19-12-2021

Predicting adult language learning success in its very initial stages.

Master Thesis NT2: Docent/Expert

STUDENT:	JOSSEKE JONKER S1024127
FIRST SUPERVISOR	MARIANNE STARREN
SECOND SUPERVISOR	ROELAND VAN HOUT

Preface

Writing this thesis has been a tremendous journey. From the moment I was introduced to the VILLA project and the incredible potential of its data, I knew I wanted to be involved and I was thrilled when I heard this was possible.
I had no idea, however, about the amount of work that was waiting for me. The raw data was out there, somewhere and the first thing I did was to make an inventory of everything I needed. In my quest to do this, I got to speak with many of the researchers involved in VILLA who were without exception so very nice and willing to help me out that I felt very honored I was given this opportunity to do this. It gave me such an invaluable insight in what it means to work at academics and what a huge undertaking the VILLA experiment had been. In particular, I remember my 1,5 hour video call with Heather Hilton who not only answered the questions I had, but gave me a briefing about how the VILLA project came about and spoke enthusiastically about the potential of the VILLA data in terms of individual learner variables. It was one of the highlights in this study! Another memorable moment was when I had a question and answering session with many of the VILLA researchers who were working on the VILLA Field Manual in Paris. I cannot emphasize enough how valuable it is to be able to work alongside your thesis supervisors and to be part of a real project! It is, in my opinion, the only way to experience academic life, and the best way to learn as everybody is working towards the same goal, striving towards the best possible product, in my case a thesis that can be used as a basis for future research!

Marianne Starren, one of my supervisors and more importantly one of the VILLA researchers, whose gratitude, enthusiasm, open personality and never-failing trust in me made me feel like I was part of the team, filled me in on just about everything about the project, which along with the input of the other researchers and the information I had through the VILLA Field Manual, was essential for the success of this study.

Another key player in this process was my second supervisor Roeland van Hout. Without him I would have given up a million times, as the amount of statistics and statistical computing needed was immense. No matter how often I got stuck, he was always there for me, helping me out by breaking things down for me in smaller more comprehensible pieces, or leading the way towards a solution, always in such a way that I could learn something from it! His perfectionism pushed me to go the extra mile and I am extremely proud of the final product, which I could not have done without his help!

I really feel that this thesis is the result of team effort and I am proud that for the duration of writing this paper, I was part of that team!

Inhoudsopgave

1. Introduction

1.1 Background

Researchers have long tried to find the answer to why some students are more successful at learning another language than others. Their quest has led to many theories but very few conclusive answers, as language acquisition processes are influenced by so many, and so many potentially confounding factors that it is hard to extract the individual ingredients (Lambert & Gardner, 1963). Apart from external factors, such as exposure duration, the teaching method and the type of input, individual factors, unique to each language learner such as motivation, learning style, cognitive abilities, prior linguistic knowledge, and language aptitude play a role (Paradis, 2011). Controlling for all of these variables in order to study isolated factors has turned out to be next to impossible in a natural language learning setting and this is why research in the area of individual learner profiles has been scarce.

The VILLA project, *Varieties of Initial Learners in Language Acquisition*, (Dimroth et al., 2013; VILLA Field Manual, 2022) managed to find a format in which this could be done. The researchers recruited one single teacher, teaching a carefully devised curriculum for both learning conditions to all of the thoroughly selected participants with no prior knowledge of the language to be taught, thus ensuring uniformity in entry level, input, teaching method and exposure duration. In this longitudinal experiment (two weeks), 162 adult learners of five different project countries, recruited from the universities of the Netherlands (Radboud Universiteit), the UK (York university), France (Université Paris8), Germany (Universität Osnabrück) and Italy (Università di Pavia) took part in a Polish language course in which they were exposed to 14 hours of monolingual input in 10 sessions, structured to such degree as to allow for the testing of learners in different linguistic areas (Dimroth et al., 2013).

The main aim of the VILLA project was to study both the initial stages of language acquisition and the role of input in this process, differentiating between students who were taught in the meaning-based learning condition without explicit grammar instruction and those who were taught in the form-focused condition in which grammar was addressed more explicitly (Dimroth et al., 2013). Studying individual learner differences was not the main concern of the project, but the target was to obtain a homogeneous sample of learners in the project countries. For this reason, strict selection criteria were applied. Moreover, an extensive battery of psychometric tests was administered to all of their selected participants to control for potentially confounding factors. The effect of these individual learner variables on language acquisition, however, was never systematically researched. That means that the VILLA data, derived from their unprecedented systematic set-up in which they administered an impressive range of psychometric tests and language tasks, provided a unique opportunity to study the longitudinal effect of individual learner variables on language acquisition in its initial stages. Previous research on the predictive value of individual learner differences was never done in the same way as the VILLA project did; in a controlled learning environment with learners from five different project countries, studying the very first stages of the language acquisition process.

1.2 Research questions

According to Ehrman et al., (2003), individual learner differences that may have an influence on the language acquisition process can be categorized in language aptitude, learning styles (cognitive ability, personality, and perceptual preference), learning strategies (experiential preference), affective factors (motivation) and demographic factors (sex, age and level of education). Then there are those learner differences that have an influence on the learning process in general, such as the working memory (Gathercole et al., 2008; Unsworth & Engle, 2005) and attention span (Fernandez-Castillo & Gutiérrez-Rojas, 2009; Gsanger et al., 2002; Riccio et al., 2003).

In this present study we will use the VILLA data to address the following three research questions:

1) Are there any differences between the countries in the way learner variables are represented?

2) Does the distinction between meaning- and form-based learning conditions play a role in predicting language learning success?

3) Can individual learner differences predict language learning success?

1.3 The present study

Answering the above research questions turned out to be more complex than we thought. As the raw data we had access to had not been processed in a systematic way, we had no other choice but to do this ourselves, resulting in a thesis with a stepwise set-up. The entire process necessary to ultimately answer the research questions can be broken down into three steps:

Step 1: Documenting and analyzing the learner variables
Step 2: Analyzing the test results of a selection of language tasks
Step 3: Analyzing the effects of a selection of individual learner differences

Step 1: Documenting and analyzing the learner variables
A large part of the work put into this thesis was made up of collecting and documenting information that was stored in various places. Fortunately, many of the researchers involved in the VILLA project were happy to assist so that missing or incongruent information could be retrieved or explained. Another invaluable source of information was the VILLA Field Manual, a document the VILLA researchers were working on at the time of writing this thesis, which is to be published in 2022. In the VILLA Field Manual (2022), the authors describe various stages and aspects of the VILLA project. In particular, the sections on the individual learner differences and the psychometric tests, written by Heather Hilton and the ones on the various language tasks, written by Rebekah Rast, were extremely helpful in creating a complete picture of the test battery employed by VILLA. All of this ultimately led to chapter 2, in which we give an extensive overview of the learner variables collected by VILLA and describe the differences of all these learner variables in the five project countries.

Step 2: Analyzing the test results of a selection of language tasks
Chapter 3 gives an overview of the language tasks employed by VILLA. We give a detailed description of the three tasks we selected for this research, and we will briefly describe the research methods used in this study. As the focus of this thesis was to study the longitudinal effect of individual learner variables on language acquisition in its initial stages and time did not permit us to analyze all of the language tasks, we made a selection of tasks that both covered different language domains and those that had been administered at a minimum of two different time intervals. The selected tasks are Word Recognition, Grammaticality Judgment I and Phoneme Discrimination, covering the language domains Lexicon, Morphology and Phonology. The data for these tasks was collected using E-Prime experimental software (Schneider et al., 2002) which enabled us to finally do an analysis across all of the project countries and the two learning conditions, something that had not been done before. All of this was a lot of work, but necessary in order to find the answers to my research questions. Chapters 4 to 6 are therefore dedicated to the comparison of the results of the language tasks between the five project countries. We will look at the progress made between the test sessions as well as the effect of form-based vs. meaning-based learning conditions.

Step 3: Analyzing the effects of a selection of individual learner difference
In chapter 7 the correlations are given both between the psychometric tests and the language tasks. Finally, the effect of a subset of learner variables on the results of the language tasks are investigated by applying forward regression.

2.Learner variables

2.1 Introduction

Researching the impact of individual differences on language learning was not part of the basic research questions of the VILLA project. The assumption was that all teams involved would sample a similar set of learners, given the selection criteria agreed upon. With this set of relevant learner variables, a comparison can be made both within and between the countries. The learner variables consisted of background data and the results of a series of psychometric tests. It is easy to check how homogenous the learner samples were in the different countries with respect to a number of background characteristics, and whether there are differences between countries. Many of the psychometric tests were administered to control for potentially confounding learner variables (Dimroth et al., 2013) and to ascertain that there were no large differences between the countries in the samples of participants. Despite of the fact that individual differences were not the primary interest of most of the researchers involved in the project, the question of the role these learner variables play in the acquisition of a new language is still a relevant one and for this reason their effects are the core research question of this thesis.

Table 2.1 gives an overview of the learner variables and their corresponding tests and/or questionnaires that were administered in the VILLA project. A distinction is made between five levels of learner information: background, cognition, language aptitude, personality, and motivation.

Table 2.1. Overview of learner variables administered through tests or questionnaires used in the VILLA project.[1]

	Variable	Instrument [2]	Administration, remarks
background			
1	general data	in-house questionnaire on age, sex, type of study, and linguistic profile	completed by each participant
2	language profile	in-house questionnaire + interview + Language Sensitivity test	preliminary screening, to eliminate subjects with prior knowledge of Polish or other Slavic languages.
tests: cognition			
3.	nonverbal intelligence	Raven's Standard Progressive Matrices (Raven 1981), extracts.	collective administration (paper & pencil answer sheet + slideshow presentation of items.
4.	executive function (working memory)	Digit span task	administered individually online, TalkBank website (MacWhinney 2012)
5.	executive function (working memory)	Letter Number Sequencing task	
6.	executive function (attention, switching and inhibition)	Flanker Task	
7.	perceptual preference	Barsch (1980) Learning Styles Inventory	paper & pencil questionnaires, completed individually (at home or at the learning venue)
8.	cognitive style[3]	(exploratory) in-house Wordlist Task	say the first ten Polish words you remember from Week 1 (recorded beginning of Week 2)
tests: language aptitude			
9.	language aptitude: grammatical inferencing	Llama Language aptitude tests (Meara 2005): Test F	computerized tests, administered individually.

[1] This overview is an adaptation of Dörnyei's (2005) list.

[2] When possible, standardized instruments were used; otherwise, existing tools (such as Gardner's motivation questionnaire) were adapted for the context of the project.

[3] This test was only administered in France and Italy, and in Germany for the group of children. Based on work from Nelson (1973; 1981) More information in VILLA Field Manual (2022) (Nelson's L1 acquisition data).

10.	word-learning skill	Llama Language aptitude tests: Test B	automatically scored. (no access to itemized answers)
11.	phonological recognition	Llama Language aptitude tests: Test D	
tests: personality			
12.	personality	NEO Five Factor Inventory FFI-3 Adult, short version (Costa & McCrae 2010)	Paper & pencil questionnaire, collective administration.
13.	experiential preference	Isalem-97 (Cahay et al. 1997)	Paper & pencil questionnaires, completed individually (at home or learning venue)
tests: motivation			
14.	motivation for Polish	Motivation Questionnaire adapted from Gardner (2004)	Paper & pencil questionnaire, completed individually (at home or learning venue)

Selection

The VILLA project defined a clear set of criteria to be met by the participants. The participants had to be students in higher education meeting the following selection criteria:

- Be between 18 to 28 years old.
- Be a native speaker of the language spoken in the project country where participants were recruited.
- Have no more than the equivalent foreign language competence required by the schooling system of the project country where participants were recruited. Bi-or multilinguals, or students with extensive L2 learning experience were excluded.
- Have no known history of dyslexia or any other language related problems such as hearing problems or reading problems.
- Be unfamiliar with the target language Polish, Russian or any other Slavic languages. This was tested prior to selection through a language sensitivity test.
- Be enrolled in majors other than Linguistics or other language studies, Cognitive Science and Psychology.

Furthermore, potential candidates had to be available for three hours every day on weekdays during the entire period of the experiment.[4] As it was crucial that all participants who had signed up for this project remained committed and attended all the sessions, students had to sign a contract and were offered an incentive upon completion of the experiment.

The recruitment procedure consisted of three parts. All interested candidates were invited to an interview session where they were asked to fill in a questionnaire, they were then briefed about the project and asked to sign a contract after which the language sensitivity test was administered. Selection took place after the interviews.
Given the project guidelines, each country selected their sample of learners. Selected candidates were randomly assigned in order of appearance to one of two learning conditions: the meaning- based input group or the form-based input group. Table 2 gives an overview of the total sample across all countries, subdivided into two learner conditions.

Table 2.2. Number of participants in the participating VILLA countries

Learner groups		Source languages				
		Dutch	**English**	**French**	**German**	**Italian**
Adult learners, meaning-based input	20	17	17	20	15	
Adult learners, form-based input		20	18	19	–	14

Table 2.2 makes clear that there was no form-based input group in Germany. Germany opted to compare two groups that differed in age; an adult group and a group with children aged 10-11.
Both groups received meaning-based instruction (Dimroth, 2013).

[4] 14 hours over 10 subsequent weekdays, interrupted by one weekend

Note of interest

Despite the above-mentioned selection criteria, some participants who should have been excluded somehow ended up being part of the study. Table 2.3 gives an overview of the students who didn't meet all of the criteria but did take part in the experiment.

Table 2.3. Overview of irregularities regarding the selection criteria

Country	Participant	Irregularity
France	1112	Extensive L2 learning: Spent 13 years in South America
	1214	Language related problems: reading issues (vision)
	1218	Extensive L2 learning: Father's L1 Creole French
The Netherlands	2109	Major: Language & Culture
	2202	Major: Language & Culture
	2205	Language related problems: hearing + reading issues
The United Kingdom	3101	Major: English Literature
	3107	Major: Linguistics
	3116	Language related problems: visually impaired
	3119	Some tests are missing
	3204	Language related problems: Dyslexia
	3211	Extensive L2 learning: Parents L1 Hindi
	3214	Extensive L2 learning: Mother Scottish
	3219	Did not do all the tests
Italy	5104	Major: Psychology
	5105	Major: Foreign Languages
	5106	Major: Psychology
	5114	Major: Psychology
	5115	Major: Psychology
	5202	Did not do all the tests (No PD, or WR data)
	5219	Did not do all the tests (No PD, or WR data)

2.2 Background

2.2.1 General data (1)

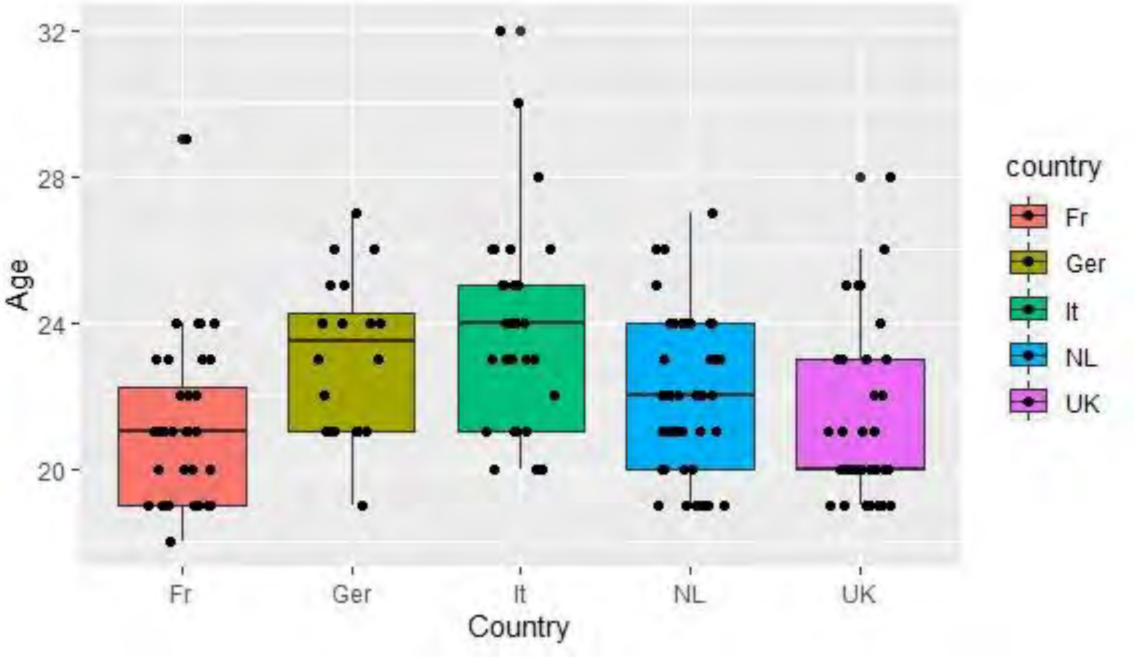
Age

The age of the participants in the different project countries ranged from 18 to 30. More details on the age distribution can be found in Table 2.4 below. Figure 2.1 gives the box plots on age for the five countries involved.

Table 2.4. Age distribution (N=162)

Country	Sample size (N)	Median years old	Youngest years old	Oldest years old
France	36`	21	18	29
Germany	20	23.5	19	27
Italy	31	24	20	32
The Netherlands	40	22	19	27
United Kingdom	35	20	19	28

Figure 2.1 Age distribution per country (N=160 with 2 missing values)[5]



Data distribution.
As can be seen in Figure 2.1, the median age of the countries ranges between 20 and 24. Italy catches the eye because of its sample with relatively older participants, including a participant of 32. The UK data are interesting, because it shows an atypical distribution., as 25 % of the participants had an age similar to the median of 20. In total there were three participants with ages that fell beyond the age range defined in the selection criteria: one in France (29) and two in Italy (30 and 32).

Statistical significance
An ANOVA test was run on age differences between countries, showing that there was a significant age difference (F(4, 155)=7.309;p<0.000).

A subsequent Tukey post hoc test shows significant age differences between France and Germany (p< 0.030), France and Italy (p<0.000), Italy and the Netherlands (p< 0.010), and Italy and the UK (p< 0.000). These distinctions point out that students are significantly younger in France, the Netherlands and the UK.

Sex

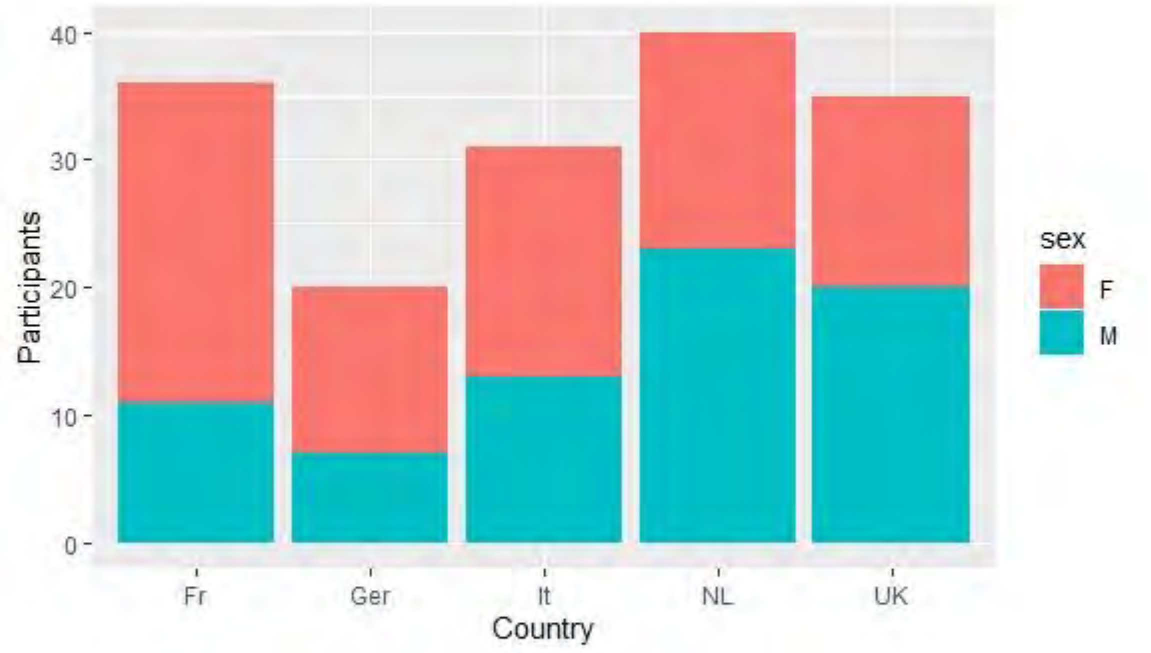
Table 2.5 gives the distribution of sex over the country samples. These differences are visualized in Figure 2.2.

Table 2.5. Number of males and females per country (N=162)

Country	Females	Males
France	25`	11
Germany	13	7
Italy	18	13
The Netherlands	17	23
United Kingdom	15	20

[5] The dots in the plot represent the individual values.

Figure 2.2 Number of males and females per country (N=162 with 0 missing values)



Data distribution.

Overall, more females participated than males.

In France the imbalance in the female/male distribution is particularly salient, with only 11 males participating compared to 25 females. Interestingly, in both the Netherlands and the UK, more males than females participated.

Statistical significance

In order to check whether there is a significant difference in the distribution of males and females in the different project countries, a Pearson's Chi Square test was run, showing no significant difference in the female/male distribution (χ2= 8.519, df = 4, p < 0.074), indicating that the found differences might be due to sample variation.

Type of Study

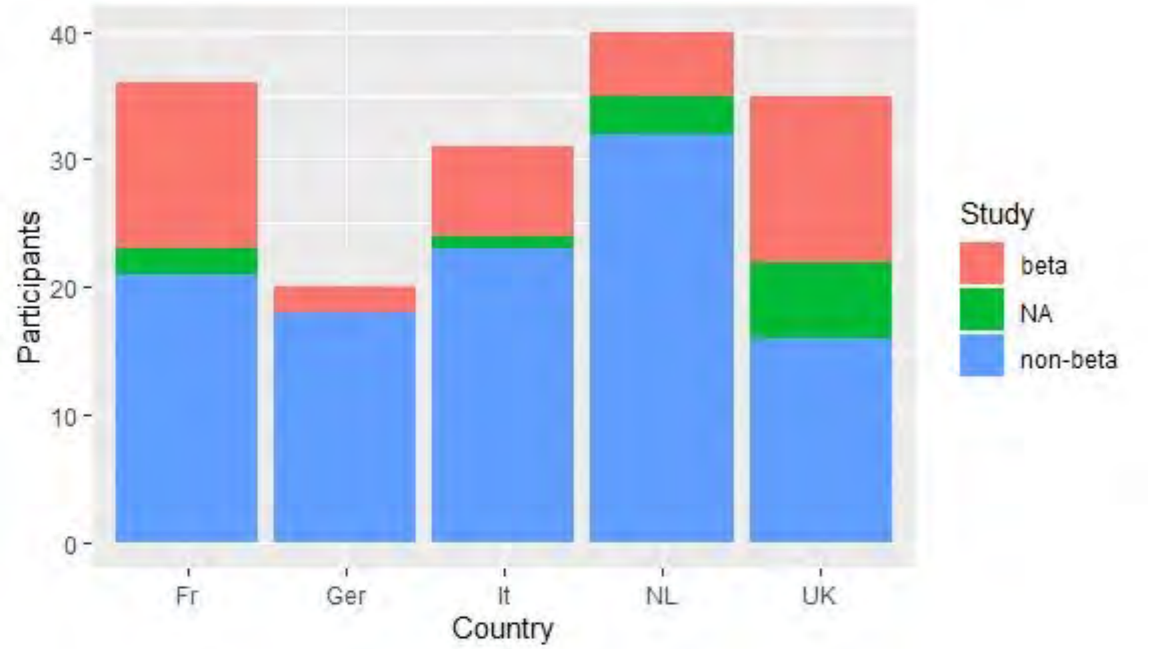
Despite the fact that students with majors in psychology, cognitive science, linguistics or any other language-studies did not fit the VILLA profile and therefore should have been excluded from the study, several of them were included in the final sample. This happened in the project countries Italy, the Netherlands and the UK. In Italy five such students participated with majors in Foreign Language (1), and Psychology (4). The Netherlands had two participants not meeting this selection criterium with students studying Language and Culture (2), The UK had two non-profile students who majored in Linguistics (1) and English Literature (1). Table 2.6 gives an overview of the distribution of all the major studies included in this project.

Table 2.6. Distribution of major studies

Major	**Country**					
	Fr	Ger	It	NL	UK	Tot
anthropology	0	0	0	2	0	2
architecture	1	0	0	0	0	1
art	1	0	0	0	0	1
art & music	0	1	0	0	0	1
arts	0	0	1	0	0	1
biology	2	1	0	0	5	8
business	6	0	0	1	1	8
business communication	0	0	0	2	0	2
chemistry	1	0	0	2	1	4
cinema	1	0	0	0	0	1
communication	1	0	0	0	0	1
computing	1	1	0	0	0	2
culture studies	0	0	0	2	0	2
dentistry	1	0	0	0	0	1
digital communication	0	0	0	1	0	1
dutch law	0	0	0	2	0	2
economics	1	0	2	0	0	3
engineering	0	0	5	0	0	5
english literature	0	0	0	0	1	1
european studies	0	1	0	0	0	2
foreign languages	0	0	1	0	0	1
geography	1	0	0	0	1	2
history	1	1	0	4	0	6
human geography	0	0	0	2	0	2
human movement sciences	0	0	0	1	0	1
humanities	0	0	3	0	0	3
international migration	0	1	0	0	0	1
language & culture	0	0	0	2	0	2
law	4	8	5	1	8	26
linguistics	0	0	0	0	1	1
maths	1	0	1	0	1	3
maths & economics	0	0	0	0	1	1
MBA	0	3	0	0	0	3
MBA business & eco	0	1	0	0	0	1
medicine	3	0	1	1	0	5
music/musicology	1	0	1	0	0	2
NA	2	0	1	3	6	12
north america studies	0	0	0	1	0	1
osteopathy	1	0	0	0	0	1
pedagogy	0	0	1	1	0	2
philosophy	0	0	2	1	0	3
physical therapy	1	0	0	0	0	1
physics	0	0	0	0	1	1
physics/astronomy	0	0	0	1	0	1
planning	0	0	0	1	0	1
political philosophy	0	0	0	0	1	1
politics	0	0	0	2	0	2
post graduate certificate in education	0	0	0	0	2	2
post graduate certificate in education (maths&science)	0	0	0	0	1	1
post graduate certificate in education (maths)	0	0	0	0	1	1
post graduate certificate in education (science)	0	0	0	0	2	2
psychology	0	0	4	0	0	4
sciences po	0	0	3	0	0	3
social geography	0	0	0	3	0	3
social work	0	0	0	1	0	1
sociology	4	0	0	2	0	6
speech therapy	1	0	0	0	0	1
sports	0	1	0	0	0	1
theology	0	1	0	0	0	1
theoretical physics	0	0	0	1	0	1
writing, directing, performance	0	0	0	0	1	1

In Figure 2.3 the distribution of beta and non-beta students across the countries is visualized.
The label "NA" was given whenever there were missing values.

Figure 2.3 Beta vs non-beta distribution (N=150 with 12 missing values)



Data distribution.

Across the board, countries seemed to have far more participants involved in non-beta studies than those involved in beta-type studies (N= 162, beta= 40, NA= 12, non-beta= 110). This difference seems most salient in Germany and the Netherlands, where respectively 90%, and 80% of its candidates were enrolled in a non-beta study. France and the UK have a more equal distribution when it comes to study type.

