3758/BF03195534>
- Wucherer, B. V., & Reiterer, S. M. (2018). Language is a girlie thing, isn't it? A psycholinguistic exploration of the L2 gender gap. *International Journal of Bilingual Education and Bilingualism*, *21*(1), 118–134. <https://doi.org/10.1080/13670050.2016.1142499>
- Yalçın, Ş., Çeçen, S., & Erçetin, G. (2016). The relationship between aptitude and working memory: An instructed SLA context. *Language Awareness*, *25*(1–2), 144–158. <https://doi.org/10.1080/09658416.2015.1122026>
- 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/10.1093/applin/ams044>