Statistical significance

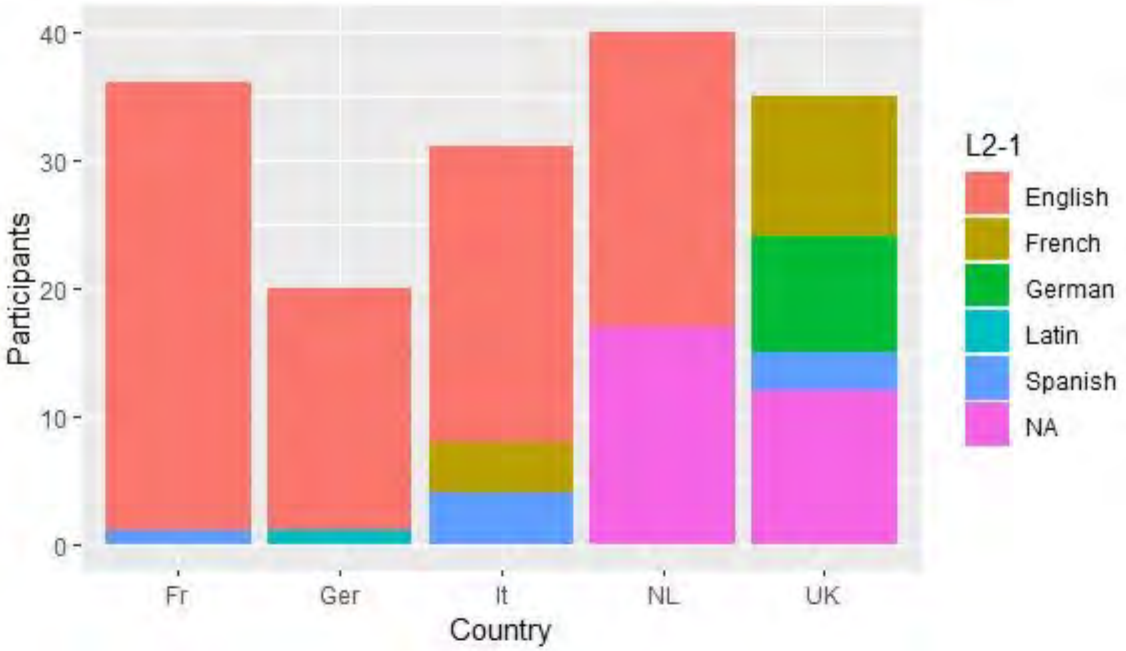
In order to check whether there is a significant difference in the distribution of beta and non-beta students in the different project countries, a Pearson's Chi Square test was run, showing a significant difference in the beta -non-beta distribution ($\chi2$ = 20.674, df = 8, p< 0.08), indicating an association between the variables that cannot be attributed to chance. This means that the countries have different distributions between beta and non-beta students.

Background languages/ L2 status:

Despite the fact that Third Language Acquisition (TLA) or Acquisition of Alternative Languages (AAL) as the field is also known is a relatively young area of expertise, studies that have been carried out so far do provide evidence that background languages other than the L1 can actually influence the acquisition of a new language (Cenoz et al., 2001; De Angelis, 2007; Ringbom, 1987; Williams & Hammarberg, 1998). For successive adult language learners, learning a third or additional language is fundamentally different from learning a second language in the sense that L3 or Ln learners can benefit from enhanced language awareness, acquired language learning strategies and increased potential for cross-linguistic interferences (CLI) between L1 and background languages that can occur in additional language acquisition (Cenoz et al.,2001; De Angelis 2007). Although there are numerous studies on CLI, relatively little is known to date about the exact way in which languages influence each other (De Angelis, 2005; Falk & Bardel, 2010; Slabakova, 2017). Factors known to affect CLI are *typological distance* (Cenoz et al., 2001; De Angelis & Selinker, 2001; Kellerman, 1983; Schepens et al., 2016), *L2 status* (Hammarberg, 2001; Williams & Hammarberg, 1998), *recency* (Hammarberg, 2001), *context* (Dewaele, 2001), *proficiency* (Ringbom, 1987; Tremblay, 2006; Williams & Hammarberg, 1998), *order of acquisition of the languages* (Dewaele, 1998) and *constraints on verbal memory* (Williams & Hammarberg, 1998).

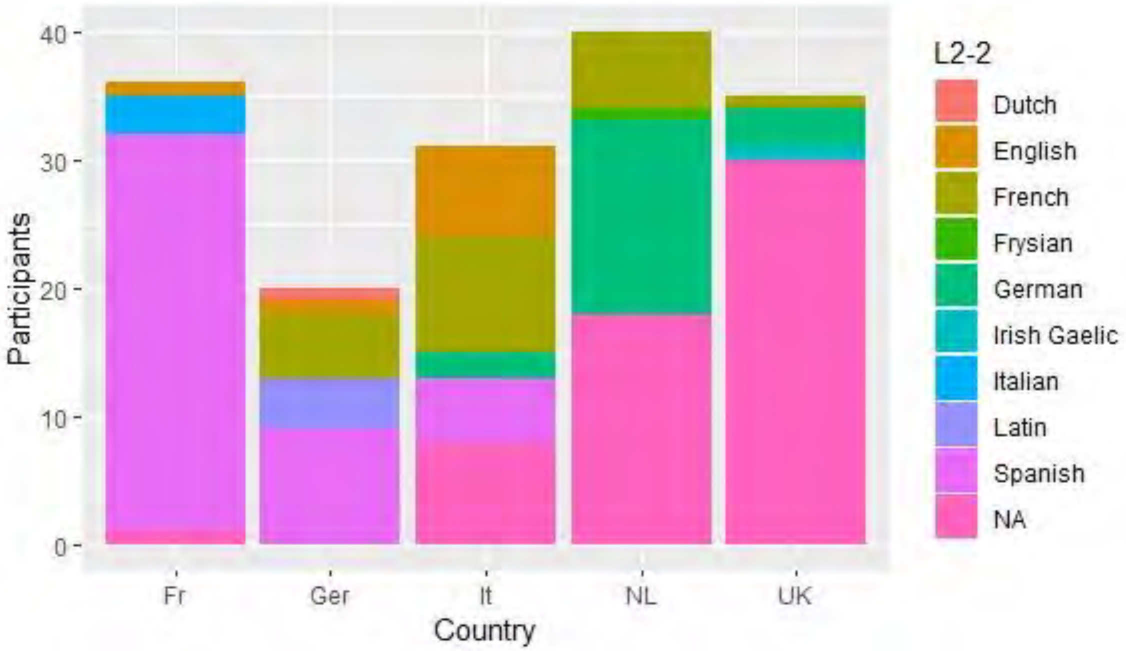
An extensive study on this topic, clearly is beyond the scope of this thesis, but it would be interesting to see if respondents' background languages play a significant role in their language success and for this purpose the distribution of the student's background languages are given in Figures 2.4, 2.5 and 2.6.

Figure 2.4 Distribution of participants' First L2 per country (N=134 with 28 missing values)



Note: The missing values (NA) in the Netherlands are due to missing data[6] (NA=17)
Two missing values in the UK are due to missing data, the remaining missing values in the UK are due to the fact that for some participants Polish was their first L2 (NA= 11).

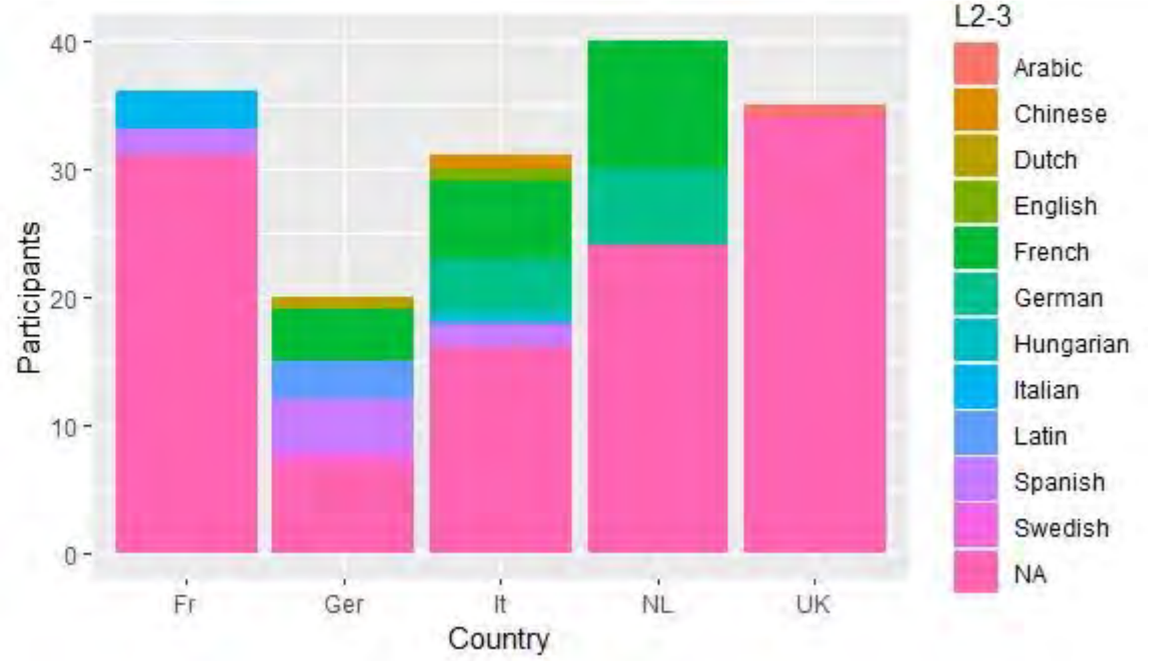
Figure 2.5 Distribution of participants' second L2 per country (N=105 with 57 missing values)



Note: The missing values (NA) in the Netherlands are due to missing data[6] (NA=17)
Two missing values in the UK are due to missing data, the remaining missing values in the UK (NA=28) are due to the fact that for some participants Polish was their first L2.
The missing values in the in the other countries are due to the fact that in these countries participants did not have a second L2.

[6] This data has yet to be digitalized and transferred into the VILLA data, but was not available to me at the time of writing this thesis.

Figure 2.6 Distribution of participants' third L2 per country



Note: The missing values (NA) in the Netherlands are due to missing data[6]
Two missing values in the UK are due to missing data, the remaining missing values in the UK (NA=30) are due to the fact that for some participants Polish was their first L2.
The missing values in the other countries are due to the fact that in these countries participants did not have a third L2.

2.2.2 Language Profile (2)

Selection criteria regarding the participants' language profile were quite strict. First of all, candidates had to have no prior knowledge of Polish or any other Slavic language. Moreover, they couldn't be bi- or multilingual or have any extensive L2 learning experience. Preferably, the candidates had to have acquired his/her second language skills through the schooling system of his/ her country (Dimroth et al., 2013). Furthermore, students with language related disabilities such as dyslexia as well as students with hearing and or reading problems were excluded from the project.

A language sensitivity test was administered in order to check whether candidates truly had no knowledge of the Polish language or any other Slavic languages. During the test, participants had to listen to a series of 18 sentences recorded by native speakers of Polish, Finnish and Russian and decide after each sentence, whether the utterance heard was in Polish or in another language. Only those students who could not distinguish Polish from the other two languages were included in the project.

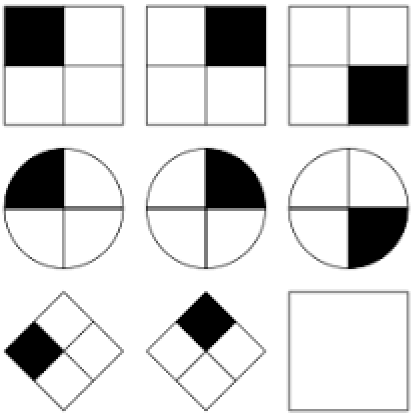
2.3 Tests: Cognition

2.3.1 Nonverbal Intelligence: Raven's Standard Progressive Matrices (3)

Nonverbal Intelligence was tested using Raven's Standard Progressive Matrices, a classic method used to assess general intelligence and abstract reasoning. The test measures the two main components of general cognitive ability; *eductive ability*, the extent to which test-takers are able to make meaning out of abstract complexity, and *reproductive ability*, the extent to which test-takers are able to absorb, store, recall and reproduce information (Raven, 2000; Spearman, 1923).

There are several versions of Raven's Progressive Matrices (RPM), but the VILLA project used the Standard Progressive Matrices (SPM). This test consists of a series of diagrams or designs where one part is missing. Test-takers are asked to identify the missing item that completes a pattern. (Raven, 2000). An example item is shown in Figure 2.4.

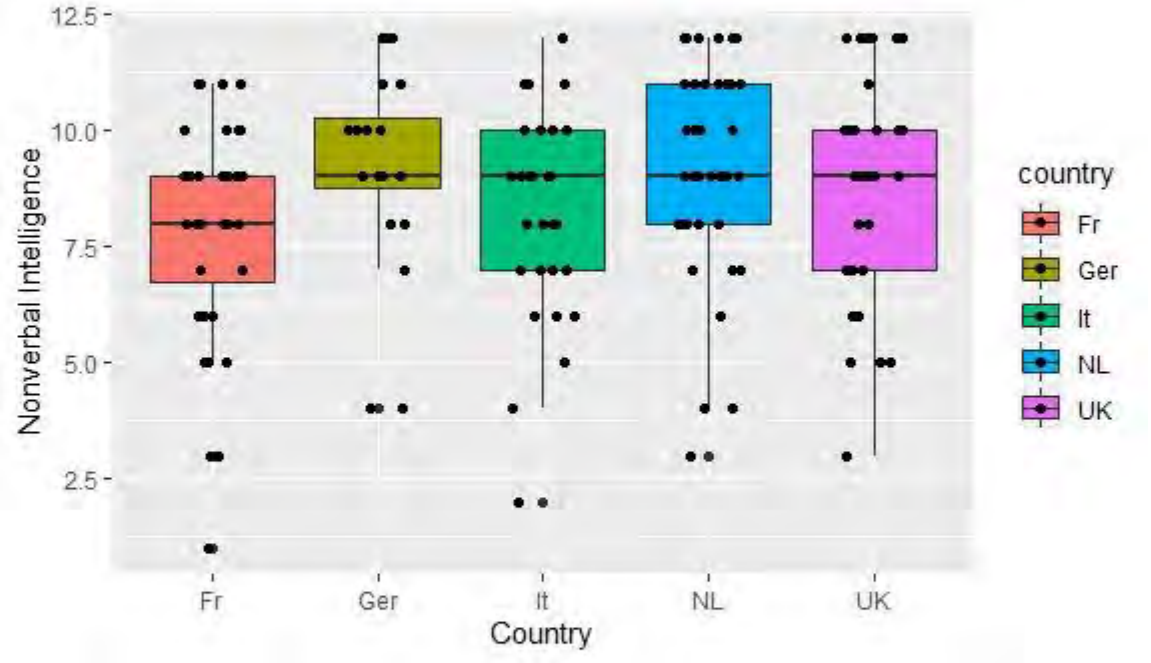
Figure 2.7 Example test item Raven



Due to time restraints, the VILLA project opted to use only a selection of 18 items, rather than the original 60.[7] This means that any conclusions drawn from this test are meant to be indicative only. Scoring of the test was fairly straightforward; respondents received one point for every item correctly answered, resulting in a maximum score of 18 on this test. An overview of the results on this test of all the participating countries can be found in Figure 2.5.

[7] According to Heather Hilton, one of the researchers involved in the VILLA project, this selection was made by a group of psycholinguists at the University of Savoie. She believes the test included a few items from each set of stimuli, working through the different levels of complexity of the original 60-item test.

Figure 2.8 Results Raven test per country (N=162 with 0 missing values)



Data distribution
Looking at the boxplots in Figure 2.8, we can see that there are more similarities than differences between the coutries. Without exception, the data of the countries seem to be skewed towards the lower scores. Moreover, apart from Germany, all countries seem to have more or less the same box sizes, indicating that the variance shown in the results between the countries is about the same. The smaller box size in Germany suggests a greater internal consistency when it comes to the scores; more students had similar scores. The country with the most outliers is France. Germany, Italy and the Netherlands each have one and the UK doesn't have any at all.

Across the board, countries seem to perform equally well on this test, with an identical median of 15 in all countries, except for France. France is the only country with a median below 15 and outliers scoring less than 10.

Statistical significance
An ANOVA test was run to find out if there were any significant differences between the countries on the variable RAVEN. This turned out not to be the case (F(4, 157)= 2.283; p<0.062).

2.3.2 Executive Function (4-6)

Executive Function (EF) is a term used to denote brain processes involved in the execution of goal-oriented behavior, that is the conscious management of cognitive processing (Diamond, 2013). Miyake et al. (2000) created a model in which they further divide these brain processes in three sub-processes: *switching* (the ability to adapt when faced with changing rules), *inhibition* (the capacity to suppress any irrelevant input) and *working memory* (the extent to which verbal or non-verbal information can be kept in mind and manipulated (Goriot, 2019).

There is still no consensus amongst researchers as to the relationship between language development and EF (Gooch et al., 2016; Tonér et al., 2021). Some studies suggest that aspects of EF seem to be essential for language development (ten Braak et al., 2018; Weiland et al., 2014; Woodland et al., 2016). Other research however claim that language plays a crucial role in the development of EFs (Botting et al., 2017; Kuhn et al, 2014; Miller & Marcovitch, 2015). There are also studies with results pointing in the direction of a dynamic relation between language and EFs (Bohlmann et al., 2015; Friend & Bates, 2014; Tonér et al., 2021). Although the exact nature of the relationship between executive functions and language development is to date still unclear, studies do frequently report correlations between the two constructs (Carlson et al., 2005; Gooch et al., 2013; Muller et al., 2009), which is why it was included in the psychometric test battery.

In order to obtain information about the scope of the executive functions of the participants, three tests were administered: the Digit Span Task, assessing attention, auditory processing, and mental manipulation (Groth-Marnat, 2009; Reynolds, 1997; Sattler & Ryan, 2009), the Letter–Number Sequencing Task, measuring working memory, mental manipulation, attention, concentration, and short-term auditory memory (Groth-Marnat, 2009; Kaufman & Lichtenberger, 1999, 2006; Lichtenberger & Kaufman, 2009; McCabe et al., 2010; Sattler & Ryan, 2009).,and the Eriksen Flanker test, evaluating the degree of selective attention, switching and inhibition (Eriksen & Eriksen, 1974; Grundy et al., 2017; Mc Dermott et al., 2007).

The above-mentioned executive functions tasks were designed by Brian MacWhinney, at Carnegie-Mellon University, and translated into the five project languages with the help of the VILLA team (VILLA Field Manual, 2022). The respondents took the test online via the *TalkBank* website[8] which generated automatic data files for each participant's responses (VILLA Field Manual, 2022).

Unfortunately, due to the intensive scheduling of the Polish language tasks on the project computers, supervised administration of these three instruments was not always possible in all project countries[9]. Test-takers couldn't take the test more than once, but in an unsupervised condition it is not unlikely they could have used unauthorized techniques for noting down the stimuli in the Digit Span or Sequencing Tasks. The reliability of the scores in those countries is therefore problematic and the data collected with these instruments need to be carefully examined for suspicious response patterns (VILLA Field Manual, 2022).

[8] The tests used in the VILLA project can be found on: https://sla.talkbank.org/tasks/#

[9] At the time of writing this thesis, it was unclear in which countries the test had been administered under supervision.

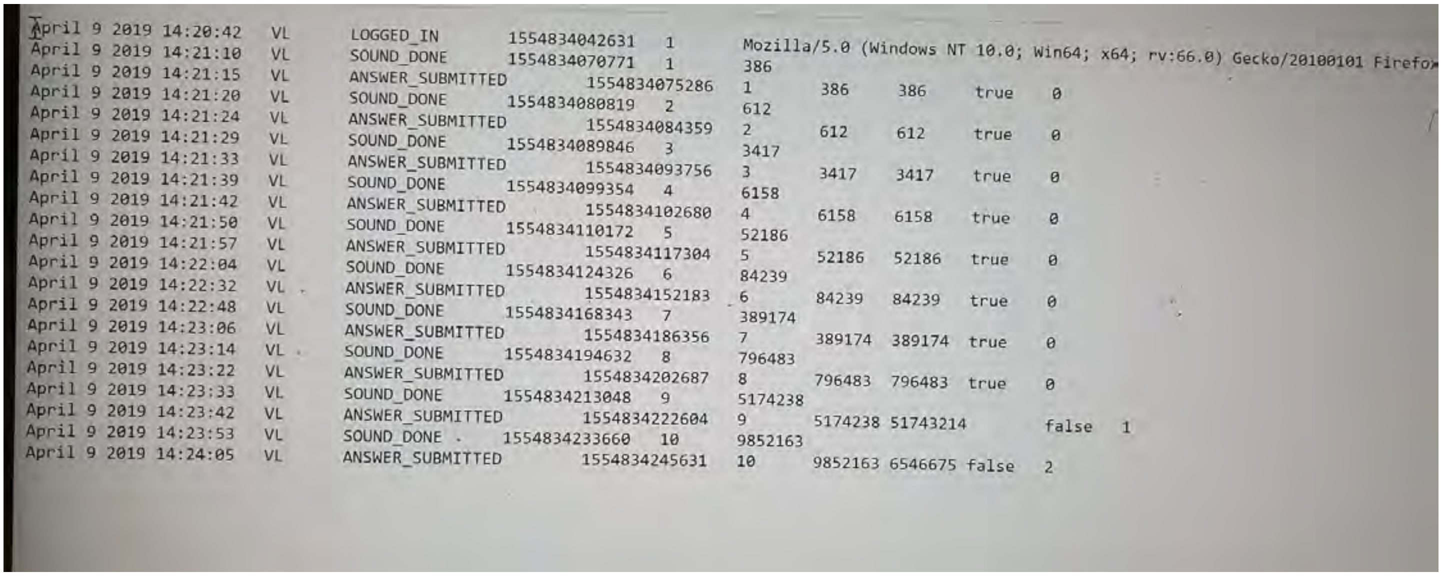
2.3.2.1 Executive Function: Digit Span (4)

In the Digit Span Test, subjects are asked to repeat a series of numbers that are read out loud to them in a certain pace. After each successfully completed trial, an additional digit is added to the sequence. If a subject fails to repeat the entire sequence in the right order, another sequence is read out to them with the same number of digits and the participant gets one more chance to get it right.

There are two varieties: the forward span and the backward span. In the forward span, subjects need to read the sequence back to the examiner in the same order and in the backward span, subjects read the sequence back to the examiner in reverse order.The VILLA project opted for the forward digit span only, which measures attention efficiency and capacity (*Brief descriptions of the most commonly used measures/testing procedures*, z.d.).

Unlike the face-to-face digit span (where the participant responds to the examiner orally), VILLA participants typed their answers (in text form) into a special answer box on the screen. Respondents were instructed to only start typing after they had heard the entire sequence, but as it was not made impossible for participants in an unsupervised setting to start typing as the sequence was being read to them, or to use other unauthorized techniques such as writing the sequence down on a piece of paper, the results of this test in those countries where students had to take the test on their own are questionable (VILLA Field Manual, 2022)[10]. Figure 2.9 shows a sample logfile of a random subject (not part of the VILLA project) and includes the presented sequences, the answers entered by the respondent and the status of the answers (true or false).

Figure 2.9 Sample logfile Digit Span Task (Talkbank)



```
April 9 2019 14:20:42   VL    LOGGED_IN         1554834042631   1     Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:66.0) Gecko/20100101 Firefo
April 9 2019 14:21:10   VL    SOUND_DONE        1554834070771   1     386
April 9 2019 14:21:15   VL    ANSWER_SUBMITTED        1554834075286   1       386      386      true    0
April 9 2019 14:21:20   VL    SOUND_DONE        1554834080819   2     612
April 9 2019 14:21:24   VL    ANSWER_SUBMITTED        1554834084359   2       612      612      true    0
April 9 2019 14:21:29   VL    SOUND_DONE        1554834089846   3     3417
April 9 2019 14:21:33   VL    ANSWER_SUBMITTED        1554834093756   3       3417     3417     true    0
April 9 2019 14:21:39   VL    SOUND_DONE        1554834099354   4     6158
April 9 2019 14:21:42   VL    ANSWER_SUBMITTED        1554834102680   4       6158     6158     true    0
April 9 2019 14:21:50   VL    SOUND_DONE        1554834110172   5     52186
April 9 2019 14:21:57   VL    ANSWER_SUBMITTED        1554834117304   5       52186    52186    true    0
April 9 2019 14:22:04   VL    SOUND_DONE        1554834124326   6     84239
April 9 2019 14:22:32   VL    ANSWER_SUBMITTED        1554834152183   6       84239    84239    true    0
April 9 2019 14:22:48   VL    SOUND_DONE        1554834168343   7     389174
April 9 2019 14:23:06   VL    ANSWER_SUBMITTED        1554834186356   7       389174   389174   true    0
April 9 2019 14:23:14   VL    SOUND_DONE        1554834194632   8     796483
April 9 2019 14:23:22   VL    ANSWER_SUBMITTED        1554834202687   8       796483   796483   true    0
April 9 2019 14:23:33   VL    SOUND_DONE        1554834213048   9     5174238
April 9 2019 14:23:42   VL    ANSWER_SUBMITTED        1554834222604   9       5174238 51743214          false   1
April 9 2019 14:23:53   VL    SOUND_DONE .      1554834233660   10    9852163
April 9 2019 14:24:05   VL    ANSWER_SUBMITTED        1554834245631   10      9852163 6546675 false   2
```

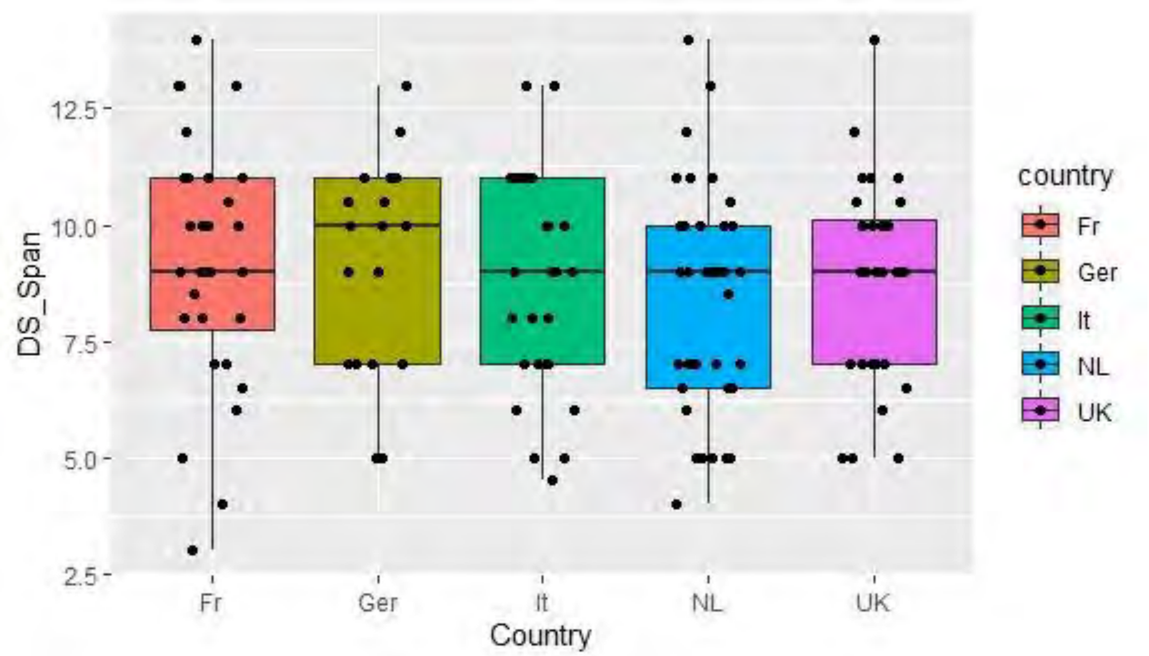
Talkbank produced two automatic scores for the Digit Span task[11]; the DS_Span, which calculates the number of digits in the last correctly repeated series and a DS_Score which counts each instance of correct repetition regardless of the number of digits in the series. The DS_Span assesses phonological capacity. The visual representation of the scores can be found in Figure 2.10. The DS_Score, represented in Figure 2.11, can be useful to distinguish between those students who were close to reaching the next level. Subjects were given one additional point if they got one digit in the following series correct and half a point if they got two wrong, but some more right. A note of warning must be given, since there was no cut-off, subjects could have start applying a strategy such as writing things down, of typing while listening (VILLA Field Manual, 2022).

[10] An overview of countries with and without supervision could not be obtained at the time of writing this thesis but would of course shed light on this issue.

[11] The Talkbank tasks were scored automatically. The exact way in which points were attributed was unclear at the time of writing this thesis. Heather Hilton advised me to look at the raw data in Talkbank, to which I had no access.

Pearson's r was calculated and rendered a very high correlation (r= 0.968) between the two automatic scores, indicating a great similarity between the two scores.

Figure 2.10 Results DS_Span (Digit Span test) per country (N= 146, with 15 missing values)



Data distribution

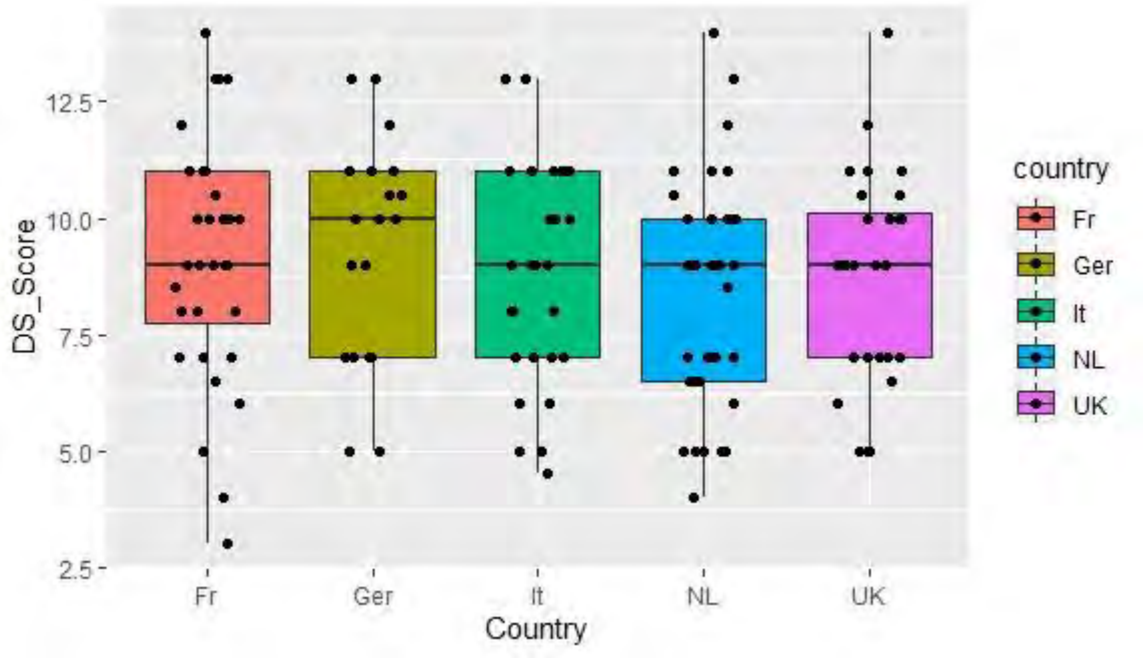
Looking at the boxplots in Figure 2.10, we can see that the data of all five countries are comparatively symmetric, indicating an even spread. Italy, the Netherlands, and the UK are similar it the way their data is distributed, with medians of six, scores ranging between four and nine and no outliers. The Netherlands and the UK are even identical in the way their data is distributed. Germany and France show different patterns. The French participants did not score less than five and had a median that was slightly higher than the median of Italy, the Netherlands, and the UK. Most of Germany's participants scored around the median of seven and with this, the country seems to be the best scoring country of the five. The only country with any outliers is France.

Germany had many participants who managed to recall 7 digits in a series, which is quite exceptional, and one has to wonder whether unauthorized techniques were used during the test, which might make the results unreliable. Looking at the other countries, we see that, with the exception of France, participants both within countries and across countries, seem to perform equally well. The French subjects, notwithstanding the three outliers, seem to perform slightly better, with a median of 6.5 and 50 % of the candidates scoring between 6 and 7.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable "DS_Span". This turned out not to be the case (F(4, 142)= 0.95; p<0.437).

Figure 2.11 Results DS_Score (Digit Span test) per country (N= 147, with 15 missing values)



Data distribution

Looking at the boxplots in Figure 2.11, we can see various degrees of symmetry amongst the countries in the way their data is distributed with Italy showing most symmetry and France the least.

Across the board, judging by the box sizes, countries showed similar degrees of variance as well.

All countries but Germany have identical medians, indicating that there's great consistency amongst the countries in the way participants scored on this test, with the German participants slightly outperforming the others.

There are no outliers.

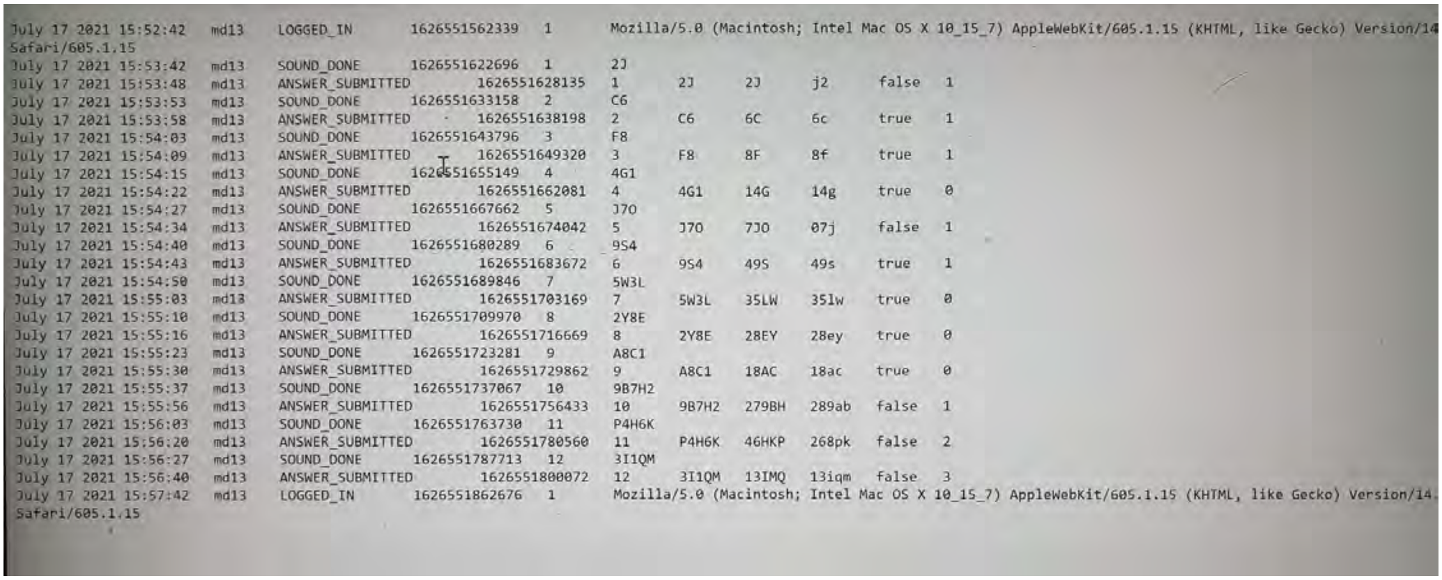
Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable "DS_Score". This turned out not to be the case (F(4, 142) = 0.931 ; p<0.448)

2.3.2.2 Executive Function: Letter Number Sequencing Task (5)

In the Letter Number Sequencing Task (LNS-task), subjects have to listen to a jumbled series of numbers and letters being read to them at the approximate rate of one item per second. Participants are then asked to not only recall the numbers and letters heard in each set, but to also put them in alphabetical and numerical order. The entire test consists of eight blocks in which respondents start out with a block of three items and get three trials to recall and order the items correctly. The list length is increased by one for each successive block. After three missed trials, the test is terminated. Figure 2.12 shows a sample logfile of a random subject (not part of the VILLA project) and includes the presented sequences, the answers entered by the respondent and the status of the answers (true or false).

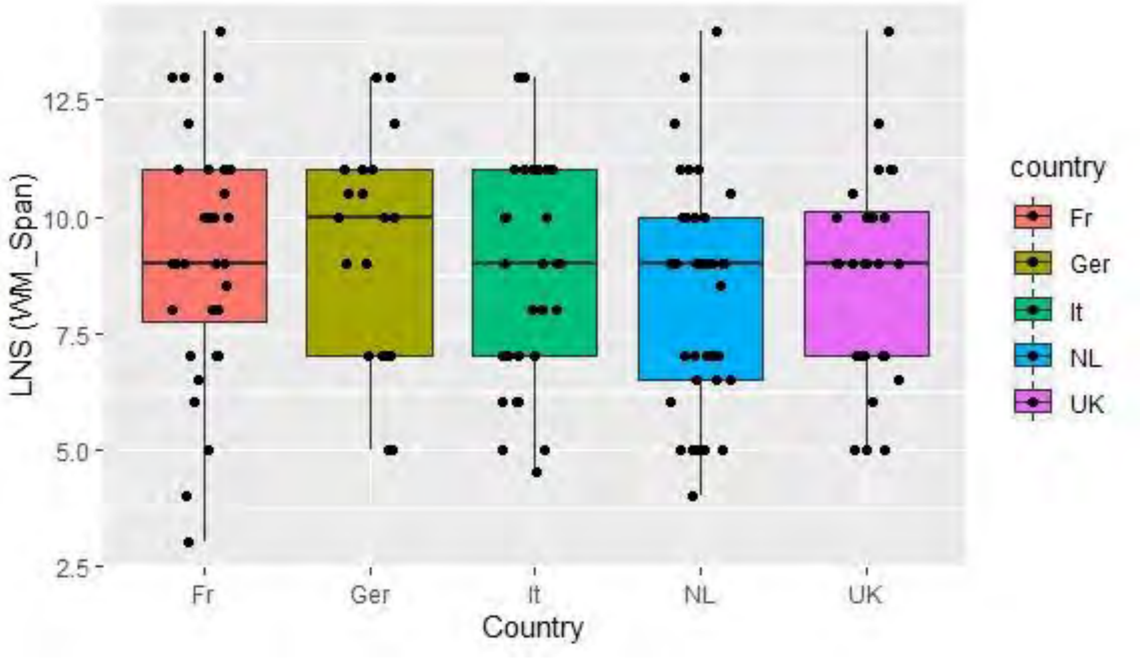
Figure 2.12 Example LNS task



```
July 17 2021 15:52:42   md13    LOGGED_IN       1626551562339   1       Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14
Safari/605.1.15
July 17 2021 15:53:42   md13    SOUND_DONE      1626551622696   1       2J
July 17 2021 15:53:48   md13    ANSWER_SUBMITTED        1626551628135   1       2J      2J      j2      false   1
July 17 2021 15:53:53   md13    SOUND_DONE      1626551633158   2       C6
July 17 2021 15:53:58   md13    ANSWER_SUBMITTED        1626551638198   2       C6      6C      6c      true    1
July 17 2021 15:54:03   md13    SOUND_DONE      1626551643796   3       F8
July 17 2021 15:54:09   md13    ANSWER_SUBMITTED        1626551649320   3       F8      8F      8f      true    1
July 17 2021 15:54:15   md13    SOUND_DONE      1626551655149   4       4G1
July 17 2021 15:54:22   md13    ANSWER_SUBMITTED        1626551662081   4       4G1     14G     14g     true    0
July 17 2021 15:54:27   md13    SOUND_DONE      1626551667662   5       J7O
July 17 2021 15:54:34   md13    ANSWER_SUBMITTED        1626551674042   5       J7O     7JO     07j     false   1
July 17 2021 15:54:40   md13    SOUND_DONE      1626551680289   6       9S4
July 17 2021 15:54:43   md13    ANSWER_SUBMITTED        1626551683672   6       9S4     49S     49s     true    1
July 17 2021 15:54:50   md13    SOUND_DONE      1626551689846   7       5W3L
July 17 2021 15:55:03   md13    ANSWER_SUBMITTED        1626551703169   7       5W3L    35LW    35lw    true    0
July 17 2021 15:55:10   md13    SOUND_DONE      1626551709970   8       2Y8E
July 17 2021 15:55:16   md13    ANSWER_SUBMITTED        1626551716669   8       2Y8E    28EY    28ey    true    0
July 17 2021 15:55:23   md13    SOUND_DONE      1626551723281   9       A8C1
July 17 2021 15:55:30   md13    ANSWER_SUBMITTED        1626551729862   9       A8C1    18AC    18ac    true    0
July 17 2021 15:55:37   md13    SOUND_DONE      1626551737067   10      9B7H2
July 17 2021 15:55:56   md13    ANSWER_SUBMITTED        1626551756433   10      9B7H2   279BH   289ab   false   1
July 17 2021 15:56:03   md13    SOUND_DONE      1626551763730   11      P4H6K
July 17 2021 15:56:20   md13    ANSWER_SUBMITTED        1626551780560   11      P4H6K   46HKP   268pk   false   2
July 17 2021 15:56:27   md13    SOUND_DONE      1626551787713   12      3I1QM
July 17 2021 15:56:40   md13    ANSWER_SUBMITTED        1626551800072   12      3I1QM   13IMQ   13iqm   false   3
July 17 2021 15:57:42   md13    LOGGED_IN       1626551862676   1       Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.
Safari/605.1.15
```

The score is traditionally computed by adding the total number of correct trials (Shelton et al, 2009).
Talkbank produced two scores[11] for the LNS-task; the WM_Span[12] (represented in Figure 2.13), which is the number of items in the last correctly sequenced series and the WM_Score[13] (represented in Figure 2.14), which is an automatic point score that keeps track of the series number. The latter is problematic as it allows subjects with various wrong answers to still get the maximum score of 8 (VILLA Field Manual, 2022).
Pearson's r was calculated and rendered a relatively weak correlation (r= 0.608) between the two automatic scores, indicating that there might be a difference in the two scores.

Figure 2.13 Results WM_Span (LNS-task) per country (N= 90, with 72 missing values)



Data distribution

Looking at the boxplots in figure 2.13, we can see that the distribution both within and across the countries is rather diverse. All countries, except the Netherlands show a symmetric distribution of their data, indicating an equal spread in

[12] The WM_Span has 72 missing values, the WM_Score only 42! An explanation to this difference had not been found at the time of writing this thesis. Roeland suggested that the WM_score might not be given if it is identical to the WM_Span.

[13] Looking at the data, in one instance, a result of WM_Span is given, but there is a missing result in WM_Score. This is in contrast with the other missing values where it is the other way around. Was this done by accident, or was this done in a systematic way?

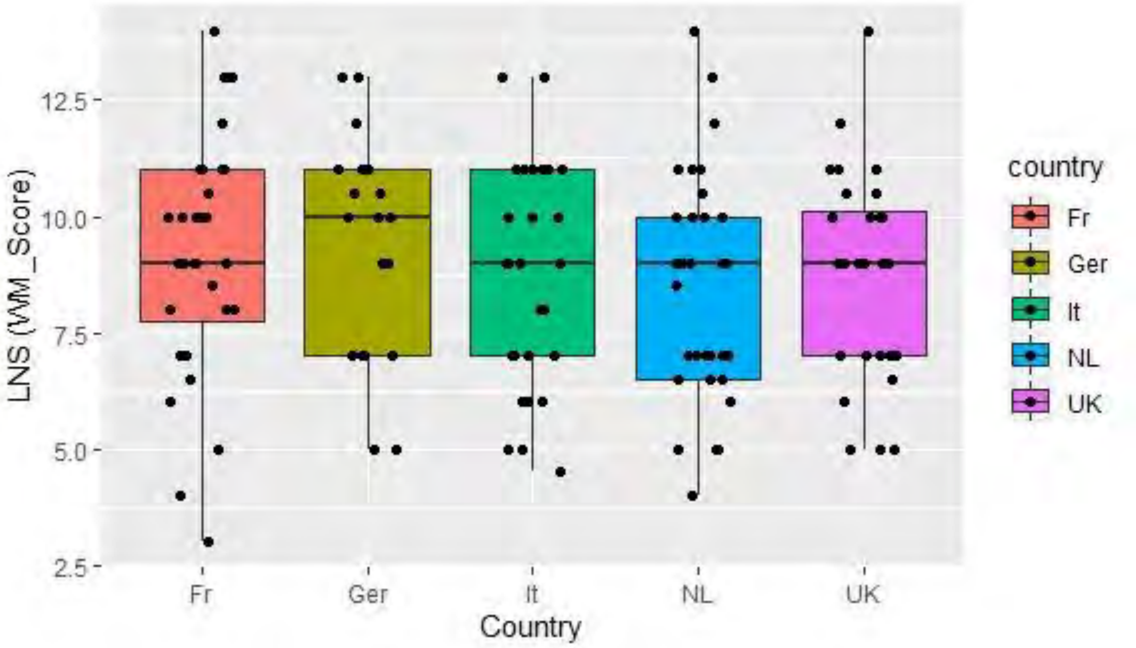
the way their respondents scored on this test. Judging by the cropped boxplot, the German participants had a high level of internal agreement, with 50% of the students scoring between 4.75 and 5.25. The Dutch data stick out due to the fact that the country had many participants with a score of 5, but also because contrary to the other countries the data seem to be skewed towards the lower scores. he Netherlands and Germany are the only countries with outliers. The Netherlands has two and Germany only one.

Interestingly, except for France, the median of all countries lies at a score of 5, indicating that most participants across the countries managed to recall and correctly order a series of at least5 jumbled letters and digits. With a median of 4, the French respondents do worse in that respect. The best-performing country on the LNS-task is Italy, with 25% of its test takers scoring within the range of 5 and 6 points and maximum scores of 7.

Statistical significance

An ANOVA test was run to find out if there were any significant effects between the countries on the variable "WM_Span", and this turned out not to be the case (F(4, 85)= 1.334; p<0.264).

Figure 2.14 Results WM_Score(LNS-task) per country (N= 122 with 40 missing values)



Data distribution

Looking at the boxplots in figure 2.14, we can see that the UK sticks out in the way their data is distributed. Whereas the data of all other countries seem to be skewed toward the lower scores, the British data is skewed towards the upper scores. Moreover, the UK only had one participant scoring at ceiling, while 50% of the respondents in the other countries managed to get that score.Most internal consistency can be found in Germany where 75% of the candidates scored 7, or at ceiling.

The fact that across countries, many students scored at ceiling, seems to support the concerns expressed with regard to this method of scoring, making the validity of the results of this test questionable. Judging by the median (M= 7), all countries, except the UK, seem to perform equally well on this test.
Germany is the only country with an outlier.

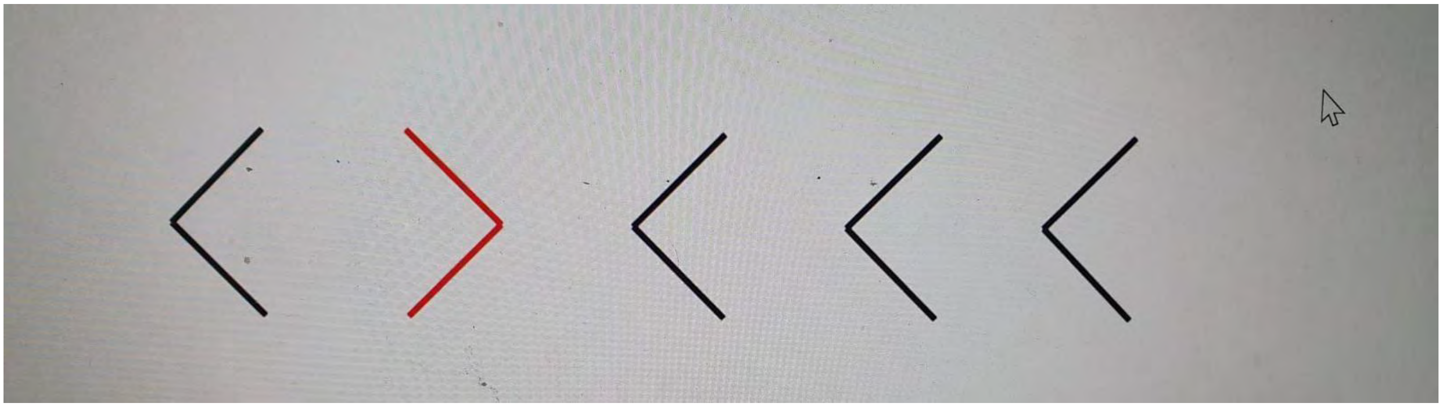
Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable "WM_Score", and this turned out not to be the case (F(4, 117)= 1.666; p<0.162).

2.3.2.3 Executive Functions: Eriksen Flanker Task (6)

In the task, participants have to look at a central target that is surrounded by distracting stimuli and give a directional physical response by pushing a left or right button. Figure 2.15 gives an example of an Eriksen Flanker Task item.

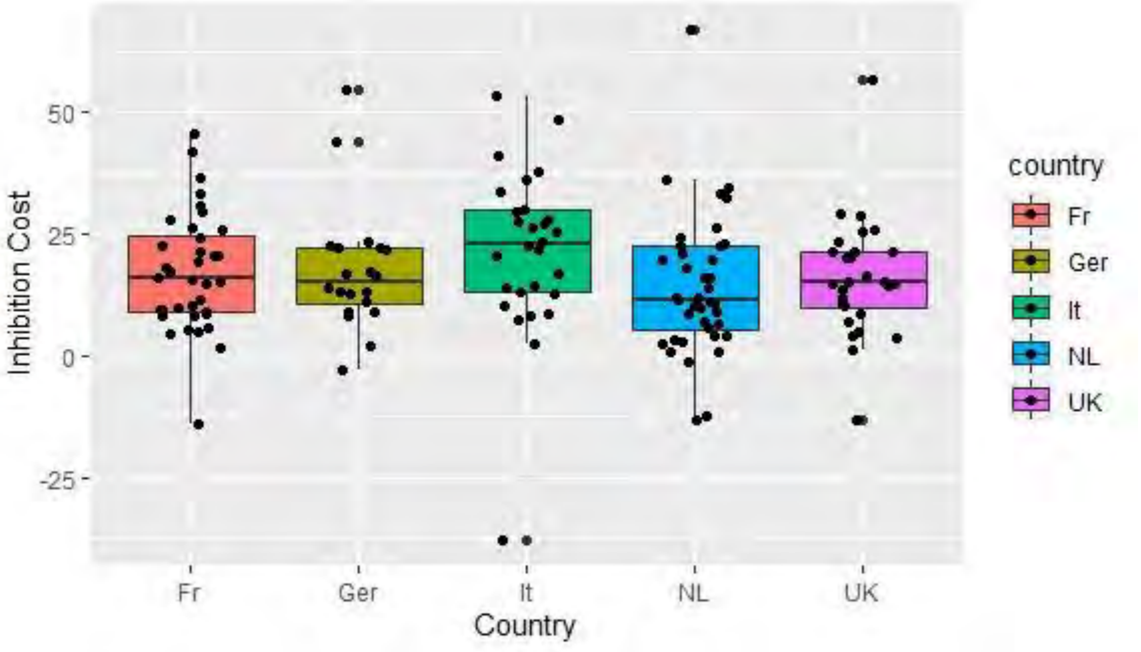
Figure 2.15 Example item Eriksen Flanker Task



Three different types of stimuli are used: congruent, incongruent and neutral (Lamers & Roelofs, 2011) Congruent, or compatible stimuli are distractors, or flankers that call for the same response as the target and therefore may facilitate the process of coming to the correct response. Incongruent, or incompatible stimuli are distractors that call for a different response as the target and may therefore hinder the process of coming to the correct response. Neutral stimuli are distractors that do not have any impact on the process of coming to the correct response.

Incongruent stimuli create conflict that need to be resolved and as a consequence, responses in incongruent trials are generally slower than those in congruent trials (Costa et al., 2009). Inhibition Cost is a measure of the extra processing time needed to suppress irrelevant information in order to give the correct response. It is calculated by subtracting the reaction time (RT) measured in the congruent trials from the RT measured in the incongruent trials. A low inhibition cost is an indication of great inhibition skills. Figure 2.16 gives a visual representation of the inhibition cost per country.

Figure 2.16 Results Eriksen Flanker Task per country (N= 154, with 8 missing values)



Note

Two values were too extreme to be included in this analysis and were therefore omitted and marked as missing values.

This was the case for the Italian participants 5105 and 5111 with values of −499.3 and 160.8 respectively. The reason remains unclear.

Data distribution

Across the board, as can be seen in Figure 2.16, countries show a fairly symmetric distribution of their data, with only Italy and the Germany showing some skewedness; Italy towards the higher scores, and Germany towards the lower scores.Judging by the slim boxes, most internal agreement can be found in Germany and the UK, both countries with a median of around 18.

All countries except France had outliers. Germany had four, two at around 31 and two at around 56. Italy had two at around -38, the Netherlands had two at around 69 and the UK had four, two at around -13 and two at around 56.

Two countries stick out when it comes to inhibition cost: Italy and the Netherlands. Whereas France, Germany and the UK all had identical medians of around 18, these two countries had deviating medians. Judging by the elevated median of close to 25, the Italian participants clearly had more issues with inhibition than the candidates of the other countries. The Dutch respondents seem to be best performing in this respect (M=12.5).

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Working Memory, and this turned out not to be the case (F(4, 147)= 1.074; p<0.372).

2.3.3 Perceptual Preference: Barsch Learning Style Inventory (7)

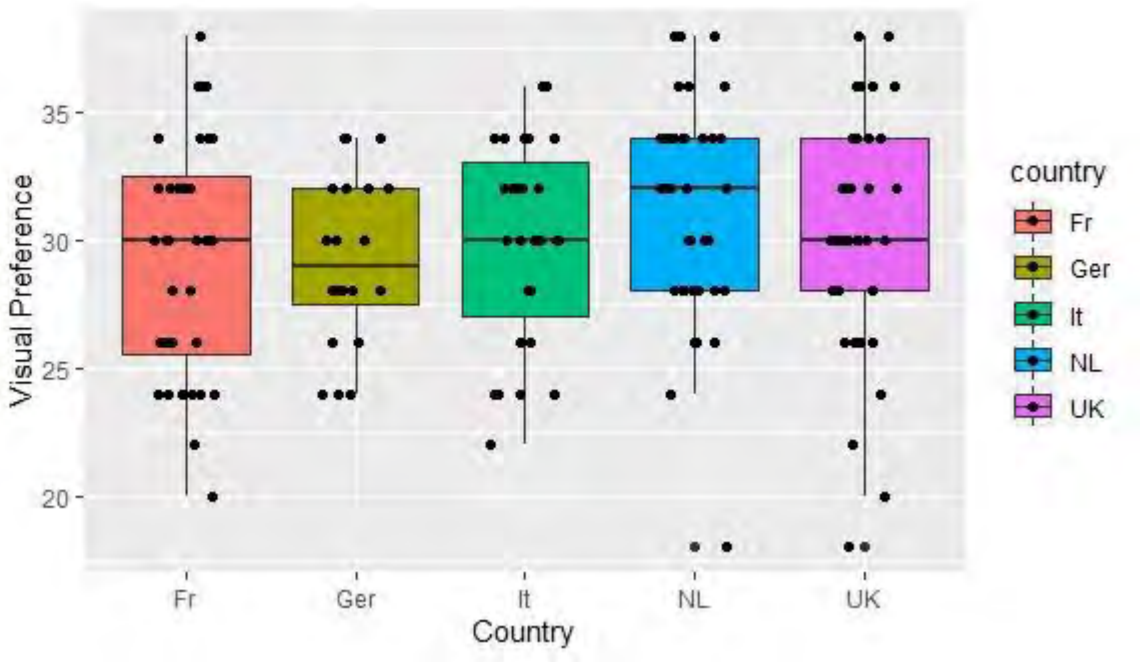
Perceptual preference was assessed using the *Barsch Learning Style Inventory* (Barsch,1991).
It is thought that one of the ways in which individuals differ is in the sense modality (visual, auditory and kinesthetic), from which they best absorb, retain and process new information (Cassidy & Eachus, 2000; Dunn, 1983; Harrison et al. 2003). More specifically, input provided in a learner's preferred learning modality could optimize learning (Zapalska & Dabb, 2002).
Researchers Dunn (1993), Zapalska & Dabb (2002) hypothesize that:

> A person who is a visual learner needs to see, observe, record, and write to best learn.
> An auditory learner prefers information that is spoken and heard, as it is in dialogue and discussion.
> A kinesthetic learner prefers to learn in an environment where material can be touched and he or she can be physically involved with the to-be-learned information.

Barsch (1991) created a relatively simple questionnaire designed to measure an individual's perceptual preference. It consists of 24 statements, eight for each modality, and respondents have to indicate how frequent they feel a statement applies to them: often, sometimes, seldom. Upon completion of the questionnaire, respondents fill in the scoring sheet and attribute the points to the corresponding sentences (often = 5, sometimes = 3, and seldom= 1). As there are eight statements for each modality, scores per modality range from 0 to 40, where a high score indicates a strong preference for that modality.

The entire questionnaire, as well as some additional information about the three primary senses used to take in information can be found in Appendix A. In Figures 2.17, 2.18 and 2.19 the results per perceptual preference per country can be found.

Figure 2.17 Results Barsch LSI, Visual Preference (N= 162 with 0 missing)



<u>Data distribution</u>
Looking at Figure 2.17, we can see more or less similar distributions of the data across the countries with boxes that are all round about the same size, indicating that the variation in scores within and across the countries is fairly similar.
The variation in spread is most salient in France and the UK with scores ranging from 20 to 35. The two countries show an almost identical distribution of their data with a median of 30, but whereas France seems to be slightly more skewed

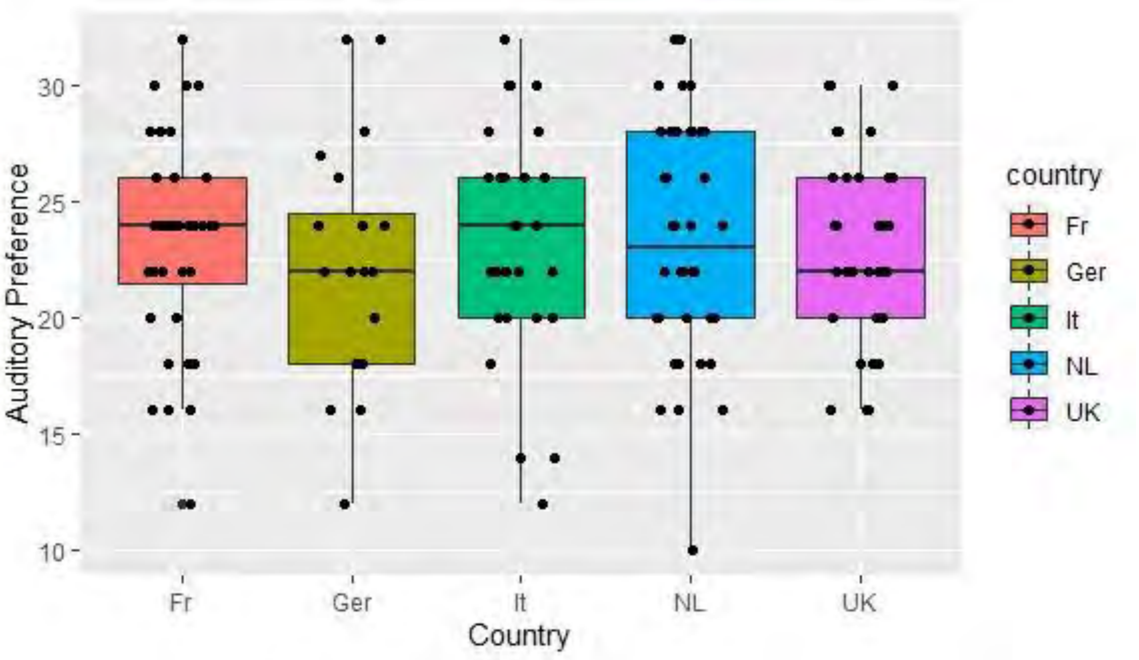
towards the lower ranges, the UK leans heavier towards the upper ranges. Germany, with a relatively cropped box, is the country with most agreement in the way their participants scored.

With one each, the Netherlands and the UK are the only countries outliers.

Across the board countries seem to score high on the preference for visual perception, with 50% of their respondents scoring around 29 or higher on this test. The Netherlands stick out when it comes to preference for the visual modality, with 75 % of its participants scoring around 28 or higher, a median of around 32 and, disregarding the one outlier, a minimum score of around 24. Looking at the median only, we could say that the preference for the visual modality wasn't as strong with the German participants as it was for the students of the other countries. However, the German sample was half the size of that of the other countries and there was more internal agreement amongst the scores within the country, indicating that more students had similar scores.

Statistical significance
An ANOVA test was run to find out if there were any significant differences between the countries on the variable Grammatical Inferencing, and this turned out not to be the case (F(4, 157)= 0.973; p<0.424).

Figure 2.18 Results Barsch LSI, Auditory Preference (N= 162, with 0 missing values)

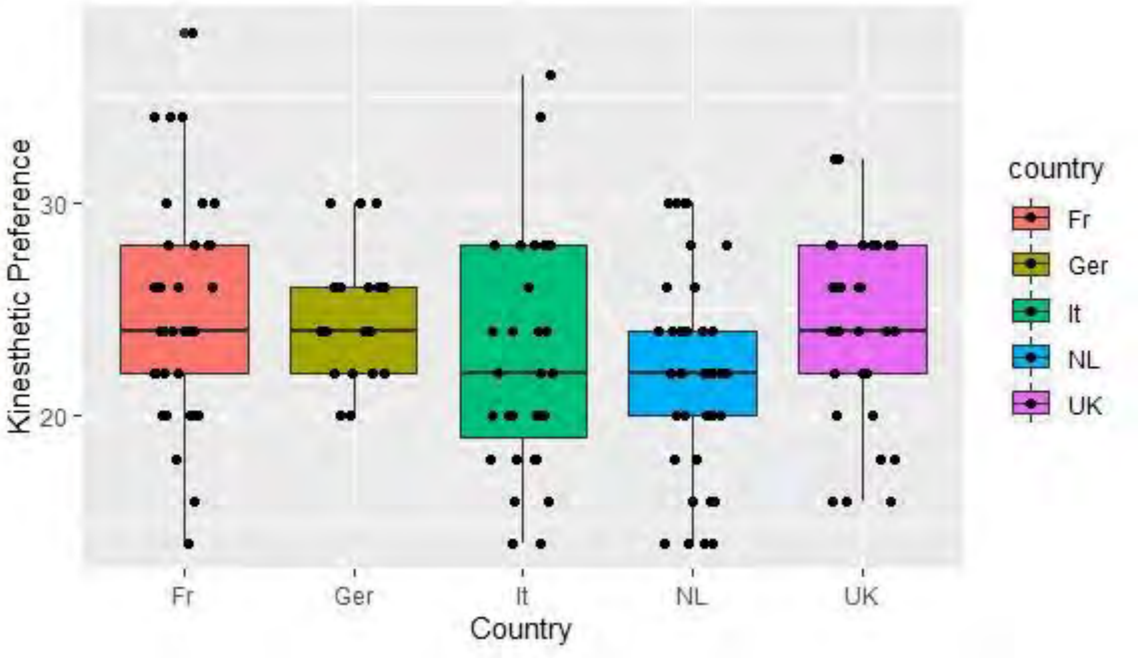


Data distribution
Looking at Figure 2.18, with the exception of France and the UK, we can see relatively tall boxes and long whiskers, indicating quite some variance in the way the respondents scored on Auditory Preference. France and the UK show more symmetry in that respect, which means there is more agreement among the participants in the importance they attribute to auditory perception The largest variance can be found in the Netherlands with scores ranging from 10 points to 32.France is the only country with an outlier.

With a median of 24, the French and Italian candidates seem to display the strongest preference for the auditory modality, closely followed by the Dutch with a median of 23. The students from Germany and the UK seem to favor this modality the least, with a median of 22.

Statistical significance
An ANOVA test was run to find out if there were any significant differences between the countries on the variable Grammatical Inferencing, and this turned out not to be the case (F(4, 157)= 0.27; p<0.897),

Figure 2.19 Results Barsch LSI, Kinesthetic Preference (N= 162, with 0 missing values)



<u>Data distribution</u>

Figure 2.19 shows great variety in the way countries attribute importance to the kinesthetic modality. Most variance can be found in Italy with scores ranging from around 14 points to 36 and a relatively tall box, indicating that there was quite some variety in the way the Italian participants scored on Kinesthetic Preference.

France and the UK show almost entirely similar data distributions with equally large boxes and an identical median of around 24.The only difference being that France has longer whiskers, indicating that at least two of the French participants showed extremes on each side of the box.

The countries with the highest internal consistency are Germany and the Netherlands. Both countries show a relatively cropped box that is about the same size. The German subjects however seem to overall show a higher preference for the kinesthetic modality compared to the Dutch candidates.

France is the only country with an outlier.

If we take the median as a measure of preference for kinesthetic perception, France, Germany and the UK, all with a median of 24, rely on this modality most. Italy and the Netherlands, both with a median of 22 seem to favor this modality the least. However, the Italian respondents seem to display a greater variety in the way they scored with a maximum score of 33 as opposed to the Dutch participants, whose maximum score on this modality was 30.

Across the board, 50 % of the participants of the various countries score about 23 points or more which is still a decent score considering the fact that the maximum score for this modality is 40

<u>Statistical significance</u>

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Kinesthetic Preference, and a significant effect was indeed found (F(4, 157)= 2.629; p<0.0365).

A subsequent Tukey post hoc test showed a near-significant difference between France and the Netherlands (t=2.746; p<0.052) indicating that the French candidates attribute more importance to the Kinesthetic modality than the Dutch do.

2.4 Tests: Language aptitude

2.4.1 The LLAMA suite (9-11)

The LLAMA test battery assesses a candidate's underlying cognitive ability (Granena, 2011).
It was developed by Meara (2005) and consists of a number of subtests that are all, with the exception of LLAMA D, based on Caroll's (1981) four-factor structure of language aptitude and Caroll & Sapon's (1959) MLAT tests. The LLAMA D test is inspired by research done by Service (1992), Service & Kohonen (1995) and Speciale et al. (2004).
The LLAMA suite consists of the following subtests: Associative Memory (LLAMA B), Phonemic Coding Ability (LLAMA D), Inductive Language Analytic Ability (LLAMA E) and Grammatical Sensitivity (LLAMA F) (Artieda & Muñoz, 2016).
Table 2.7 below, gives an overview of the entire LLAMA suite along with the underlying cognitive abilities that it tests. The VILLA project however, only used the subtests LLAMA B,D and F.

Table 2.7 Purported aptitude abilities captured by the LLAMA test suite.

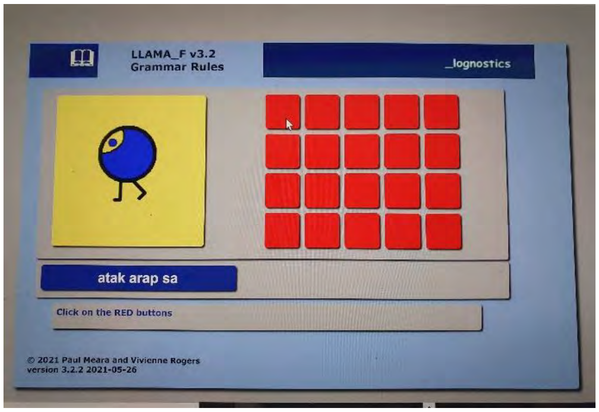
LLAMA test	Underlying cognitive ability
	Analytic learning ability**:**
LLAMA B *Vocabulary learning* LLAMA E *Sound-symbol correspondence* LLAMA F *Grammatical inferencing*	- gained by linguistic experience in one's L1 - allows for strategy use and problem solving techniques - learning happens by analysis - equated to explicit learning aptitude
	Sequence learning ability:
LLAMA D *Sound recognition*	- discovery of language structure by by detecting statistical properties in input - learning is unintentional and uncontrolled, and happens by analogy - equated to implicit learning aptitude

Adapted by Artiede et al. (2016) from Granena (2011).

2.4.2. Language Aptitude: Grammatical Inferencing (9)

The subtest LLAMA F was used to assess the subject's grammatical inferencing skills.
The test consists of two parts: the training phase and the test phase. During the training phase, participants are exposed to sentences in a new language that are accompanied by a graphic illustration of their meaning (see Figure 2.20) from which subjects have to work out the grammatical rules. Knowledge of the newly acquired grammar is then assessed during the test phase (see Figure 2.21) where candidates have to decide on the morpho-syntactic acceptability of new sentences in the same language. All candidates received the instructions in their L1 (VILLA Field Manual, 2022).
The results of the LLAMA F subtest are visualized in Figure 2.22.

Figure 2.20 Screenshot training phase LLAMA F

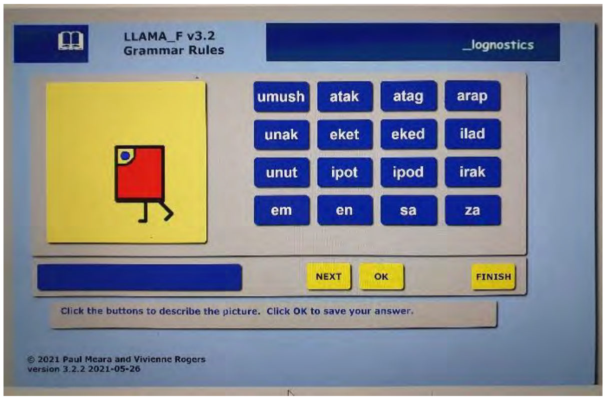


atak arap sa

Click on the RED buttons

© 2021 Paul Meara and Vivienne Rogers
version 3.2.2 2021-05-26

Figure 2.21 Screenshot testing phase LLAMA F



Click the buttons to describe the picture. Click OK to save your answer.

© 2021 Paul Meara and Vivienne Rogers
version 3.2.2 2021-05-26

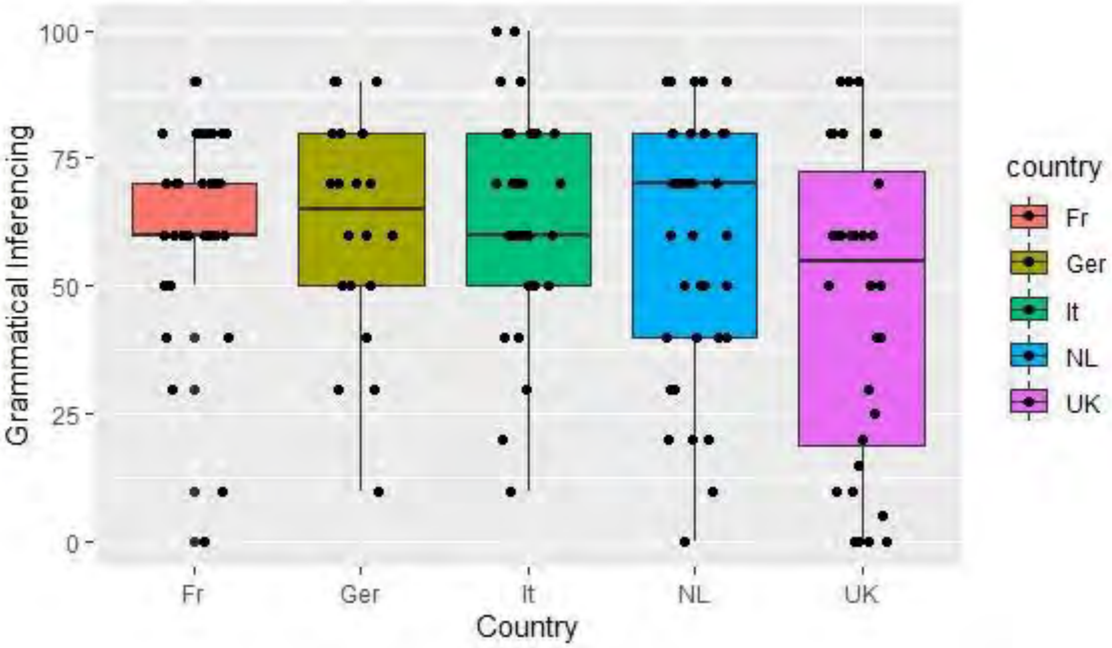
The illustrations make the test language independent, but also come with some limitations. As a result, the LLAMA F subtest relies more on agreement features than on word order (Artieda & Muñoz, 2016; Meara, 2005).

Scores for LLAMA F range between 0 and 100 and should be interpreted as in Table 2.8

Table 2.8 Key LLAMA F (Meara, 2005)

Scoring range	Interpretation
0-15	a very poor score, probably due to guessing
20-45	an average score, most people score within this range
50-65	a good score
75-100	an outstandingly good score. Few people manage to score in this range

Figure 2.22 Results LLAMA F, Grammatical Inferencing (N= 158, with 4 missing values)



Data distribution

Looking at the boxplots in figure 2.22, we can see that there is quite some variance both within and between the countries in the way the participants scored on the LLAMA F test. The medians of all countries are all very different with France having the lowest median of around 59, which according to Meara's (2005) key is still a good score, and the Netherlands being the best scoring country with a median of around 69, which according to Meara's (2005) key lies on the upper side of that very same range.

If we study the variance within the countries, we can see extended box sizes below the median in the Netherlands and the UK, indicating a large diversity of scores with the UK clearly having the largest selection of participants that score below the median of 56. Contrary to Germany and Italy that seem to display a diversity of individual scores in the scores above the median. France stands out due to a number of reasons. First of all, it is the only country with any outliers, and it counts five of them. Second, if we disregard the outliers, it has an extremely high minimum score compared to the other countries. And thirdly, it seems to have a fairly large number of participants with scores similar to the median or higher. Moreover, looking at the box size, there is not a lot of variance in the individual scores. France is the only country with any outliers at five different positions.

Across the board, with the exception of France, countries display a diverse spread when it comes to Grammatical Inferencing, indicating some variety in the way individual participants performed on this test. If we go by the median as a measure of success, the Netherlands seems to be the best scoring country. The distribution of the scores of the Dutch participants however is rather skewed towards the lower ranges and in this respect, France seems to be the best performing country, as it shows a symmetric distribution and a small box with scores in the upper ranges, indicating a large agreement amongst the scores of its subjects. Disregarding the five outliers, the French participants did not score below 50 with many students scoring 58 or higher.

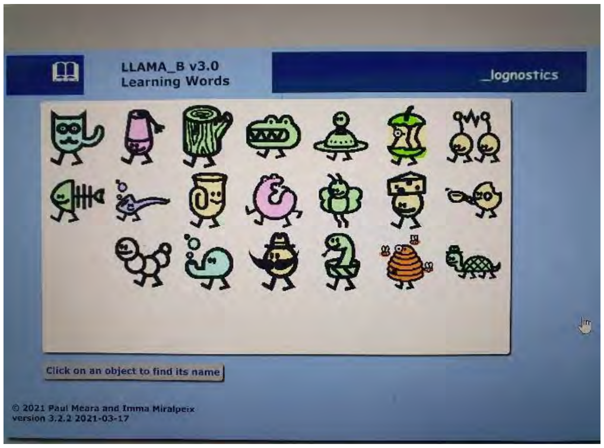
Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Grammatical Inferencing, and this turned out not to be the case (F(4, 153)= 2.422; p<0.051).

2.4.3 Language Aptitude: Word Learning Skill (10)

The subtest LLAMA B was used to assess word learning skill.
The test consists of a training phase, in which subjects get two minutes to memorize 20 pseudowords presented as the names of various fantasy creatures (see Figure 2.23), and a test phase in which candidates need to make new form-meaning associations and are assessed on how well they do this (see Figure 2.24). All candidates received the instructions in their L1 (VILLA Field Manual, 2022).
Results of the subtest LLAMA B can be found in Figure 2.25.

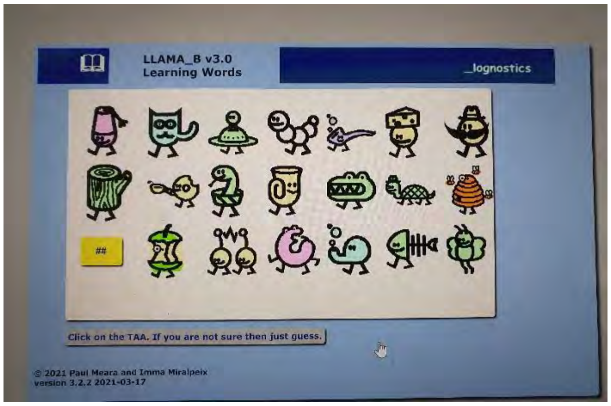
Figure 2.23 Screenshot training phase LLAMA B



Click on an object to find its name

© 2021 Paul Meara and Imma Miralpeix
version 3.2.2 2021-03-17

Figure 2.24 Screenshot testing phase LLAMA B



Click on the TAA. If you are not sure then just guess.

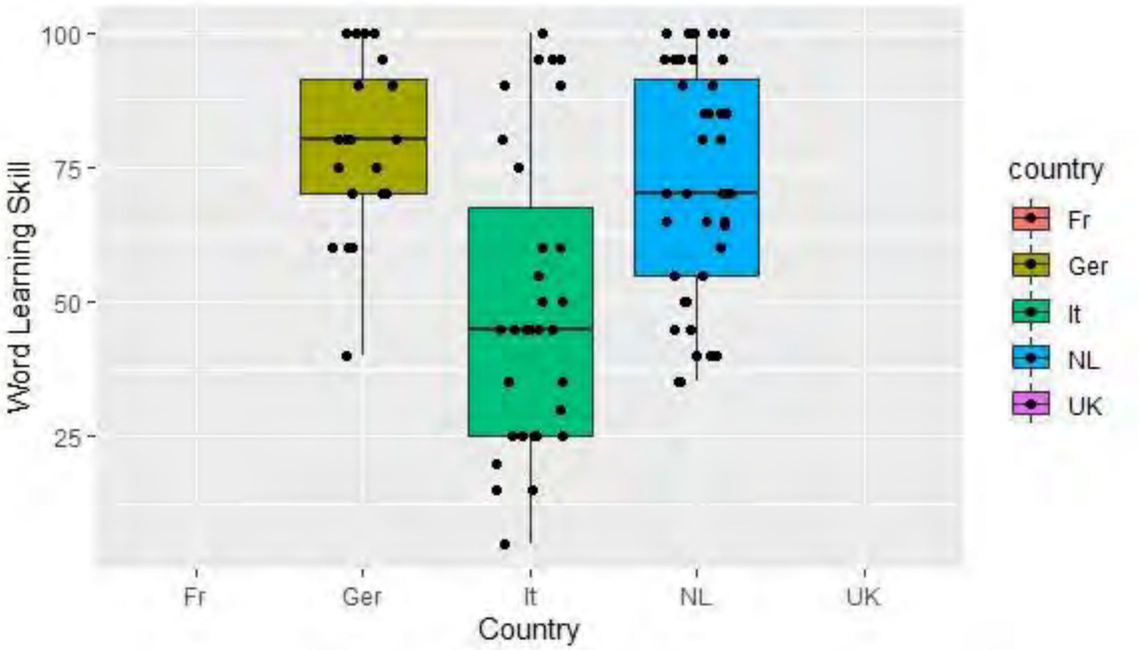
© 2021 Paul Meara and Imma Miralpeix
version 3.2.2 2021-03-17

Candidates score five points for each object correctly identified. There is no correction for guessing (Rogers et al., 2017). Scores for LLAMA B range between 0 and 100 and should be interpreted as in Table 2.9.

Table 2.9 Key LLAMA B (Meara, 2005)

Scoring range	Interpretation
0-20	a very poor score, probably due to guessing
25-45	an average score, most people score within this range
50-70	a good score
75-100	an outstandingly good score. Few people manage to score in this range, unless they are using a formal mnemonic system.

Figure 2.25 Results LLAMA B test, Word Learning Skill, per country. (N= 91, with 71 missing values)



Note: The data set had 71 missing values that were not included in Figure 2.25.
This large number of missing values is mainly due to the fact that the test was not administered in France and the UK.

Data distribution

Looking at the boxplots in Figure 2.25 we can see there is quite some variance in the scores, both between the countries and within. Italy seems to be the country with most variation in the scores, with scores ranging from 6 to the maximum score of 100. There was more agreement amongst the German participants judging by the fairly compact box and the symmetric distribution within the box.
The Dutch subjects seem to do better than the Italian students with scores ranging from 36 to 100 even though scores vary quite a bit within the country.
No country has any outliers.

All countries seem to have a couple of participants who managed to get the maximum score on this test, which according to Meara's (2005) interpretation is highly exceptional. Germany clearly is the best performing country when it comes to word learning. Fifty percent of the German subjects score 81 or higher on this test, falling in the range which Meara (2005) labelled "an outstandingly good score". Moreover, the country has a high level of agreement amongst its participants, indicating that many students had similar high scores. Even the minimum score of 39 is still higher than that of other countries. Having said this, Germany, with about half the number of adult subjects, did have a small sample size.
The second-best performing country is the Netherlands with a median of around 68 and a minimum score of around 36.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Word Learning Skill, and a significant effect was indeed found (F(4, 88)= 12.59; p=0.002).

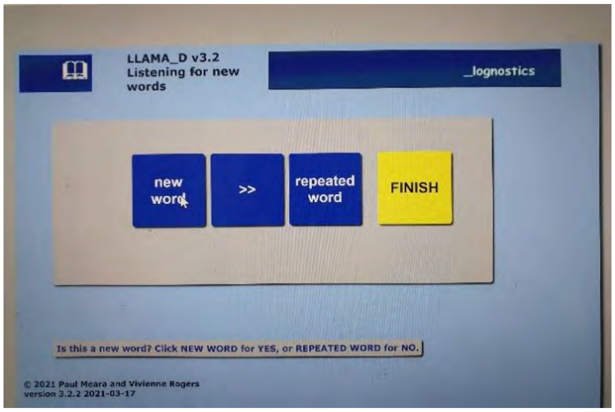
A subsequent Tukey post hoc test showed significant differences between Germany and Italy (t=4,432; p=0.000) and Italy and the Netherlands (t=4,159; p= 0.000), indicating that there was a significant difference in word learning skills between the Italian students and both the Dutch and the German respondents in the sense that the Dutch and the German participants outperformed the Italian learners.

2.4.4 Language Aptitude: Phonological Recognition (11)

The subtest LLAMA D is a measure of Phonological Recognition. The subtest is loosely based on work from researchers Service (1992), Service & Kohonen (1995) and Speciale et al. (2004) who suggest that the ability to recognize oral patterns contributes to language learning as it is involved in acquiring vocabulary and signaling grammatical features (Artieda & Muñoz, 2016; Meara, 2005).

In the test, subjects are exposed to a number of orally presented words in a new language and have to decide whether they think it is a new word or a repeated word. Figure 2.26 shows a screenshot of this test. All candidates received the instructions in their L1. (VILLA Field Manual, 2022). Results of the subtest LLAMA D can be found in Figure 2.27.

Figure 2.26 Screenshot LLAMA D



Is this a new word? Click NEW WORD for YES, or REPEATED WORD for NO.

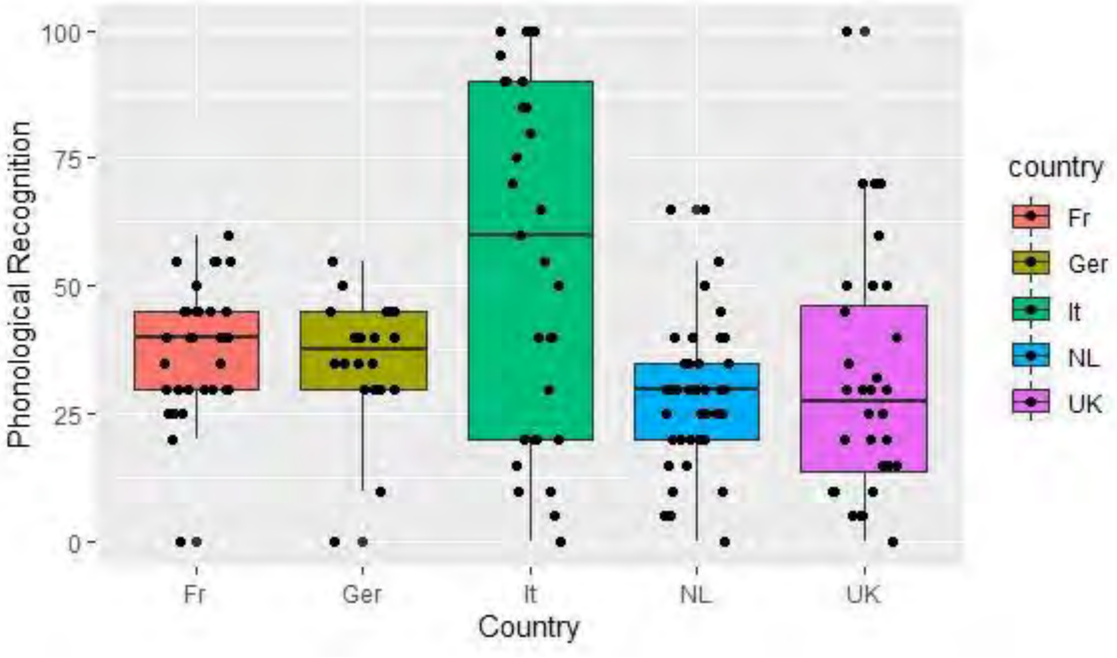
© 2021 Paul Meara and Vivienne Rogers
version 3.2.2 2021-03-17

Scores for LLAMA D range between 0 and 100 and should be interpreted as in Table 2.10.

Table 2.10 Key LLAMA D (Meara, 2005)

Scoring range	Interpretation
0-10	a very poor score, probably due to guessing
15-35	an average score, most people score within this range
40-60	a good score
75-100	an outstandingly good score. Few people manage to score in this range.

Figure 2.27 Results LLAMA D test, Phonological Recognition, per country. (N= 159, with 3 missing values)



Data distribution

Looking at figure 2.27, Italy immediately jumps out. The tall box indicates great variety in scores of its participants with students performing very poorly on this test to students who excel. Italy is also the only country without any outliers. Distribution-wise, the country shows a fairly symmetrical spread with about as many students scoring above as below the median of around 60, which is way higher than those of the other countries. France and Germany seem to have similar distributions with a fairly large internal consistency in scores, a median of around 38 and a symmetric spread. Apart from the fact that the median is quite a bit lower, the distribution of The Netherlands seems to bear strong resemblance to that of France and Germany. The variance shown in the UK is not as great as in Italy, but clearly bigger than in the other countries with scores ranging from 0 to 69. All countries except from Italy have outliers. France and Germany haver outliers in the lower ranges, whereas the Netherlands and the UK have outliers in the upper ranges.

The Italian participants clearly outperform the others when it comes to Phonological Recognition. Fifty percent of the Italian subjects score around the median of 60 or higher, which according to Meara (2005) is a good to even outstandingly good score. Italy is the only country with scores that fall within the range of excellent and even the lower ranges are similar to the scores of other countries or higher. The worst performing countries are the Netherlands and the UK with 50% of their subjects scoring within the range of average or below. Disregarding the outlier, all of the French subjects scored within the ranges of average and good (Meara, 2005), whereas some of the German participants had a very poor score.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Phonological Recognition, and a significant effect was indeed found (F(4, 154)= 8.619; p<0.000). A subsequent Tukey post hoc test showed significant differences between France and Italy (t=−3.724; p<0.003), Germany and Italy (t= −3.455 ; p<0.007), Italy and the Netherlands (t= 5.431; p< 0.000), and the Netherlands and the UK (t=−4.749; p< 0.000), indicating that the Italian participants were significantly better at phonological recognition than the French, the Germans and the Dutch. Moreover, the British respondents performed significantly better than the Dutch participants.

2.5 Tests: Personality

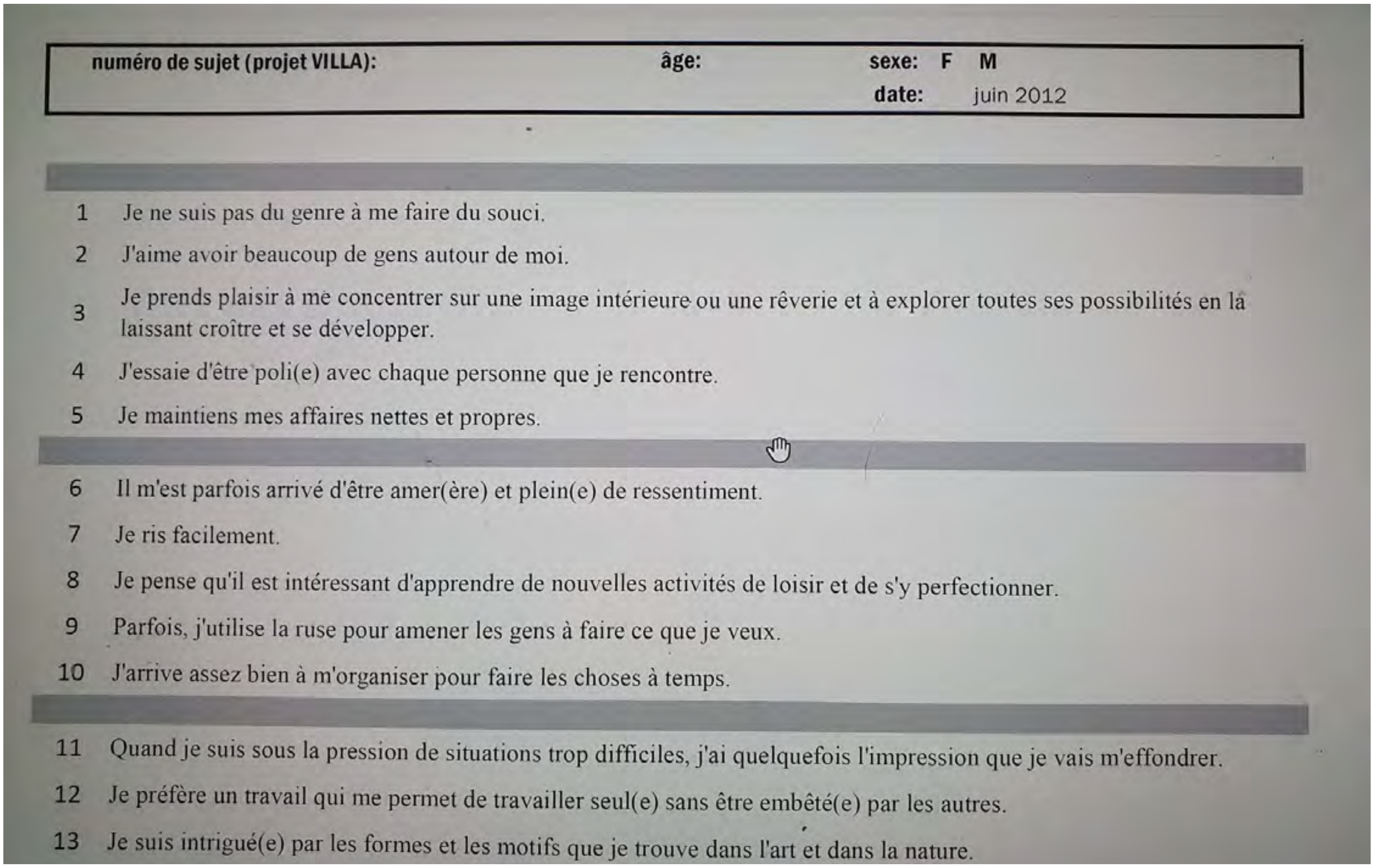
2.5.1 Personality (12-13)

Personality was documented through two different tests, NEO FFI—3 and ISALEM 97 measuring personality dimensions and experiential learning respectively. The fact that the VILLA project opted for these tests is no surprise considering the fact that personality and learning style are thought to be closely related (Ehrman & Oxford, 1995;Furnham et al., 1999; Ibrahimoglu et al., 2013), even though scientists still debate on the exact nature of their relation (Busato et al., 2000; Chamorro-Premuzic et al., 2007; Furnham, 1992; Jackson & Lawty-Jones, 1996;, von Wittich & Antonakis, 2011; Zhang, 2003).With regard to the relationship between personality and language learning, no real consensus has been reached and more research is necessary. There is research concluding that the lack of any significant findings indicates that a direct relation between personality variables and language acquisition does not seem very probable (Lalonde & Gardner, 1984), whereas other studies show a clear relationship between personality and language achievement (Ehrman & Oxford., 1995; Macintyre & Charos,1996)

2.5.2 Personality: NEO FFI-3 (12)

Personality traits were assessed using the NEO FFI-3, the adult version of the shortened *NEO Five Factor Index* (McCrae & Costa, 1987). The test consists of 60 questionnaire-type items, subdivided in 12 blocks in which each of the five personality factors: neuroticism, openness to experience, extraversion, agreeableness, and conscientiousness is represented. An excerpt of this test can be found in Figure 2.28.

Figure 2.28 Excerpt NEO FFI-3 questionnaire



numéro de sujet (projet VILLA): **âge:** **sexe: F M**
date: juin 2012

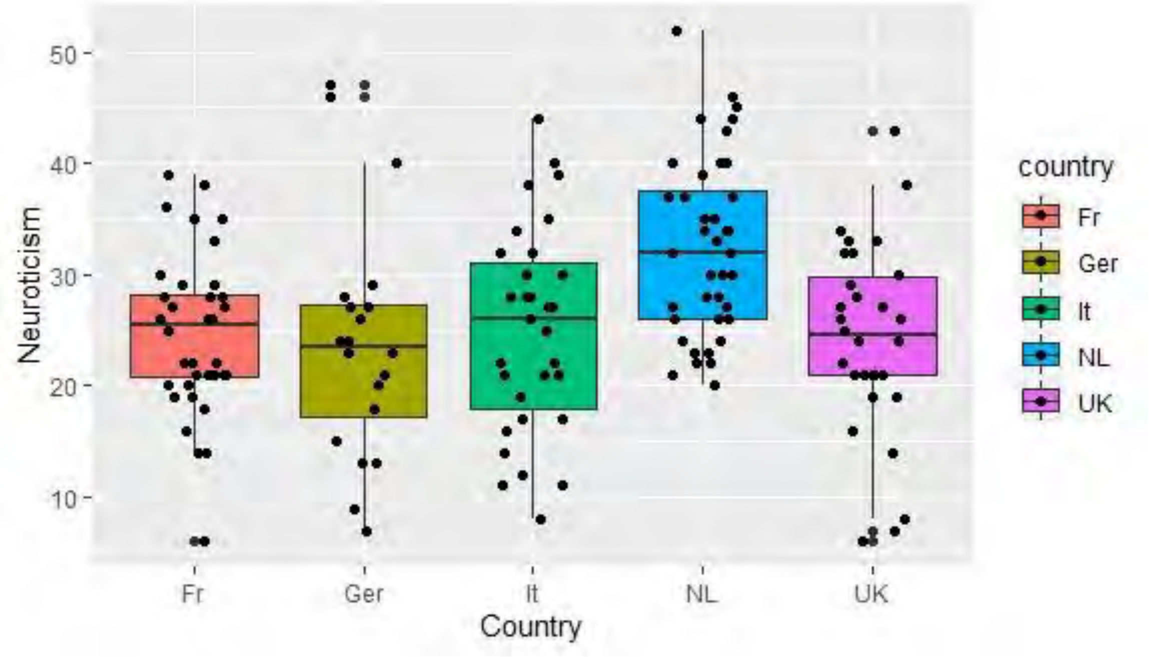
1 Je ne suis pas du genre à me faire du souci.
2 J'aime avoir beaucoup de gens autour de moi.
3 Je prends plaisir à me concentrer sur une image intérieure ou une rêverie et à explorer toutes ses possibilités en la laissant croître et se développer.
4 J'essaie d'être poli(e) avec chaque personne que je rencontre.
5 Je maintiens mes affaires nettes et propres.

6 Il m'est parfois arrivé d'être amer(ère) et plein(e) de ressentiment.
7 Je ris facilement.
8 Je pense qu'il est intéressant d'apprendre de nouvelles activités de loisir et de s'y perfectionner.
9 Parfois, j'utilise la ruse pour amener les gens à faire ce que je veux.
10 J'arrive assez bien à m'organiser pour faire les choses à temps.

11 Quand je suis sous la pression de situations trop difficiles, j'ai quelquefois l'impression que je vais m'effondrer.
12 Je préfère un travail qui me permet de travailler seul(e) sans être embêté(e) par les autres.
13 Je suis intrigué(e) par les formes et les motifs que je trouve dans l'art et dans la nature.

Participants were asked to rate the statements on the answer sheet using a Likert-type scale ranging from "totally disagree" to "totally agree". The answer sheets were then collected and scored manually with the aid of the scoring key leading to a total score for each of the five personality factors. In Appendix C both the answer sheet and the scoring key can be found. Figures 2.29, 2.30, 2.31. 2.32 and 2.33 give the visual representation of the results per personality factor per country.

Figure 2.29 Results NEO FFI-3, Neuroticism, per country. (N= 157, with 5 missing values[14])



Data distribution

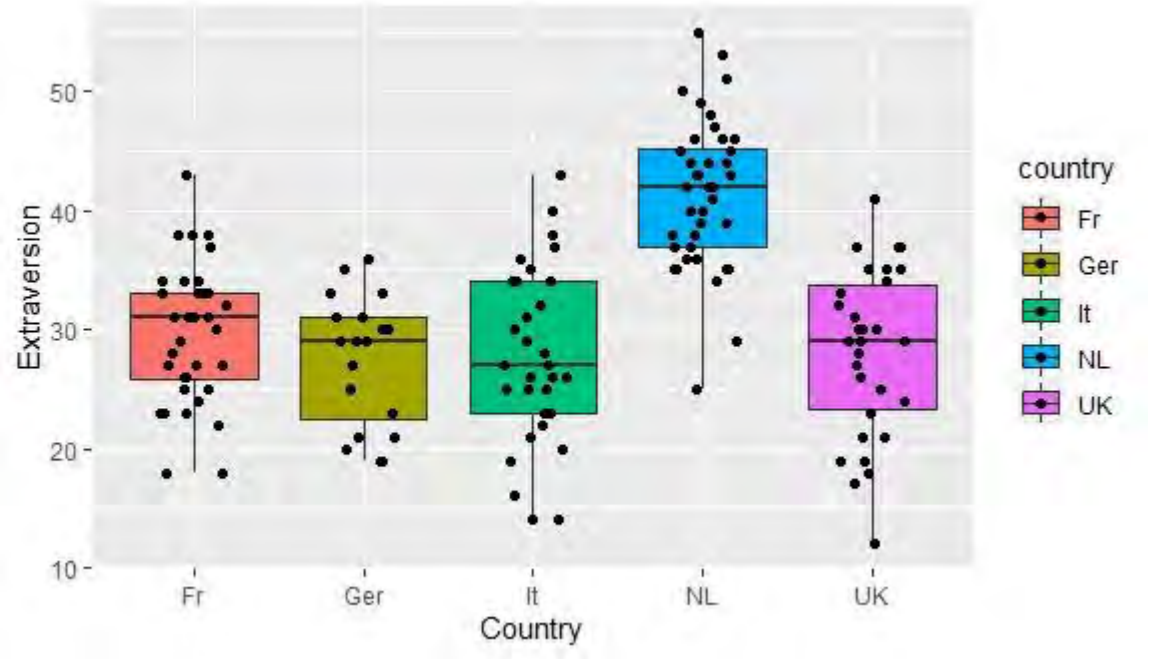
Looking at Figure 2.29, The Netherlands immediately jumps out as best performing country with its high median (32.5) and a relatively large level of internal agreement, judging by the size of the box. The data distribution seems to be slightly skewed towards the upper scores, with 50 % of the participants scoring in the range of 32.5 till 52.5 and none scoring less than 20 on the personality factor Neuroticism. France, Italy, and the UK have medians around 25 with France and Italy showing some skewedness towards the upper scores and the UK towards the lower scores giving the impression that the UK is the worst performing country on the variable Neuroticism. Germany bears resemblance to France, Italy and the UK, but has a lower median (23). Across the countries, France seems to have the greatest internal consistency and Italy seems to be the country with the most variance in the way its participants scored. Outliers can be found in France, Germany and the UK.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Neuroticism and a significant effect was indeed found (F(4, 152)= 6.057; p<0.000). A subsequent Tukey post hoc test showed significant differences between France and the Netherlands (t=−3.838; p<0.002), Germany and the Netherlands (t= −3.501; p<0.005), Italy and the Netherlands (t= −3.524; p< 0.005) and the Netherlands and the UK (t=3.858; p<0.002) , indicating that the Dutch participants scored significantly better than any of the other countries.

[14] Five participants from the UK ended up not having a score on this test for unknown reasons (3203, 3205, 3206, 3214 and 3215)

Figure 2.30 Results NEO FFI-3, Extraversion, per country. (N= 157, with 5 missing values[14])



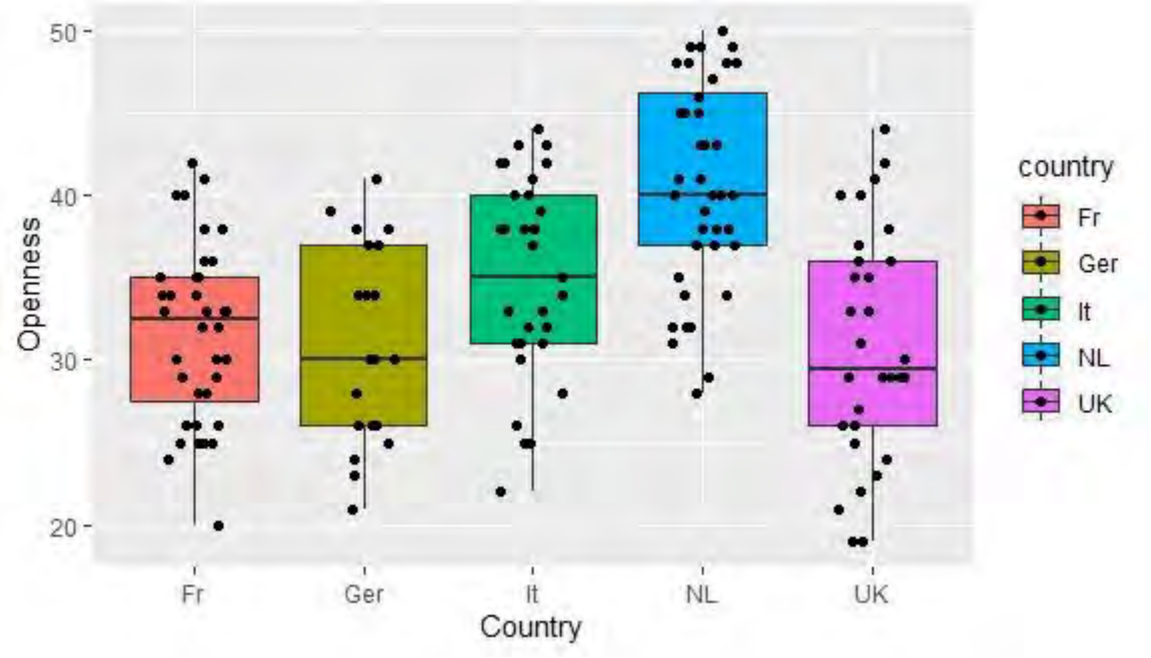
<u>Data distribution</u>

Looking at Figure 2.30, The Netherlands immediately jumps out as best performing country with its high median (42.5) and a relatively large level of internal agreement, judging by the size of the box. The data distribution seems to be slightly skewed towards the lower scores, with none scoring less than 25. Fifty percent of the participants scored in the range of 42.5 till 55.0 on the personality factor Extraversion. France seems to be the second-best performing country with a median of around 32 and scores ranging from 18 to 43. Despite their varying medians (around 27 and 29 respectively), Italy and the UK show similar patterns in the way their data is distributed with scores ranging from 13 to 43, the only difference being the fact that the Italian participants seem inclined to score above the median whereas the British participants have a tendency to score below the median. The data distribution of the German participants shows less variance compared to the other variance with scores ranging from 19 to 36 and a mean of 29. None of the countries had any outliers.

<u>Statistical significance</u>

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Extraversion and a significant effect was indeed found (F(4, 152)= 30.57; p<0.000). A subsequent Tukey post hoc test showed significant differences between France and the Netherlands (t=−7.872; p<0.000), Germany and the Netherlands (t= −7.816; p<0.000), Italy and the Netherlands (t= −8.821; p< 0.000) and the Netherlands and the UK (t=8.488; p<0.000) indicating that the Dutch participants scored significantly better than any of the other countries.

Figure 2.31 Results NEO FFI-3, Openness, per country. (N= 157, with 5 missing values[14])



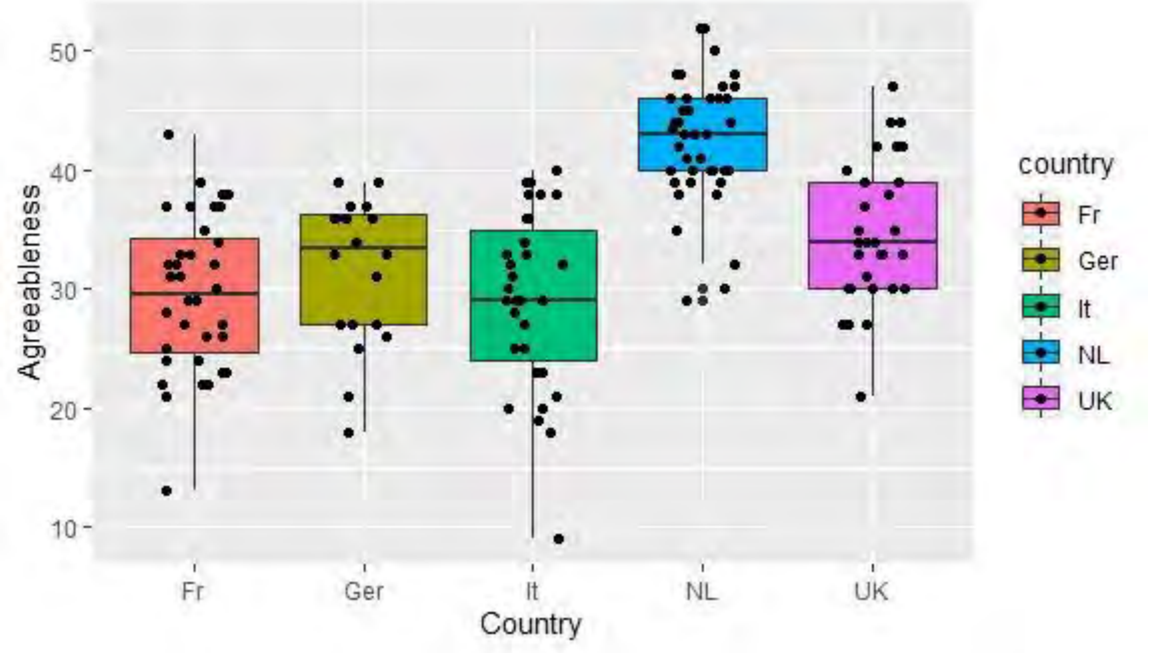
Data distribution

Looking at Figure 2.31, The Netherlands immediately jumps out as best performing country with its high median (40.0). The data distribution seems to be slightly skewed towards the lower scores, with none of the participants scoring less than 27. Fifty percent of the participants scored in the range of 40.0 till 50.0 on the personality factor Openness. Italy seems to be the second-best performing country with a median of around 35 and scores ranging from 22 to 44. Germany and the UK are similar in the way their data is distributed, both countries are skewed toward upper scores, have a median of around 30 and have relatively tall boxes, indicating varying scores among their participants. France seems to be the third best performing country on this variable with a median of around 33, a relatively high level of internal agreement and scores ranging from 20 to 42. None of the countries had any outliers.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Openness and a significant effect was indeed found (F(4, 152)= 16.08; p<0.000). A subsequent Tukey post hoc test showed significant differences between France and the Netherlands (t=−6.370; p<0.000), Germany and the Netherlands (t= −5.708; p<0.000), Italy and the Netherlands (t= −3.881; p< 0.000) and the Netherlands and the UK (t=6.551; p<0.000), indicating that the Dutch participants scored significantly better than any of the other countries.

Figure 2.32 Results NEO FFI-3, Agreeableness, per country. (N= 157, with 5 missing values[14])



Data distribution

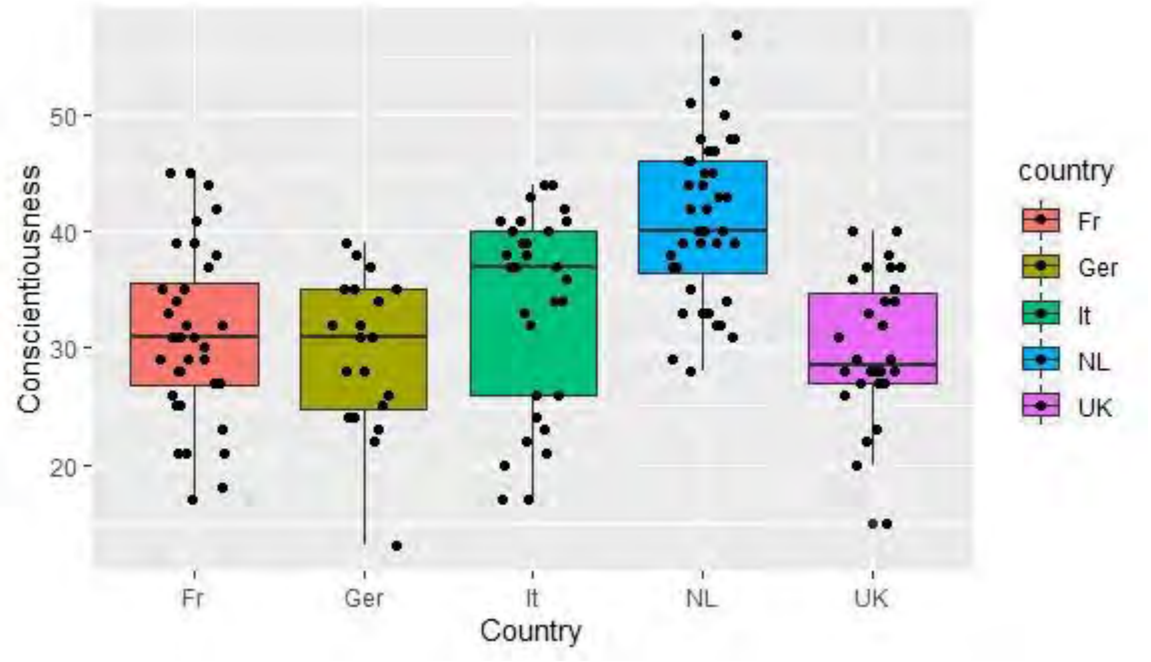
Looking at Figure 2.32, The Netherlands immediately jumps out as best performing country with its high median (42.5), it's high level of internal agreement and scores ranging from 32 to 52 on the personality factor Agreeableness. The UK seems to be the second-best performing country with a median of around 34 and scores ranging from 21 to 47. France and Italy are similar in the way their data is distributed, both countries are skewed toward the lower scores and have a median of around 29. Germany shows some resemblance to the UK in the sense that their medians are both at around 34 and their boxes are equally big, indicating a comparable level of internal agreement. Nevertheless, when comparing the countries in the way their data is distributed, the British participants seem to be inclined to score above the median whereas the German participants tend to score below the median. Another difference is the scoring range; the UK had scores ranging from 21 to 47 and Germany performed in a range of 17.5 to 39. The Netherlands was the only country with any outliers.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Agreeableness and a significant effect was indeed found (F(4, 152)=27.61; p<0.000).

A subsequent Tukey post hoc test showed significant differences between France and the Netherlands (t=−8.832; p<0.000), Germany and the Netherlands (t= −6.160; p<0.000), Italy and the Netherlands (t= −8.854; p< 0.000) and the Netherlands and the UK (t=5.114; p<0.000), indicating that the Dutch participants scored significantly better than any of the other countries. Moreover, significant effects were found between France and the UK (t= −3.211; p<0.01) and Italy and the UK (t= −3.450; p<0.006), indicating that the British participants scored significantly better than the French and the Italian participants.

Figure 2.33 Results NEO FFI-3, Conscientiousness, per country. (N= 157, with 5 missing values[14])



Data distribution

Looking at Figure 2.33, The Netherlands immediately jumps out as best performing country with its high median (40.0), and scores ranging from 28 to 57.5 on the personality factor Conscientiousness. France and Italy seem to be fighting for the runner-up position, both countries with scores ranging from 17.5 to 45. The countries differ however in the way their data is distributed. Even though Italy has the highest median of the two (37.5 and 31 respectively), its data is skewed towards the lower scores, indicating that the Italian participants were inclined to score below the median. France, however, shows a rather symmetric distribution with as many participants scoring above as below the median and a higher level of internal agreement. Germany shows some resemblance to France in the sense that both countries have similar levels of internal agreement and a median of around 31, the comparison stops there, as the German data is skewed towards the lower scores with a minimum score of 12.5 and a maximum score of 39. The UK seems to be the worst-performing country with a median of around 28 and scores ranging from 20 to 40. Outliers were found in the UK only.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable Conscientiousness and a significant effect was indeed found (F(4, 152)=14.96; p<0.000). A subsequent Tukey post hoc test showed significant differences between France and the Netherlands (t=−5.991; p<0.000), Germany and the Netherlands (t= −5.785; p<0.000), Italy and the Netherlands (t= −4.251; p< 0.000), and the Netherlands and the UK (t=6.269; p<0.000), indicating that the Dutch participants scored significantly better than any of the other countries.

2.5.3 Personality: ISALEM-97 (13)

Experiential learning is defined as the way an individual interacts with the world around him, and more importantly the way he/she absorbs and processes information such that learning takes place or problems can be solved (Therer, 1998). The *ISALEM-97* (Inventaires des Styles d'Apprentissage du Laboratoire d'Enseignement Multimédia) was used to assess the participants experiential preference and identifies four different learning styles: intuitive, methodical, reflexive, and practical [15](Therer, 1998).

The original French questionnaire was developed at the Université de Liège and translated into the four other project languages with permission from the authors. It is based on Kolb's experiential learning model (Kolb, 1984) and his Learning Style Inventory (Kolb, 1985), but differs in the sense that it provides more concrete questionnaire items and no longer makes use of keywords in the answers given, which makes the response possibilities less transparent (Cahay et al. 1997b:8-9).

The *ISALEM-97* questionnaire is made up of 12 items, 12 situations that can be found in both educational settings and everyday life. Each situation is accompanied by four possible reactions, each corresponding to a different learning style. The degree to which learning styles are implemented is represented in an X-value and a Y-value with X and Y being the two axes of experiential preference; respectively the *intuitive –methodical* axis, and the *analytic* – *practical* axis (see Appendices B and C).

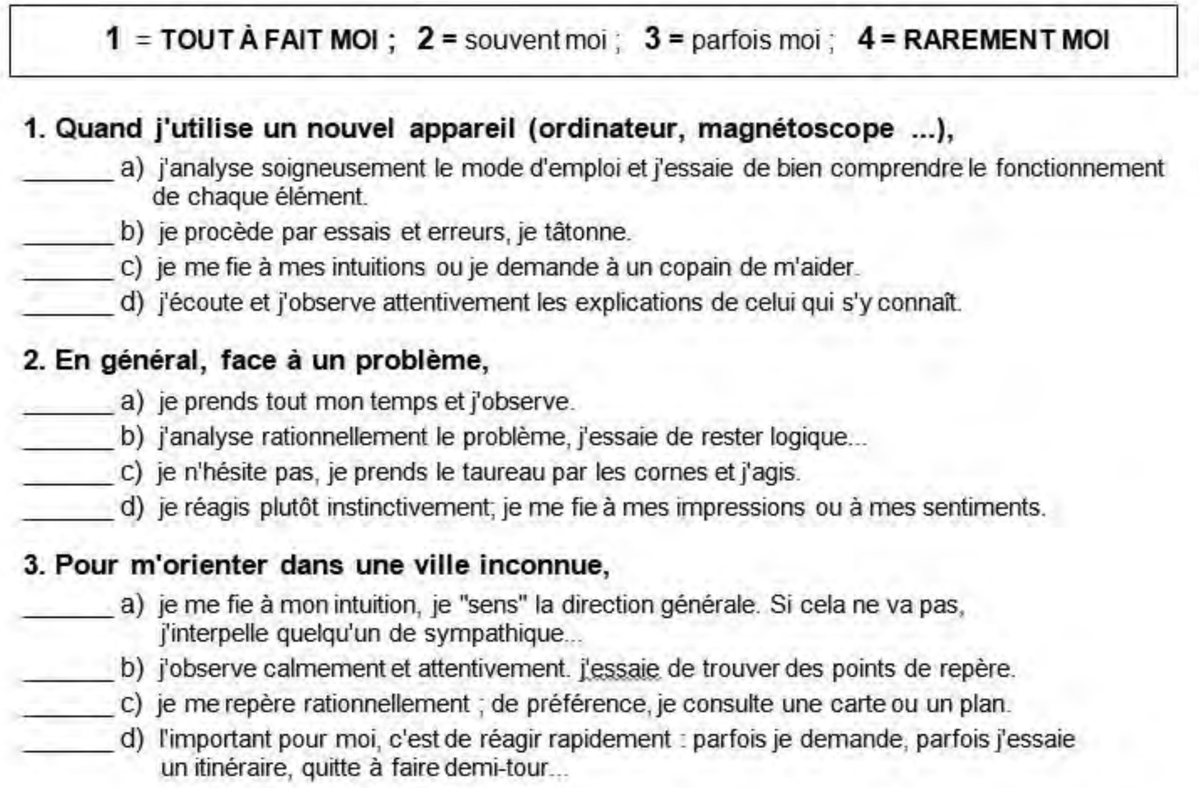
The X-value is a candidate's score on the x-axis that says something about whether a person is more inclined to respond to situations that arise in a intuitive or a more methodical way. Likewise, the Y-value says something about a person's inclination to respond in an analytic or a more practical way. The candidates' position in the grid ultimately defines his or her experiential learning profile: intuitive/analytic, intuitive/practical, methodical/analytic or methodical/practical (See Appendix D for a more detailed description of these four learning profiles.

The scoring procedure consists of three steps: rating, defining x and y values and defining learner profile. In the first step, candidates rate the four options according to their preference on the following scale: 1= totally me, 2= frequently me, 3= sometimes me 4= rarely me, using each number but only once (Therer, 1998). Upon completion, students transfer their answers onto the scoring sheet (see Appendix B) through which their personal y and an x-value is calculated. In the last step, respondents transfer their x and y values onto the grid in order to define their learner profile.

Figure 2.34 shows a few examples from the *ISALEM-97* questionnaire.
A complete version of the questionnaire (with instructions and background research) can be found at
http://www.lem.ulg.ac.be/StyleApprent/StyleApprent_CG/page_01.htm

[15] In the scoring sheet also referred to as respectively I-score (Intuition), Ab-score (Abstraction, Ac-score (Action) and R-score (Reflection)

Figure 2.34 Excerpt from the *ISALEM-97 questionnaire*



1 = TOUT À FAIT MOI ; 2 = souvent moi ; **3 =** parfois moi ; **4 = RAREMENT MOI**

1. Quand j'utilise un nouvel appareil (ordinateur, magnétoscope ...),

_______ a) j'analyse soigneusement le mode d'emploi et j'essaie de bien comprendre le fonctionnement de chaque élément.

_______ b) je procède par essais et erreurs, je tâtonne.

_______ c) je me fie à mes intuitions ou je demande à un copain de m'aider.

_______ d) j'écoute et j'observe attentivement les explications de celui qui s'y connaît.

2. En général, face à un problème,

_______ a) je prends tout mon temps et j'observe.

_______ b) j'analyse rationnellement le problème, j'essaie de rester logique...

_______ c) je n'hésite pas, je prends le taureau par les cornes et j'agis.

_______ d) je réagis plutôt instinctivement; je me fie à mes impressions ou à mes sentiments.

3. Pour m'orienter dans une ville inconnue,

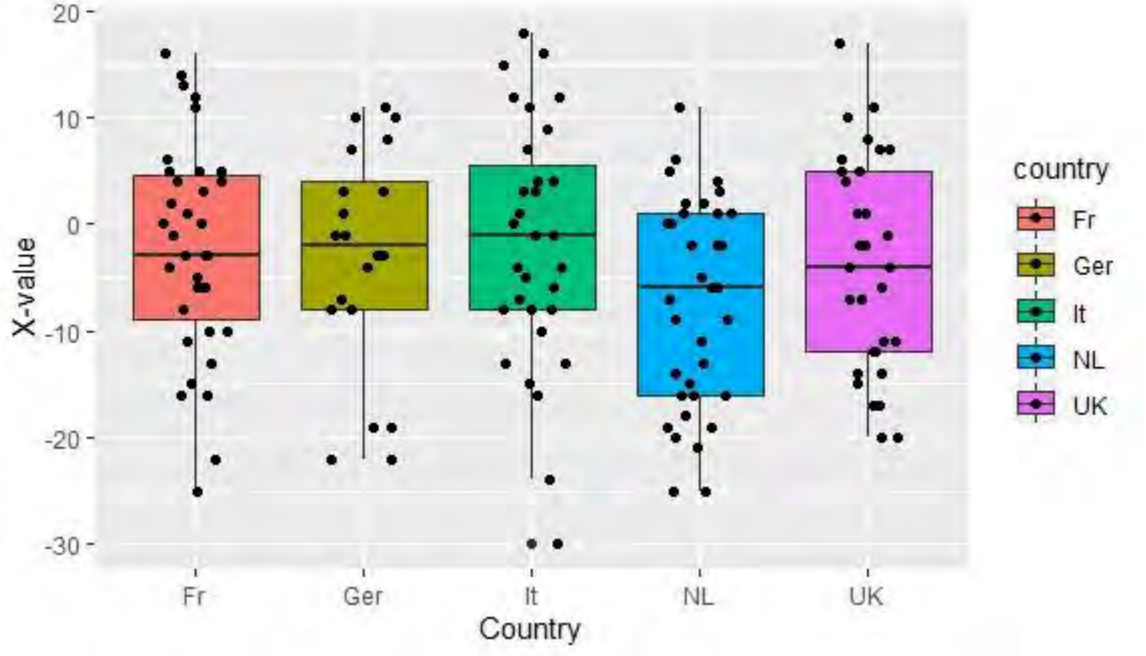
_______ a) je me fie à mon intuition, je "sens" la direction générale. Si cela ne va pas, j'interpelle quelqu'un de sympathique...

_______ b) j'observe calmement et attentivement. j'essaie de trouver des points de repère.

_______ c) je me repère rationnellement ; de préférence, je consulte une carte ou un plan.

_______ d) l'important pour moi, c'est de réagir rapidement : parfois je demande, parfois j'essaie un itinéraire, quitte à faire demi-tour...

Figure 2.35 Results ISALEM, X-value, per country. (N= 154, with 8 missing values)



Data distribution

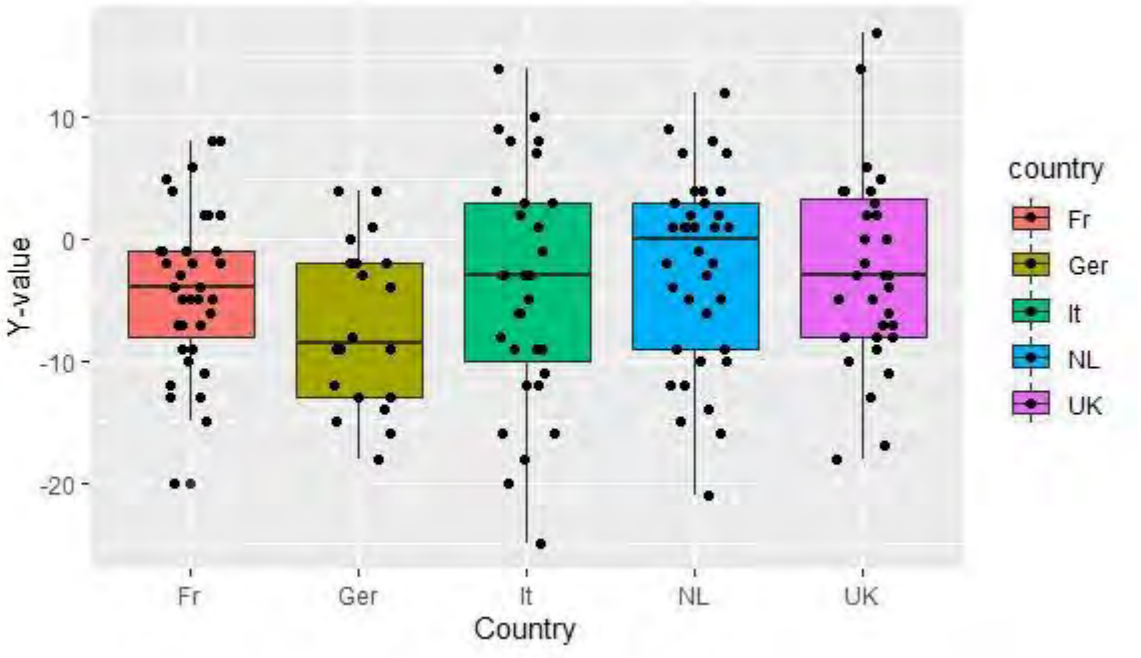
Looking at Figure 2.35, we can see that there is not a lot of difference between the countries in the way their data is distributed. France, Germany and Italy show the most similarity with medians of around −2, indicating a very mild preference for the intuitive approach and a relatively symmetric spread. This preference for the intuitive approach seems to be slightly stronger with the British and Dutch participants, judging by their medians of −4 and -6 respectively. Italy is the only country with an outlier.

Overall, participants of all countries, except the Netherlands don't seem to have a very clear preference for one particular side of the axis and hover somewhere between 5 and 12. 75% of the Dutch candidates however seem to have a more outspoken preference for the intuitive learning style with scores ranging from 0 to 25.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable X-value, and this turned out not to be the case (F(4, 149)= 1.408; p<0.234).

Figure 2.36 Results ISALEM, Y-value, per country. (N= 154, with 8 missing values)



Data distribution

Looking at Figure 2.36 we can see a division between countries with a clear preference for the practical learning style (France and Germany) and those with a less outspoken preference (Italy, the Netherlands, and the UK). Interestingly, disregarding the one outlier in France, France and Germany are also the countries with the most internal consistency and less extreme minimum or maximum values, which might be a indication that this approach is looked upon favorably, or is perhaps even stimulated in the culture of these countries. France is the only country with an outlier.

Overall, judging by the median and the way their data seem to be skewed towards this side of the axis, countries seem to favor the practical learning style.

Statistical significance

An ANOVA test was run to find out if there were any significant differences between the countries on the variable X-value, and this turned out not to be the case (F(4, 149)= 1.346; p<0.256),

2.6 Tests: Motivation (14)

2.6.1 Motivation for Polish

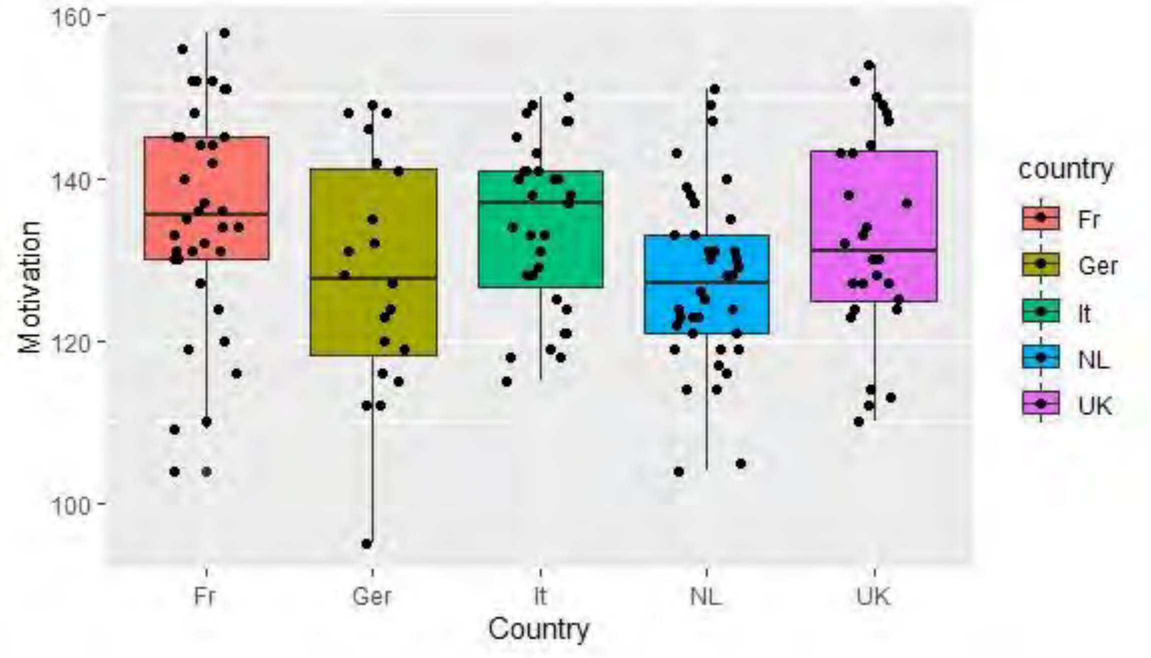
Motivation for learning the language was checked at the end of the two-week experiment by means of an in-house motivation questionnaire. The questionnaire was an adapted version of an instrument used in previous language acquisition studies (Hilton et al. 2008; Rast et al. 2011), which in turn was adapted from Gardner's (2004) *Attitude and Motivation Test Battery*, in 2005. It consisted of 30 items with statements reflecting one of Gardner's (2004) ten different motivational or emotional reactions: interest for foreign languages, attitude towards Polish, attitude towards learning Polish, motivational intensity, integrative orientation, instrumental orientation, anxiety in using Polish, classroom anxiety, emotional reactions to the lessons, emotional reactions to the Polish tests (VLLA Field Manual, 2022). Annex 5 gives the questionnaire with the coded item types and the scoring key. Subjects had to indicate to which extent the statements applied to them, using a Likert-type scale ranging from "totally disagree" to "totally agree" with no neutral option. Depending on whether the questions were formulated in a positive or a negative way, the maximum number of points were attributed to "totally agree" and "totally disagree respectively. An example of the type of statements and the type of Likert scale used is shown in Figure 2.37. The questionnaires were scored manually and the individual response data was entered into a spreadsheet.

Figure 2.37 gives a visual representation of the results per country.

Figure 2.37 Excerpt from the *motivation questionnaire*

Statement	Scale				
	pas du tout d'accord	pas d'accord	plutôt pas d'accord	plutôt d'accord	tout à fait d'accord
Le polonais est une langue qui me plaît					
Le polonais est une langue facile à prononcer					
C'est inutile d'étudier le polonais					

Figure 2.38 Results Motivation questionnaire per country. (N= 153, with 9 missing values)



Data distribution:

There is quite a bit of variance between the countries when it comes to motivation for the Polish language. The Netherlands seems to have a symmetrical distribution with a median of 126 and no outliers. The data of the UK, Germany and France all seem to be skewed towards the upper ranges of their respective medians. Italy sticks out with the highest median score but seems to be skewed towards the lower scores. The within-country variance seems to be highest in Germany and the UK. France is the only country with an outlier of 105. Most students in this country scored 135 or higher.

France and Italy seem to score highest on motivation for the Polish language, with 50% of their participants scoring respectively 135 or more and 137 or more. The worst performing country when it comes to motivation is Germany with 25% of their candidates scoring in the range of 119 to 128 and 25% scoring even less.

Statistical significance

ANOVA revealed some significant differences for motivation across the countries (F(4, 148)= 2.873; p=0.025).
A subsequent Tukey post hoc test showed a significant difference between France and the Netherlands (t=2.875; p=0.037), indicating that the French students were significantly more motivated than the Dutch participants were.

2.7 Overview of most salient differences in learner variables

Given that the VILLA teams assumed that the strict selection criteria would lead to a homogenous sample of learners and only administered the psychometric tests to control for potentially confounding learner variables, it is interesting to see if the VILLA researchers were right in their initial assumption, or whether the samples ended up being less homogenous than they had expected.
Table 2.11 gives an overview of the most salient differences in learner variables between the countries.

Table 2.11 Overview of the most salient differences in learner variables between the countries

Learner variable	Observed difference between the project countries
Background/ General data	
Age	**Younger students in France, the Netherlands and the UK** Significant age differences between France and Germany, France and Italy, Italy and the Netherlands, and Italy and the UK, indicating that the participants are significantly younger in France, the Netherlands and the UK.
Sex	No significant differences.
Type of study	**Significant differences between the countries** Further analyses are needed to find out more details
Background languages/ L2 status	Not yet analyzed.
Psychometric Tests: Cognition	
Raven	No significant differences.
Digit Span	No significant differences.
Letter Number Sequencing (LNS)	No significant differences.
Flanker Inhibition Cost	No significant differences.
Barsch LSI Visual Preference	No significant differences.
Barsch LSI Auditory Preference	No significant differences.
Barsch LSI Kinesthetic Preference	**The French, the only participants to have a preference for the kinesthetic modality** Near-significant difference between France and the Netherlands, indicating that the French participants seem to attribute more importance to the kinesthetic modality than the Dutch do.
Psychometric Tests: Language Aptitude	
Llama B Vocabulary learning	**The Dutch and Germans better than the Italians.** Significant differences between Germany and Italy, and Italy and the Netherlands, indicating that the Dutch and the Germany respondents outperformed the Italian learners.
Llama D Sound recognition	**Italians outperform the French, the Germans, and the Dutch/ the Brits better than the Dutch** Significant differences between France and Italy, Germany and Italy, Italy and the Netherlands and the Netherlands and the UK, indicating that the Italian participants were significantly better at sound recognitions than the French, the German and the Dutch. Moreover, the British respondents performed significantly better than their Dutch counterparts.

Llama F Grammatical Inferencing	No significant differences.
Psychometric Tests: Personality	
NEO FFI-3 Neuroticism	**The Dutch outscore the other countries** Significant differences between France and the Netherlands, Germany and the Netherlands, Italy and the Netherlands, and the Netherlands and the UK, indicating that the Dutch participants scored significantly better than the other countries.
NEO FFI-3 Extraversion	**The Dutch outscore the other countries** Significant differences between France and the Netherlands, Germany and the Netherlands, Italy and the Netherlands, and the Netherlands and the UK, indicating that the Dutch participants were significantly better than any of the other countries.
NEO FFI-3 Openness	**The Dutch outscore the other countries** Significant differences between France and the Netherlands, Germany and the Netherlands, Italy and the Netherlands, and the Netherlands and the UK, indicating that the Dutch participants scored significantly better than any of the other countries.
NEO FFI-3 Agreeableness	**The Dutch outscore the other countries and the Brits do better than the French and Italians** Significant differences between France and the Netherlands, Germany and the Netherlands, Italy and the Netherlands, and the Netherlands and the UK, indicating that the Dutch participants scored significantly better than any of the other countries. Moreover, significant effects were found between France and the UK, and Italy and the UK, indicating that the British participants scored significantly better than the French and the Italian participants.
NEO FFI-3 Conscientiousness	**The Dutch outscore the other countries** Significant differences between France and the Netherlands, Germany and the Netherlands, Italy and the Netherlands and the Netherlands and the UK, indicating that the Dutch participants scored significantly better that any of the other countries.
Isalem-97 X-value	No significant differences.
Isalem-97 Y-value	No significant differences.
Psychometric Tests: Motivation	
Motivation for Polish	**The French more motivated than the Dutch** Significant differences between France and the Netherlands, indicating that the French participants were significantly more motivated to learn the Polish language than the Dutch participants were.

Clearly the sample is less homogenous than the VILLA researchers were hoping for as there are some significant differences in learner variables between the countries

3. Language tasks

3.1. Introduction

In order to document the progress made by the participants in their acquisition of the Polish language, twelve language tasks were administered to measure differences in linguistic proficiency. These experiments can be broken down into *reaction time forced choice tasks* (Phoneme Discrimination, Lexical Decision, Word Recognition and Grammatical Judgement I), *controlled oral production tasks* (Question and Answer task and Sentence Imitation, *paper and pencil tasks* (Grammaticality Judgment II, Sentence Puzzle, Picture Verification and Cloze Test), and *recorded complex oral production tasks* (Route Direction and Film Retelling) (Dimroth et al., 2013).

Given the fact that the VILLA project set out to study the effects of input on the acquisition process, a longitudinal test design was chosen in which all experiments except for the Cloze test, the Route Direction test and the Film Retelling test were administered more than once so that students' performance could be observed over time. The latter three tests were given during the wrap-up session only as they were either considered to be too hard for the students to do at any earlier stage of the acquisition process, or because they measured properties of the Polish language that were not originally part of the experimental design and therefore not scripted to be observed, but that occurred frequently enough to be regarded as part of the foundations of acquisition (e.g. subject verb agreement and the (non-)use of anaphoric pronouns) (Dimroth et al., 2013).Details about the tests as well as the longitudinal setup can be found in Table 3.1.

A full description of all the language tasks employed by the VILLA project can be found in Dimroth et al. (2013), but is beyond the scope of this thesis. The focus of this thesis is to study the effects of learner variables on the acquisition of a new language and for this reason a selection was made from tests that both covered different language domains and those that had been administered at a minimum of two different time intervals. The next sections are devoted to three language domains and their corresponding selected tasks: Lexicon (*Word Recognition*), Morphology (*Grammatical Judgement* I) and Phonology (*Phoneme Discrimination)*.

Table 3.1. Language tasks employed by the VILLA project

Language tasks	**Description**	
reaction timed forced choice tasks (computer + audio, software: E-prime16)		
1. Phoneme Discrimination	Task:	Deciding whether the sound strings heard are the same or different.
	Focus:	Minimal Pairs
	Purpose:	Measuring sensitivity to phonological contrasts.
	Domain:	Phonology
	Test Frequency:	T0, T3 and T7.[17]
2. Lexical Decision	Task:	Deciding whether the sound strings heard are Polish or not.
	Purpose:	Measuring ability to distinguish between Polish (sounding) words and Chinese words
.	Domain	Lexicon/ Phonology
	Test Frequency:	T0, T2, T4, T6 and T817
3. Word Recognition	Task:	Identifying words heard both in a sentence and in isolation.
	Purpose:	Measuring the candidate's growing sentence segmentation capacities[18]
	Domain	Lexicon
	Test Frequency:	T0, T5 and T9[17]

[16] Schneider et al., 2002

[17] T = Time of testing. T0=prior to the first day of contact, T1-T9= input sessions of 1.5 hours each, T10= the wrap-up session

[18] Shoemaker & Rast 2013

4. Grammaticality Judgment I	Task:	Deciding whether a sentence is grammatically correct or not.
	Focus:	Nominal Case
	Purpose:	Measuring knowledge about case marking.
	Domain:	Morphology
	Test Frequency:	T3 and T7[17],
paper and pencil tasks		
5. Grammaticality Judgment II Alternative name: Verbal Morphology	Task:	Deciding whether a sentence is grammatically correct or not.
	Focus:	Subject-Verb-Agreement, Word Order
	Purpose:	Measuring learners' ability to detect violations in subject-verb-agreement.
	Domain:	Morphology
	Test Frequency:	T10 [17]
6. Sentence Puzzle	Task:	Re-ordering the words in the sentence such that a logical response is created.
	Focus:	Written Word Order
	Purpose:	Discovering the learners' type of developing grammar rules. In particular: the influence of learners' L1 and Ln on their preferred word order in Polish.
	Domain:	Syntax
	Test Frequency:	T5 and T8[17]
7. Picture Verification	Task:	Deciding whether the audio response corresponds to the picture or not.
	Focus:	Morphosyntax/Argument Roles
	Purpose:	Discovering the learners' type of developing grammar rules. In particular: the learners' interpretation of argument roles.
	Domain:	Syntax
	Test Frequency:	T6 and T9[17]
8. Cloze test	Task:	Completing the dialogues by adding personal pronouns where necessary.
	Focus:	Personal Pronouns
	Purpose:	Investigating the acquisition of the rules for use of overt anaphoric pronouns.
	Domain:	Discourse
	Test Frequency:	T10[17]
recorded controlled oral production tasks		
9. Sentence Imitation	Task:	Listening to a sentence and repeating that same sentence after a moment of distraction.
	Focus:	Morphosyntax/Argument Roles
	Purpose:	Discovering the learners' type of developing grammar rules. In particular: facilitation of dealing with marked OVS-sentences in comparison to default SVO-sentences.
	Domain:	Syntax
	Test Frequency:	T6[17]
10. Oral Question -Answer task (Q&A) Alternative name: Picture Production task	Task:	Listening to a question and giving a response with the aid of pictures.
	Focus:	Nominal Case
	Purpose:	Measuring productive use of case marking.
	Domain:	Morphology
	Test Frequency:	T3 and T7[17]

recorded complex oral production tasks		
11. Elicited production I: Route Direction(RD)	Task:	Giving directions through a role play
	Focus:	Semi-structured Free Production
	Purpose:	Comparing learners' utterance structure to Klein & Perdue's (1997) Basic Variety.
	Domain:	Discourse
	Test Frequency:	T10[17]
12. Elicited production II: Finite Story (FS)[19] Alternative name: Film Retelling (FR)	Task:	Reconstructing a movie narrative.
	Focus:	Free Production
	Purpose:	Comparing learners' utterance structure to Klein & Perdue's (1997) Basic Variety.
	Domain:	Discourse
	Test Frequency:	T10[17]

3.2 Lexicon

Word Recognition was one of two tasks employed to document the learner's lexical development. It was administered at three different time intervals; at point zero, prior to the course (T0), and twice during the course, on day 5 (T3) and day 9 (T9). The Word Recognition task was a reaction time, forced choice experiment in which participants were asked to listen to a Polish sentence, followed by a Polish word in isolation after which the participants had to indicate whether the word heard was present in the sentence or not (Dimroth et al., 2013).

Some notes on the Word Recognition task

Word Recognition tasks, provide two types of measures used in SLA research to assess lexical development in various ways (Foster-Cohen et al., 2006). Response accuracy is used to index the size of learner vocabulary (Meara 1996, Meara & Milton 2002) and response speed is used to measure the learner's development of lexical processing skills (Fukkink et al.,2005, Schoonen et al. 2003, Segalowitz & Hulstijn, 2005; van Gelderen et al. 2004). However, it should be noted, that despite its widespread use, the test is not free of criticism. Validity of the Yes-No format for example has been researched with varying results (Cameron, 2002; Eyckmans, 2004; Meara & Buxton, 1987; Mochida & Harrington, 2006; Shillaw, 1996). Other objections raised include methodological challenges in the measurement of response speed and variability (Laufer & Nation, 2001; Sternberg, 1998) , statistical issues in the use of response time measures (Foster-Cohen et al., 2006; Harley 2001; Luce, 1986) and the role individual differences in base information processing rates play on L2 processing (Faust et al., 1999; Segalowitz & Frenkiel-Fishman 2005).

Procedure (VILLA Field Manual, 2022)

E-Prime experimental software (Schneider et al., 2002) was used to create and present the experimental protocol. The test was administered on either desktop computers or laptops and stimuli were presented binaurally through headphones, in randomized order. The entire experiment consisted of 106 trials and lasted approximately 12 minutes.

In each trial, participants listened to a sentence in Polish followed by a Polish word in isolation and participants reported whether the word in isolation was present or not in the sentence by pressing 1 or 2 respectively. Prior to the real experiment, participants completed a training session with 10 trials in order to familiarize them with the procedure. The items included in the experiment were different from the ones used in the training session. There was no set response limit, however, participants were instructed to respond as fast as they could without sacrificing accuracy.

Table 3.2 gives an overview of the variables of the experiment. The lexical items varied systematically along two features: frequency and transparency. Items with a high frequency were present in the instructional input, whereas items

[19] Dimroth (2012)

with a low frequency were absent in the instructional input. Highly transparent items are items with a high cognate status, that is, items that carry great resemblance to words known in any of the five source languages.[20]

Table 3.2. Variables in the Word Recognition task

Independent variables		Dependent variables
Frequency:	High (20+ tokens) and Low[21]	Accuracy
Transparency:	High and Low	Reaction time
Test:	T0, T5 and T9	

Stimuli

The stimuli consisted of 48 test sentences, in which the target word was present in the sentence, and 48 distracter sentences, in which the target word was not present in the sentence. The target words were categorized in four different groups depending on the degree of transparency and frequency: HT/ HF; LT/LF; HT/LF and LT/HF and contained 12 items each. Appendix H gives an overview of the target words and test sentences used.

Results from earlier VILLA-studies on lexical development[22]

3.3 Morphology

The Grammaticality Judgment-I task was one of three tasks employed by the VILLA project to document the acquisition of different properties of Polish inflectional morphology and measured learners' knowledge of case marking. Participants were tested at two time intervals during the course; after 4,5 hours and after 10,5 hours of exposure to the Polish language (T3 and T7). The reaction time forced choice experiment required participants to listen to a sentence and then indicate whether they thought the sentence heard was grammatical or not. The task included items varying in transparency and frequency.

Some notes on the Grammaticality Judgment task

Grammaticality Judgment tasks (GJT), also known as acceptability tests (Ionin & Zyzik, 2014) have long been used in Second Language Acquisition (SLA) research as a promising tool to measure learners' underlying language skills (Shiu et al., 2018). It is popular for a number of reasons: it is fast, a test session generally takes about 10 minutes (Slik et al.,2021), it is relatively easy to administer to a large number of participants (Shiu et al., 2018), it can be employed on a wide variety of languages due to the fact that grammaticality is considered to be a cross linguistic feature (Kail et al., 2012) and it can be used to assess features of the target language that are deemed too difficult to be elicited from the learner in activities designed to generate production. (Loewen, 2009). It is not surprising therefore that the VILLA project decided to include the task in their test battery. However, it should be noted, that despite its popularity and widespread use, the task is not free of criticism. Among the objections raised are the influence of GJT design features such as time constraints, task stimulus and task modality on learners' performance (Shiu et al., 2018), and concerns about the construct validity of the test (Godfroid et al.,2015; Slik et al.,2021).

Procedure (VILLA Field Manual, 2022)

E-Prime experimental software (Schneider et al., 2002) was used to create and present the experimental protocol. The test was administered on either desktop computers or laptops and stimuli were presented binaurally through headphones, in randomized order. The entire experiment consisted of 104 trials and lasted approximately 10 minutes. In each trial, candidates had to listen to two different types of sentences in which some contained a grammatically incorrect copula construction with a double nominative used in the wrong context and others had a grammatically correct construction with a nominative and an instrumental NP. Participants were asked to listen to a sentence and then

[20] The transparent items in this study were carefully selected. An item was considered to be transparent if 50 % of the native speakers of any of the five source languages was able to recognize it in a test done prior to the VILLA experiment. (Dimroth et al., 2013).

[21] Items with a high frequency were present in the instructional input, whereas items with a low frequency were absent in the instructional input.

[22] This section could not be finished at the time of writing this thesis as I did not have access to the required information. More information about this topic can be found in the VILLA Field Manual (2022).

indicate whether they thought the sentence was grammatically correct or not by pressing 1 or 2 respectively on their keyboards.

Prior to the real experiment, participants completed a training session with 8 trials in order to familiarize them with the procedure. The items included in the experiment were different from the ones used in the training session. There was no set response limit, however, participants were instructed to respond as fast as they could without sacrificing accuracy.

Table 3.3 gives an overview of the variables of the experiment. The lexical items varied systematically along three features: frequency, transparency and gender, related to the nominative NP. Items with a high frequency were present in the instructional input, whereas items with a low frequency were absent in the instructional input. Highly transparent items are items with a high cognate status, that is, items that carry great resemblance to words in other languages[20]. Polish has gender inflection, meaning that nouns take on different forms depending on their gender.

Table 3.3. Variables in the Grammaticality Judgment-I task

Independent variables		Dependent variables
Frequency:	High (20+ tokens) and Low[21]	Accuracy
Transparency:	High (transparent) and Low (opaque)	Reaction time
Gender:	Masculine and Feminine	
Test:	T3 and T7	

Stimuli

The stimuli consisted of 64 test sentences subdivided in two categories of 32 sentences each; nouns of profession and nouns of nationality, as well as 32 distracter sentences. The target sentences were further categorized in four different groups per category, depending on the degree of transparency and frequency: HT/ HF; LT/LF; HT/LF and LT/HF and contained eight target words each. At target word level, there was another subdivision in noun gender, with four target words that were feminine and 4 that were masculine per group in each category.Appendices I and J give an overview of the test sentences used per category.

Results from earlier VILLA studies on morphology[22].

3.4 Phonology

In the interest of establishing learners' sensitivity to phonological contrasts in the Polish language, one of the tasks administered was a reaction time, forced choice Phoneme Discrimination task, in which students were tested at three different time intervals, at point zero, prior to the course (T0), and during the course, after respectively 4,5 (T3) and 10,5 (T7) hours of exposure to the Polish language. The task also served another purpose, namely observing the influence of source languages and other acquired languages on the process of perceiving or learning to perceive phonological distinctions.

In the Phoneme Discrimination task, participants had to listen to minimal pairs; pairs of syllables from the Polish phonemic inventory that were either identical, or differed in exactly one phoneme, after which they were asked to decide whether the pairs in the trials were identical or different (VILLA Field Manual, 2022).

Some notes on the Phoneme Discrimination task

The Phoneme Discrimination task, also known as the Auditory Discrimination task, is used to track changes in discrimination accuracy of phonetic and phonological contrasts. It is often used to investigate L2 phoneme discrimination. Critics claim that the ability to discern phonological contrasts in an L2 language does not provide direct evidence for L2 category acquisition. They say that the way phonological contrasts are categorized, i.e., assimilated to the L1 phonological system, defines the degree to which new phonological contrast in L2 are detected (Best 1993, 1994, 1995; Faris et al., 2018; Tyler, 2019). If results are to provide any direct evidence for L2 category acquisition, a categorization task should be administered along with the Phoneme Discrimination task (Faris et al., 2018; Tyler, 2019).

Procedure (VILLA Field Manual, 2022)

E-Prime experimental software (Schneider et al., 2002) was used to create and present the experimental protocol. The test was administered on either desktop computers or laptops and stimuli were presented binaurally through headphones, in randomized order. The entire experiment consisted of 249 trials and lasted approximately 12 minutes. In each trial, participants had to listen to pairs of syllables and were asked to report whether they heard two instances of the same syllable, or two different syllables by pressing 1 or 2 respectively on their keyboards. Prior to the real experiment, participants completed a training session with 9 trials in order to familiarize them with the procedure. The items included in the experiment were different from the ones used in the training session. After each pair of syllables, there was an interstimulus interval of 250 msec. There was no set response limit, however, participants were instructed to respond as fast as they could without sacrificing accuracy. Table 3.4 gives an overview of the variables of the experiment.

Table 3.4. Variables in the Phoneme Discrimination task

Independent variables	Dependent variables
Test: T0, T3 and T7	Accuracy
	Reaction time

Stimuli

The stimuli consisted of pairs of CV-syllables that contained six sibilants from the Polish phonemic inventory followed by /a/. See Table 8 for an overview of the six sibilants used, including their phonemic properties. The pairs were recorded by a female speaker of Polish and presented in all possible combinations and in both orders; for example: /sa/ - /za/ and /za/-/sa/. The experiment was made up of 240 trials of which 60 distractor trials with 30 same pairs (/sa1/ -/sa2/) and 30 different pairs (/sa/ - /za/) and 180 test trials that were administered in three different presentations with 30 unique pairs of identical syllables per presentation (90 trials in total) and 30 unique pairs of different syllables per presentation (90 trials in total).

Table 3.5. Overview of the phonemic properties of the stimuli used in the P.D. task

STIMULI	ALVEOLAR	ALVEO-PALATAL	RETROFLEX
Unvoiced	/sa/ są	/ɕa/ sia	/ʂa/ sza
Voiced	/za/ zą	/ʑ/ zia	/ʐa/ rza

Results from earlier VILLA studies on phonology[22]

3.5 Measuring accuracy: d prime (d')

Word Recognition, Grammaticality Judgment and Phoneme Discrimination are examples of so-called forced choice tasks, in which participants respond to the stimuli presented with either a yes or no answer. For all three tasks we focused on accuracy to measure the performance of the participants. We limited the analyses to the overall test performance in relation to time intervals, leaving more detailed analyses of item effects (frequency, transparency, gender) to others. We measured accuracy by using d prime (d'), a sensitivity measure originating from signal detection theory (SDT), that takes both the hits and false rejections into account.

In Yes/No forced choice tasks there are two different experimental test situations: signal trials, in which the signal is present, and noise trials, in which the signal is absent. Considering that there are also two possible responses ("yes" or "no"), there are four possible outcomes to each of the trials: hits, misses, false alarms and correct negatives.

Responses are labeled a “Hit”, when participants correctly go for the option “Yes”, when the signal is indeed present in the trial. If, however, the signal was present, but participants choose to respond with “No”, the result is a “Miss”. Likewise, “False Alarms” occur when respondents say they observed the signal, while it was absent in the trial. When the signal is absent from the trial and the student correctly indicates that it was absent the result is a “Correct Negative”. Table 3.6 gives an overview of the signals used in the three language tasks as well as the possible outcomes of the trials.

Table 3.6 Overview of the signals and responses in the WR-, GJ- and PD tasks

Signal	Response	
WR: Isolated word was in sentence		
GJ: Sentence is grammatically correct		
PD: Pairs are identical		
	Yes	No
Present	Hit	Miss
Absent	False Alarm	Correct Negative

The most obvious way to measure performance on these tasks is to calculate the net score, that is, the number of hits minus the number of false alarms. The problem with this way of computing performance is that it results in a heavily skewed score in cases with many false alarm (Haatveit et al., 2010). For this reason, results were analyzed using d-prime (d’), which measures an individual’s sensitivity, or discriminability; that is, an individual’s ability to distinguish between target stimuli (signal) and distractor stimuli (noise) by taking into account the relative proportion of hits minus false alarms (Haatveit et al., 2010). The method was derived from Signal Detection Theory (SDT) (Green et al., 1966; Macmillan, 1993; Swets, Tanner, & Birdsall, 1961) and has become the standard way of assessing awareness (Vermeiren & Cleeremans., 2012).

D’prime (d’) as a measure of sensitivity is defined as the difference between the means of both signal and noise in standard deviation units (Stanislaw & Todorov, 1999) and uses the formula: $d' = Z_{Hit} - Z_{FA}$, where Z is the Z-score, a standardized score indicating the distance from the mean (=0) when measured in standard deviation units.

From the above, it follows that high sensitivity is the result of maximizing hits and minimizing false alarms and the better the sensitivity, the better the participant is able to discriminate between target and non-target stimuli in a task (Haatveit et al., 2010). As far as interpretation of d’ is concerned, a d’ close to zero indicates inability to distinguish signals from noise and can thus be interpreted as a lack of conscious access (Haatveit et al., 2010; Stanislaw & Todorov, 1999; Vermeiren & Cleeremans, 2012). A positive d’ on the other hand indicates a better than chance performance on the task with larger values suggesting a correspondingly greater ability to detect presence or absence of the signal (Haatveit et al., 2010; Stanislaw & Todorov 1999; Vermeiren & Cleeremans, 2012). Negative values of d’, indicating that a participant scored less than 50% accuracy on hits and false alarms generally occur as a result of sampling errors, or response confusion such as responding *yes* when intending to respond *no* (Haatveit et al., 2010; Vermeiren & Cleeremans, 2012).

3.6 Statistical analyses

All analyses were done with R studio, using version R4.1.1. Appendix G gives an overview of the packages used. We decided to opt for an overall d’ analysis of all three tasks and to exclude reaction times. The linguistic-internal independent variables were not included in the analyses so as to keep enough power to analyze the other independent variables involved. In chapters 4 to 6 we analyze the effects of country, learning condition and time of measurement on the outcomes of the three tasks. We used generalized linear modelling (*glm, R*) and post-hoc comparisons (*emmeans*) and the packages *ggplot2* and *sjPlot* to visualize results.

Chapter 7 is dedicated to correlations. In the first two sections the correlations between the various language tasks and the psychometric tests are analyzed in order to establish the extent to which they measure different constructs. In the last section we analyze the effects of a selection of learner variables on the leaners’ results of the three language tasks.

In order to limit the number of missing values, we decided to include only the tests assessing cognition, language aptitude and motivation, thus excluding the background data and the results on the personality tests. We also included the data on the learning conditions. Analyses were done using a forward regression procedure with standardized results so as to define a subset of statistically significant comparable predictor variables for each of the language task's sessions, after which we applied regression analysis (*lm)* in combination with *sjPLot* to visualize the results.

3.7 Presentation results

In the following three chapters we present the results of the three language tasks analyzed in this thesis. Chapter 4 describes the results on the Word Recognition task, in chapter 5 the results for the Grammaticality Judgment task can be found and in chapter 6 we present the results on the Phoneme Discrimination task. In these three chapter the results are presented in two ways.

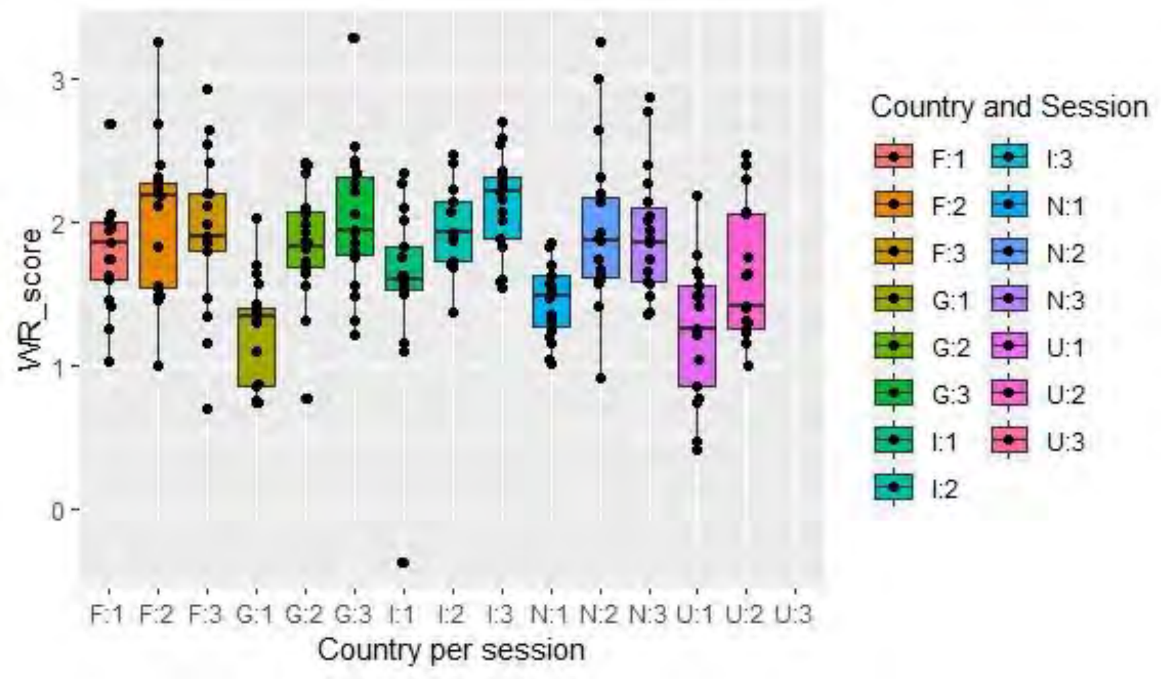
Due to the fact that Germany opted to compare adults with children in the VILLA project and given that the adults were placed in the Meaning Based Group (M) and the children in the Form-Focused group (F), a comparison of results for both learning conditions between the five countries could not be made, as there is no German adult data for the Form-Focused condition. For this reason, we decided to present two types of results: the results of the M-group, in which the performance of all countries can be compared, and the results of both learning conditions (M+F) of all countries with the exception of Germany, in which a comparison is made between the performance observed from the countries in both learning conditions.

4. Word Recognition

4.1 The effects of time and country in the MB (meaning-based) condition

The results for the M-group (MB condition) are visualized in Figure 4.1, with three time points (T0, T5 and T9) and five countries. As explained in Chapter 3, the performance of the participants was measured with d'.

Figure 4.1 Word Recognition scores of the M-group per session per country (N= 141, with 21 missing values).



Note: The data set had 21 missing values that were not included in Figure 4.1.
This large number of missing values is mainly due to the fact that for unknown reasons, the UK-data from the third session of all of the participants in the MB -group were missing.

Looking at Figure 4.1, we can see that overall, except for France, countries seem to have improved throughout the test sessions. The French participants seem to have regressed to almost starting level after their initial improvement in session 2. With a score of almost 1.8, France does have the highest median in the first session, indicating that the French participants seemed to outperform the other countries on the Word Recognition task. Germany and Italy show an increase in performance as time progresses, whereas the Netherlands seem to stagnate after session 2. Of the five countries, Germany seems to have made the biggest progress judging by the results from the first and the second session. Considering the missing data for the UK participants for the third session, not much can be said about their progress.

An ANOVA test was run to find out if there were any significant differences related to the countries and the sessions on Word Recognition. The two main effects were significant: country (F(4) = 3.114; p<0.02) and session (F(2)=48.946; p < 0.001). Additionally there was a significant interaction between country and session (F(7)=2.453;p<0.02).

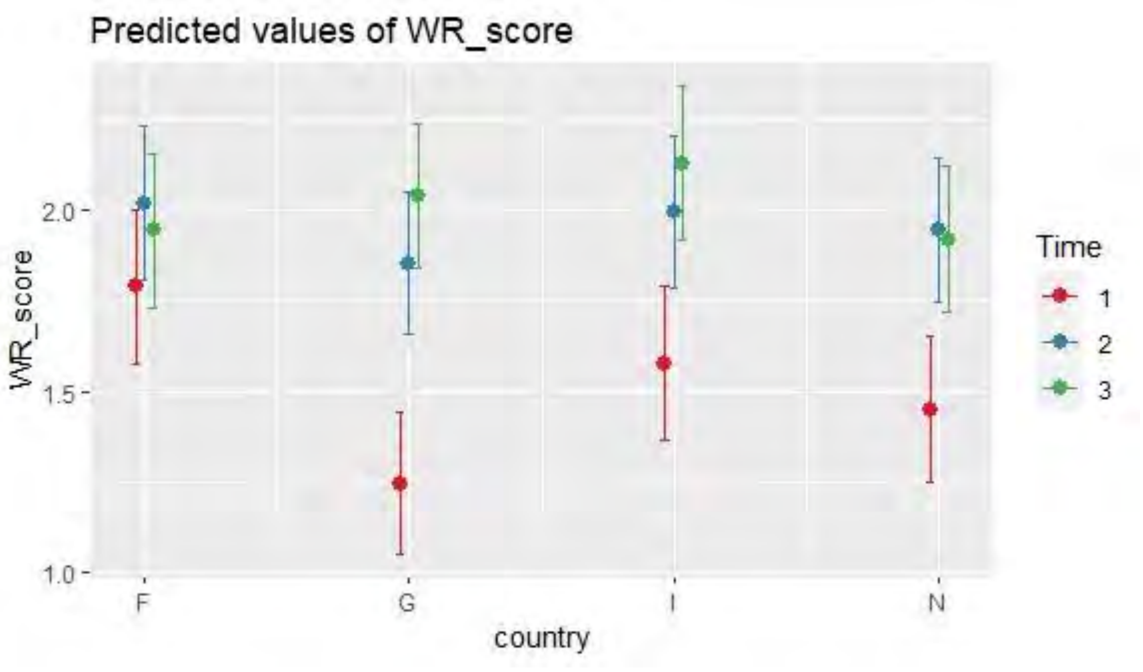
In the next step, we carried out a pairwise comparison according to the Tukey method, with the Kenward-Roger correction for *dfs* (package *emmeans*). Significant progress was made within country by Germany, Italy, and the Netherlands between the sessions S1 (T0) and S2 (T5) and the sessions S1 and S3 (T9). In none of the countries, significant progress was made between the sessions S2 (T5) and S3 (T9).

Table 4.1. Overview of progress in Word Recognition, made throughout the sessions in the M-group, all countries.

Country	Progress made (S1-S2)	Progress made (S2-S3)	Progress made (S1-S3)
France	No significant progress	No significant progress	No significant progress
Germany	**t(153)= –5.569; p<0.001**)	No significant progress	**t(155)= –7.165; p<0.001**)
Italy	**t(153)= –3.537; p<0.04**)	No significant progress	**t(153)= –4.663; p<0.001**)
The Netherlands	**t(153)= –4.423; p<0.002**)	No significant progress	**t(153)= –4.179; p<0.004**)
The UK	No significant progress	No data available for T3	No data available for T3

Moreover, looking at the default settings of the participants (S1) of the five countries, significant effects were found between France and Germany (t(180)=3.644; p<0.03) and France and the UK (t(180)=3.488; p<0.04), indicating that the French participants were significantly better at the first session of the Word Recognition task than both the German and the British participants, which is interesting given the fact that France did not show any significant progress throughout the sessions. Table 4.1, and Appendix K-1 provides more details on progress over time per country. This progress and the differences between the countries are visualized in the predicted scores and their confidence intervals in Figure 4.2.

Figure 4.2 Plot of d' scores and their confidence intervals for the M-group of Word Recognition. All sessions, all countries except the UK [23]

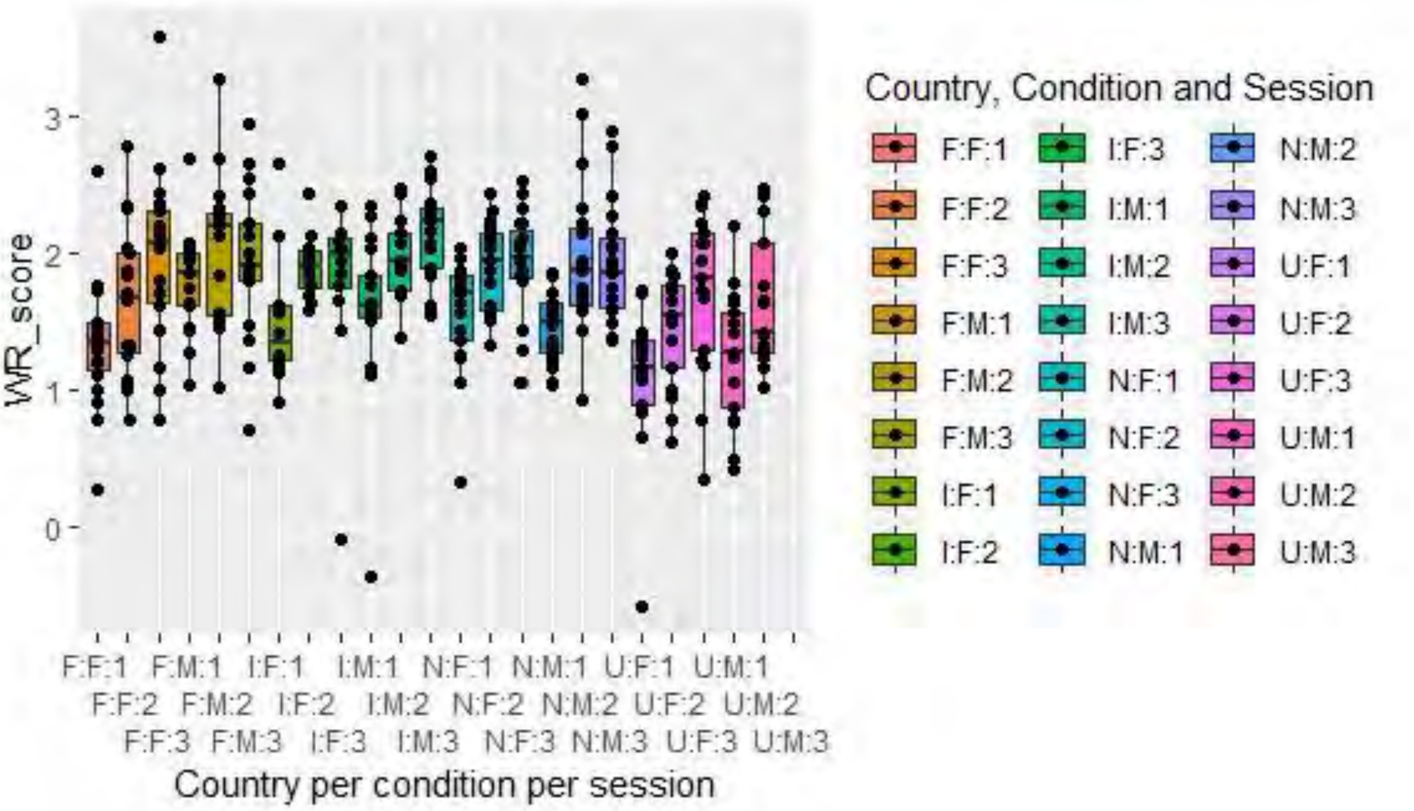


4.2 The effects of time and learning condition in four countries

Figure 4.3 visualizes the results for the four countries involved, with three time points (T0, T5 and T9) and two learning conditions (Form-Focused and Meaning-Based).

[23] The UK-data of the M-group is missing and therefore, in order to obtain a more comprehensible plot, this plot is based on an analysis done without the UK.

Figure 4.3 Word Recognition-scores across conditions and sessions; all countries except Germany (N= 128, with 34 missing values).



Note: The data set had 34 missing values that were not included in Figure 4.3.
This large number of missing values is mainly due to the fact that for unknown reasons, the UK-data from the third session of all of the participants in the MB -group seems to be missing.

Looking at figure 4.3, we can see that across the board, with the exception of the Netherlands, the participants of the MB-condition tend to start off with a higher score at the first test session of the Word Recognition task, compared to the participants in the FB-condition. For the Dutch participants, the inverse situation applies: the F-group starts off with a better score during the first session than the M-group. With a median of 1.7, the Netherlands seems to be the best-scoring country in the F-group in the first session (T0). For the M-group, this is France, with a median of 1.8. Whereas all countries in the F-group seem to have an upward trend in scores across the sessions, this is not the case for all countries in the M-group, where both France and the Netherlands seem to show a downward trend between session 2 (T5) and 3 (T9).

An ANOVA test was run to find out if there were any significant differences related to the countries, the sessions, and the learning conditions on Word Recognition. The three main effects were significant: country (F(89) = 5.325; p<0.001), session (F(63)=59.339;p < 0.001) and learning condition (F(47)=4.792; p< 0.03). Additionally, there was a significant interaction between country, session and learning condition (F(73)= 2.635; p<0.02).

In the next step, we carried out a pairwise comparison according to the Tukey method, with the Kenward-Roger correction for *dfs* (package *emmeans*). For the F-groups, significant progress was made within country by France and the UK between the session S1 and S3. For the M-groups, significant progress was made within country by Italy between S1/S3 and the Netherlands between S1/S2 and S1/S3. No significant effects were found within the countries concerning the learning conditions.

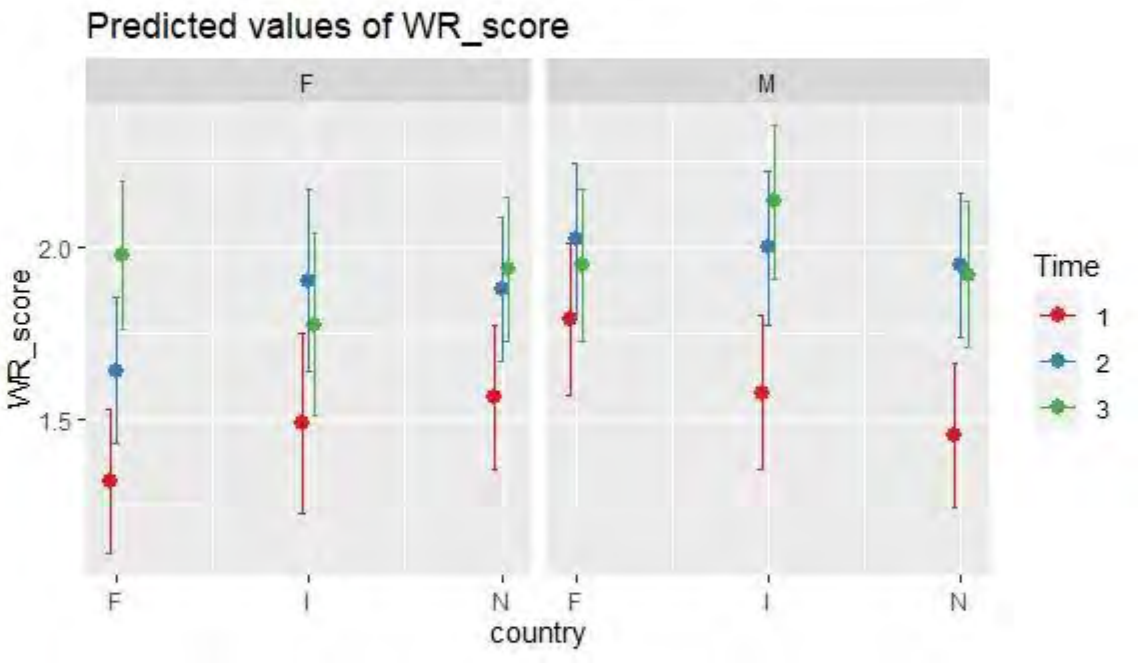
One significant effect across countries was found for the first testing session between The UK and France, where the French participants in the M-group outperformed the British participants in the F-group (F(264)=-4.290; p<0.005). Table 4.2, and Appendix K-2 provides more details on progress over time per country.

Table 4.2. Overview of progress made on Word Recognition throughout the sessions in the F- and M-groups. All countries except Germany.

Country	Progress (S1- S2) F- group	Progress (S1-S2) M- group	Progress S2-S3) F/ M- group	Progress (S1-S3) F-group	Progress (S1-S3) M-group
France	Not significant	Not significant	Not significant	**t(245)=–5.680; p< 0.001**	Not significant
Germany	Excluded	Excluded	Excluded	Excluded	Excluded
Italy	Not significant	Not significant	Not significant	Not significant	**t(240)= –4.624; p<0.002**
The Netherlands	Not significant	**t(240)= –4.386**; **p< 0.04**	Not significant	Not significant	**t(240)= –4.144; p<0.01**
The UK	Not significant	Not significant	No data available for M-T3	**t(240)= –5.214**; **p< 0.001**	No data available for M-T3

Progress and differences between the countries are visualized in the predicted scores and their confidence intervals in Figure 4.4.

Figure 4.4 Plot of d' scores and their confidence intervals on Word Recognition. Two conditions, all sessions, all countries except Germany and the UK[23].



4.3 Conclusion

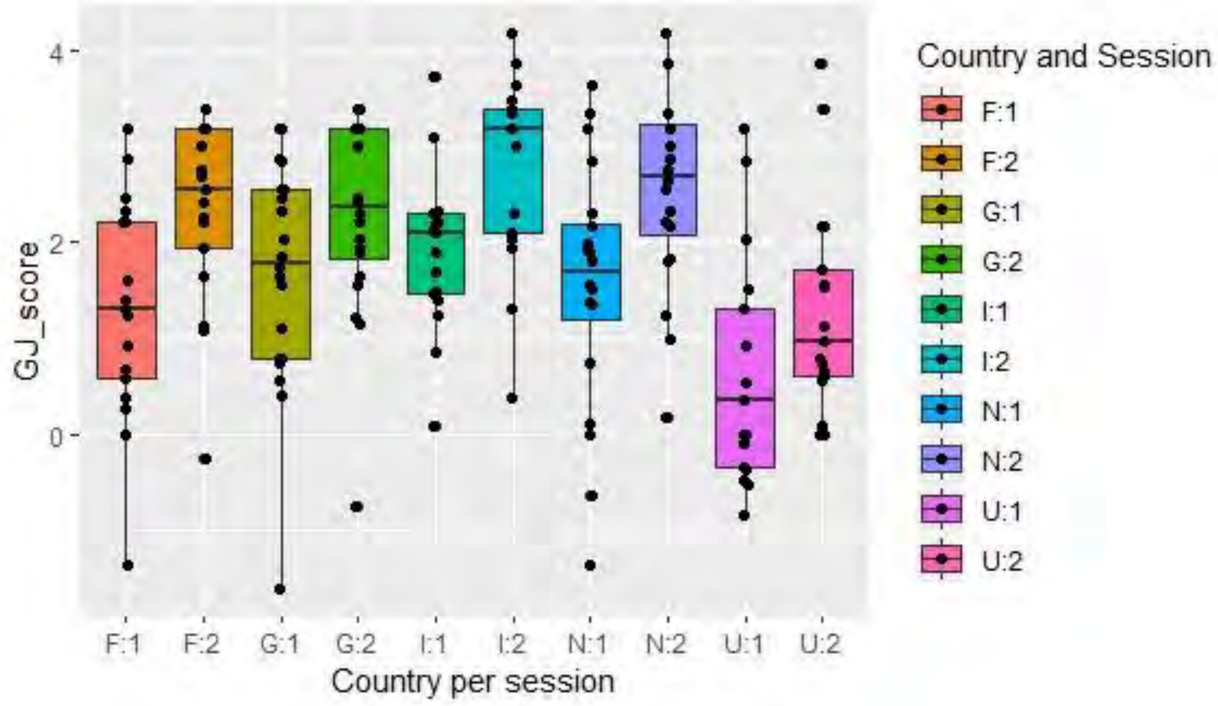
Looking at the results on the Word Recognition task we can conclude that there were no significant effects of learning condition within the countries. Performance wise, some interesting effects were found. In the analysis of the five countries in the MB-condition, the French participants scored significantly higher in the first session (T0) than both the German and the British participants did, indicating that by default their Word Recognition skills were better than their German and British counterparts, considering the fact that none of the respondents had been exposed to the Polish language at that point. In the MB-condition, significant progress was made by Italy between sessions S1(T0) and S3 (T9), after 13,5 hours of Polish input. The Dutch participants made significant progress both after 4,5 hours of Polish input, between session S1 and S2, and between S1 and S3, after having been exposed to the language for 13,5 hours. In the F-condition, significant progress was made by both France and the UK between the first and the last session, after 13,5 hours of Polish input.

5. Grammaticality Judgment

5.1 The effects of time and country in the MB (meaning-based) condition

The results for the M-group (MB-condition) are visualized in Figure 5.1, with two time points (T3 and T7) and five countries. As explained in Chapter 3, the performance of the participants was measured with d'.

Figure 5.1 Grammaticality Judgment scores of the M-group per session per country (N= 162 with 0 missing values).



Looking at the boxplots in Figure 5.1, we can see that all countries seem to have made some progress in session 2. The UK is the poorest performing country both in terms of their starting position (M=0.5) and their improvement made (M=1.0). The best performing country in that respect is Italy with medians of around 2.0 and 3.3 respectively. Not surprisingly, considering the fact that the two languages are closely related, Germany and the Netherlands seem to perform equally well on the first session of grammaticality judgment task with Median scores of around 1.7, however the Netherlands seem to slightly outperform the Germans when it comes to improvement made (M= 2.6 and M= 2.4 respectively). Italy and the Netherlands each have outliers in the first session. In the second session, outliers can be found in France, Germany, the Netherlands, and the UK. Across the board, countries seem to have all improved in session 2. France is the only country with a median below 15 and outliers scoring less than 10.

An ANOVA test was run to find out if there were any significant differences related to the countries and the sessions on Grammaticality Judgment. There were significant effects on country (F(4) = 4.155; p<0.004) and session (F(1)=224.325; p < 0.001) Additionally there was a significant interaction between country and session (F(4)=2.626;p<0.040).

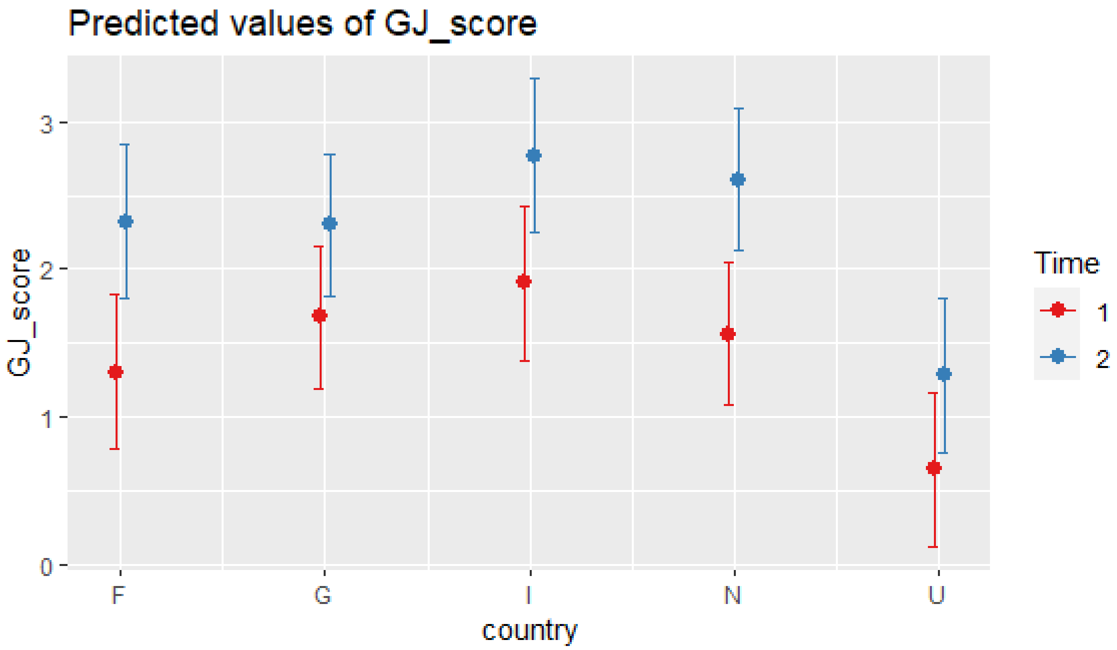
In the next step, we carried out a pairwise comparison according to the Tukey method, with the Kenward-Roger correction for *dfs* (package *emmeans*). Significant progress was made by all countries between the sessions (p<0.001). Table 5.1, and Appendix K-3 provides more details on progress over time per country.

Table 5.1. Overview of progress made throughout the sessions in the M-groups. All countries.

Country	Progress made (S1-S2)
France	t(86)= –7.897; p<0.001)
Germany	t(86)= –5.251; p<0.001)
Italy	t(86)= –6.644; p<0.001)
The Netherlands	t(86)= –8.785; p<0.001)
The UK	t(86)= –4.988; p<0.001)

Moreover, the Italian participants outscored the UK participants in both sessions (t(96.7)=3.363; p<0.04) and t(96.7)= 3.928; p<0.006). In the second session, the Dutch participants had significantly higher scores compared to the British participants (t(96.7)= 3.647; p<0.002). The differences between the countries along with their progress are visualized in the predicted scores and their confidence intervals in Figure 5.2.

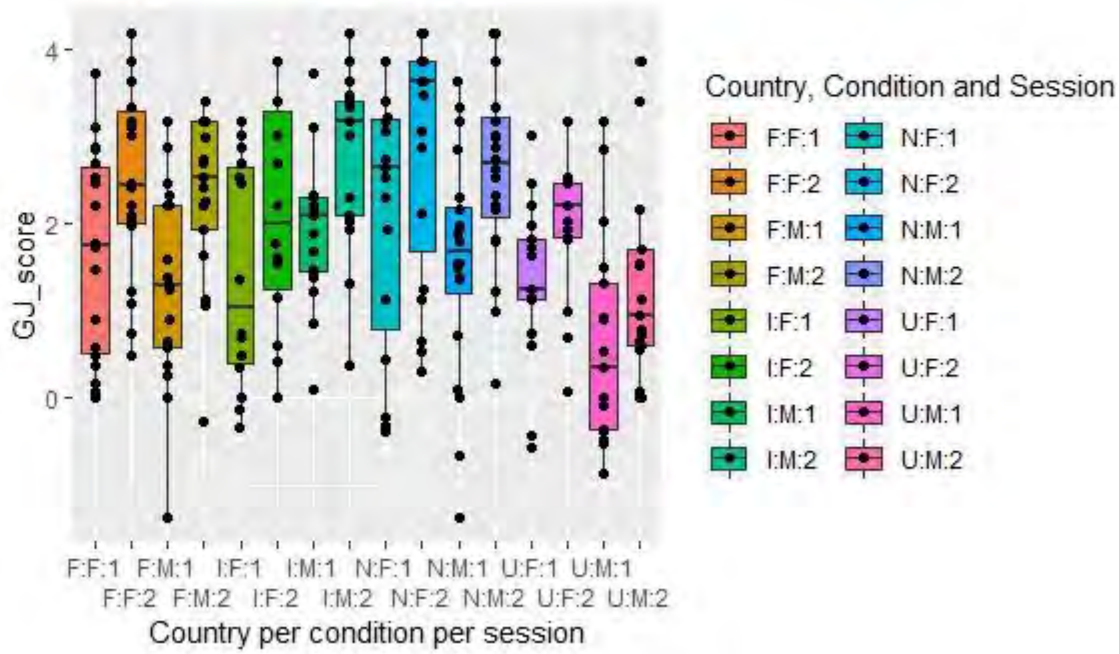
Figure 5.2 Plot of d' scores and their confidence intervals on Grammaticality Judgment for the M-group. All sessions, all countries.



5.2 The effects of time and learning condition in four countries

Figure 5.3 visualizes the results for the four countries involved, with two time points (T3 and T7) and two learning conditions (Form-Focused and Meaning-Based).

Figure 5.3 Grammaticality Judgment scores across two conditions and sessions.
All countries except Germany. (N= 156, with 6 missing values)



Looking at Figure 5.3, we can see that countries across the board made progress between the sessions, independent of learning condition. With a median of close to 2.7 as a default score, the Dutch participants in the F-group seem to outperform the other countries in that group. In the M-group the best performing country in the first session seems to be Italy (M=2.1). With respect to the learning condition, most countries seem to do better with the form-focused input (France, the Netherlands, and the UK), whereas Italy seems to get higher initial scores with the meaning-based input, without implicit grammar instructions.

An ANOVA test was run to find out if there were any significant differences related to the countries, the sessions and the learning conditions on Grammaticality Judgment. There were significant effects on country (F(3) = 4.685; p<0.004) and session (F(1)=370.6035;p < 0.001), but there were no significant effects on learning condition, nor were there any significant interactions.

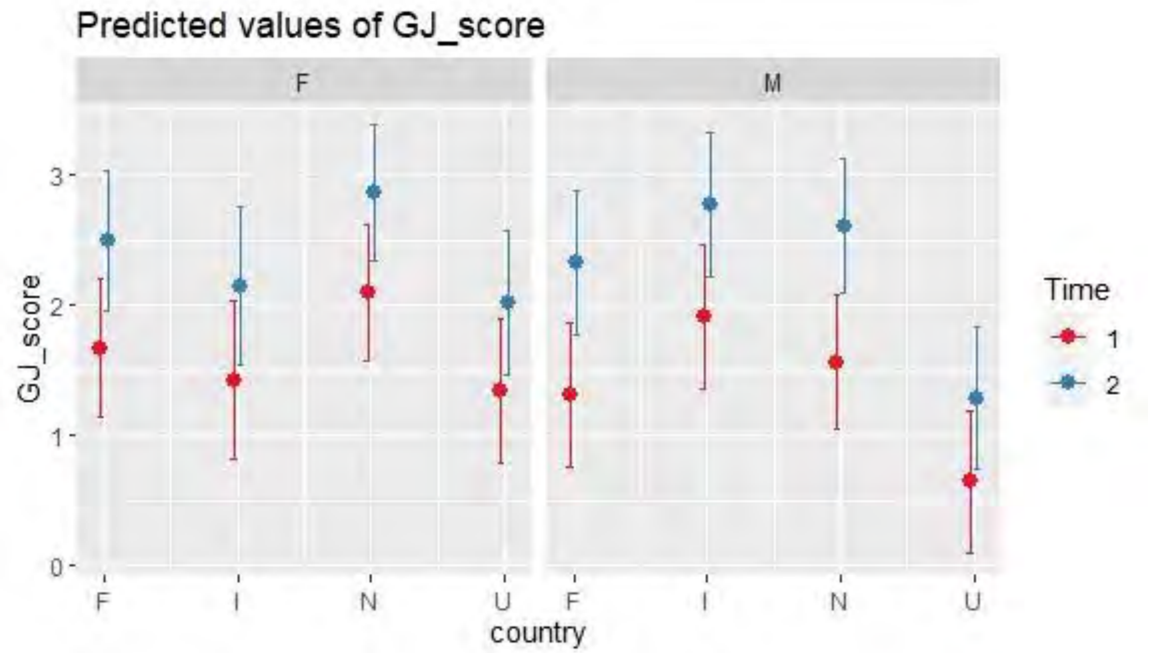
In the next step, we carried out a pairwise comparison according to the Tukey method, with the Kenward-Roger correction for *dfs* (package *emmeans*), showing significant progress made by all countries between the sessions (p<0.001). Moreover, the results from the second session of the UK participants in the F-group were significantly better than the results from the first session of the UK participants in the M-group (t(144)=3.454; p<0.05). The Dutch participants from the F-group scored significantly better during both sessions than the UK participants from the M-group in those sessions (t(144) =3.747; p<0.02 and t(1440=4.075; p<0.007). In the M-group, both the Dutch and the Italian participants outperformed the UK-participants in the second session (t(144)=3.729; p<0.02) and t(144)=3.462; p<0.05). Table 5.2, and Appendix K-4 provides more details on progress over time per country.

Table 5.2. Overview of progress made throughout the sessions in the F- and M-groups. All countries except Germany.

Country	Progress made (S1- S2) F-group	Progress made (S1-S2) M-group
France	**t(131)= –7.034; p< 0.001**	**t(131)= –8.404; p< 0.001**
Germany	Excluded	Excluded
Italy	**t(131)= –5.390; p< 0.001**	**t(131)= –7.070; p< 0.001**
The Netherlands	**t(131)= –6.718; p< 0.001**	**t(131)= –9.384; p< 0.001**
The UK	**t(131)= –5.591; p< 0.001**	**t(131)= –5.308; p< 0.001**

Progress and differences between the countries are visualized in the predicted scores and their confidence intervals in Figure 5.4.

Figure 5.4 Plot of d' scores and their confidence intervals on Grammaticality Judgment.
Two conditions, all sessions, all countries except Germany.



5.3 Conclusion

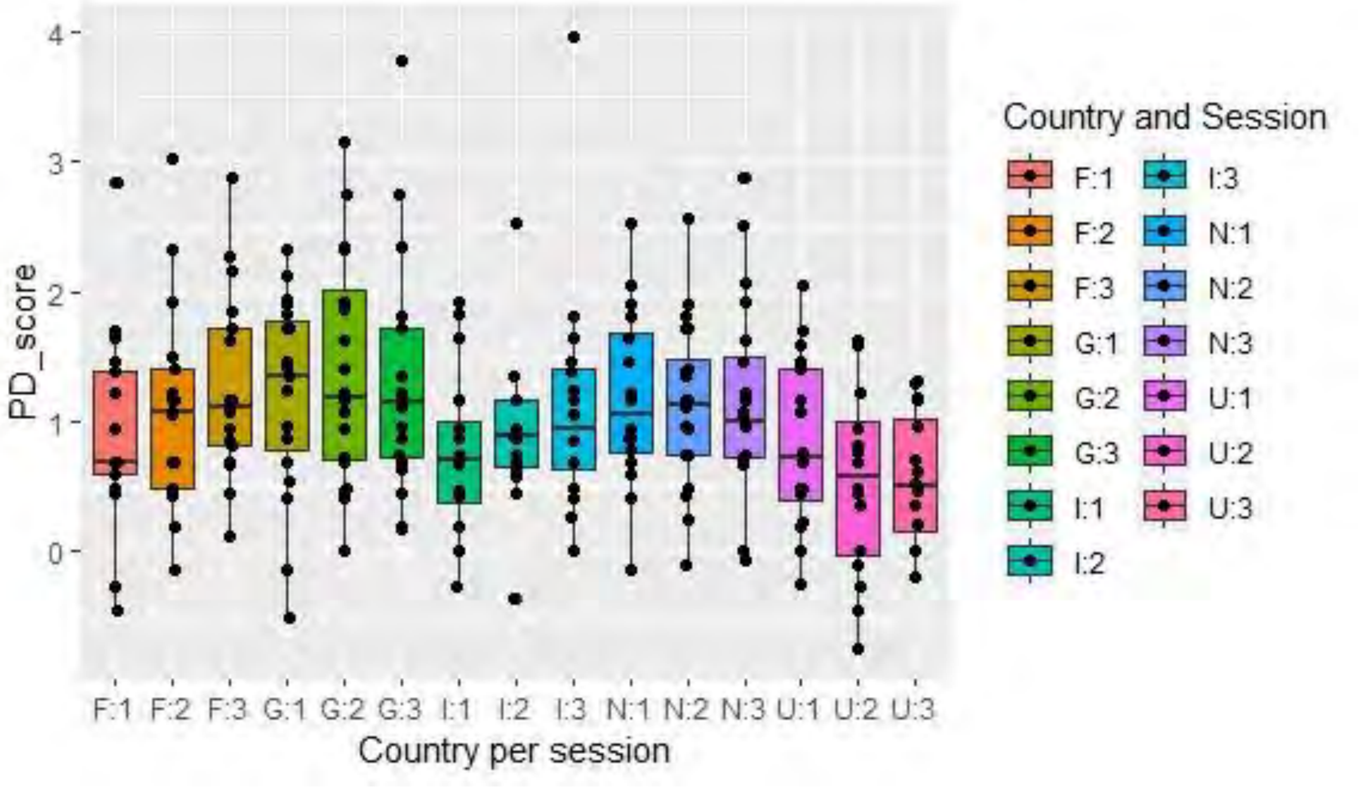
Looking at the results on the Grammaticality Judgment task we can conclude that there were no significant effects of learning condition within the countries. Performance wise, some interesting effects were found. In the second session of the M-group, both the Dutch and the Italian students scored significantly higher than the British participants. The main overall result is that all countries made significant progress.

6. Phoneme Discrimination

6.1 The effects of time and country in the MB (meaning- based) condition

The results for the M-group (MB condition) are visualized in Figure 6.1, with three time points (T0, T3 and T7) and five countries. As explained in Chapter 3, the performance of the participants was measured with d'.

Figure 6.1 Phoneme Discrimination scores M-group per session per country (N= 155,with 7 missing values).

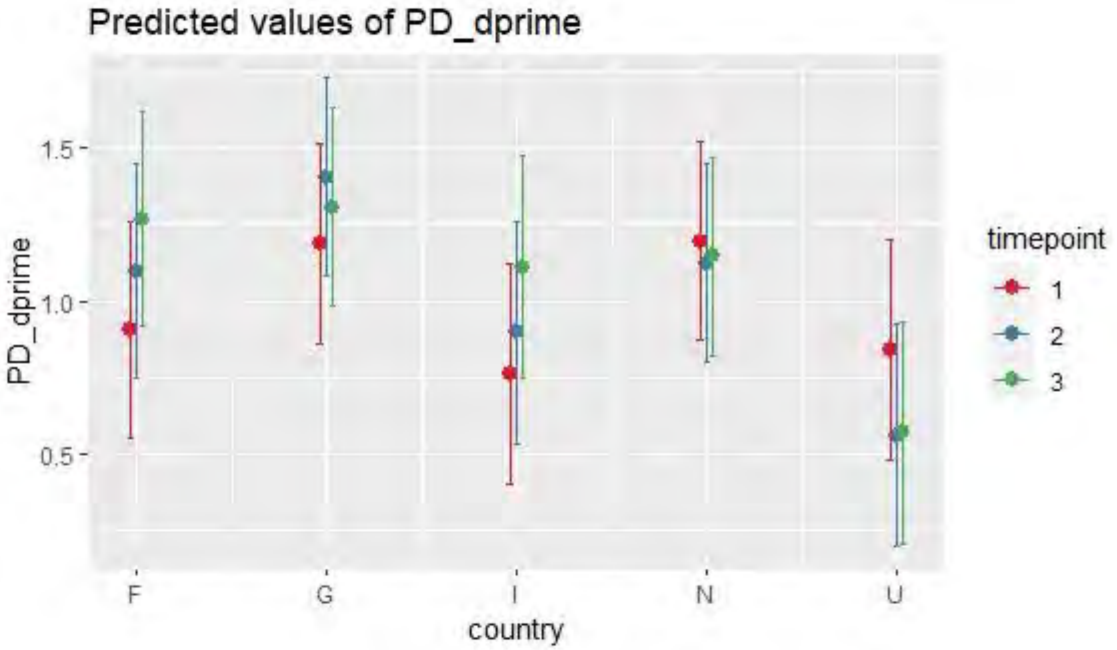


Looking at the data in Figure 6.1 we can see some countries progressing throughout the sessions (France and Italy) and others following a pattern of an initial good score, followed by a declining trend (Germany and the UK). The Netherlands sticks out in that sense, as they initially show some progress in their results between T1 and T2 after which they return to more or less the same result obtained in the first session. With a median of almost 1.3, the German participants seem to have the best initial score on the Phoneme Discrimination task in the first session, whereas the Italian participants seem to have the lowest initial score (M= 0.7).

An ANOVA test was run to find out if there were any significant differences related to the countries and the sessions on Phoneme Discrimination. There were significant effects on country ((F(4) = 2.6531; p<0.04), but none on session, nor were there any significant interactions between country and session.

In the next step, we carried out a pairwise comparison according to the Tukey method, with the Kenward-Roger correction for *dfs* (package *emmeans*), showing that here were in fact no significant effects at all. Appendix K-5 provides more details on progress over time per country. The plot of predicted scores and their confidence intervals is shown in Figure 6.2

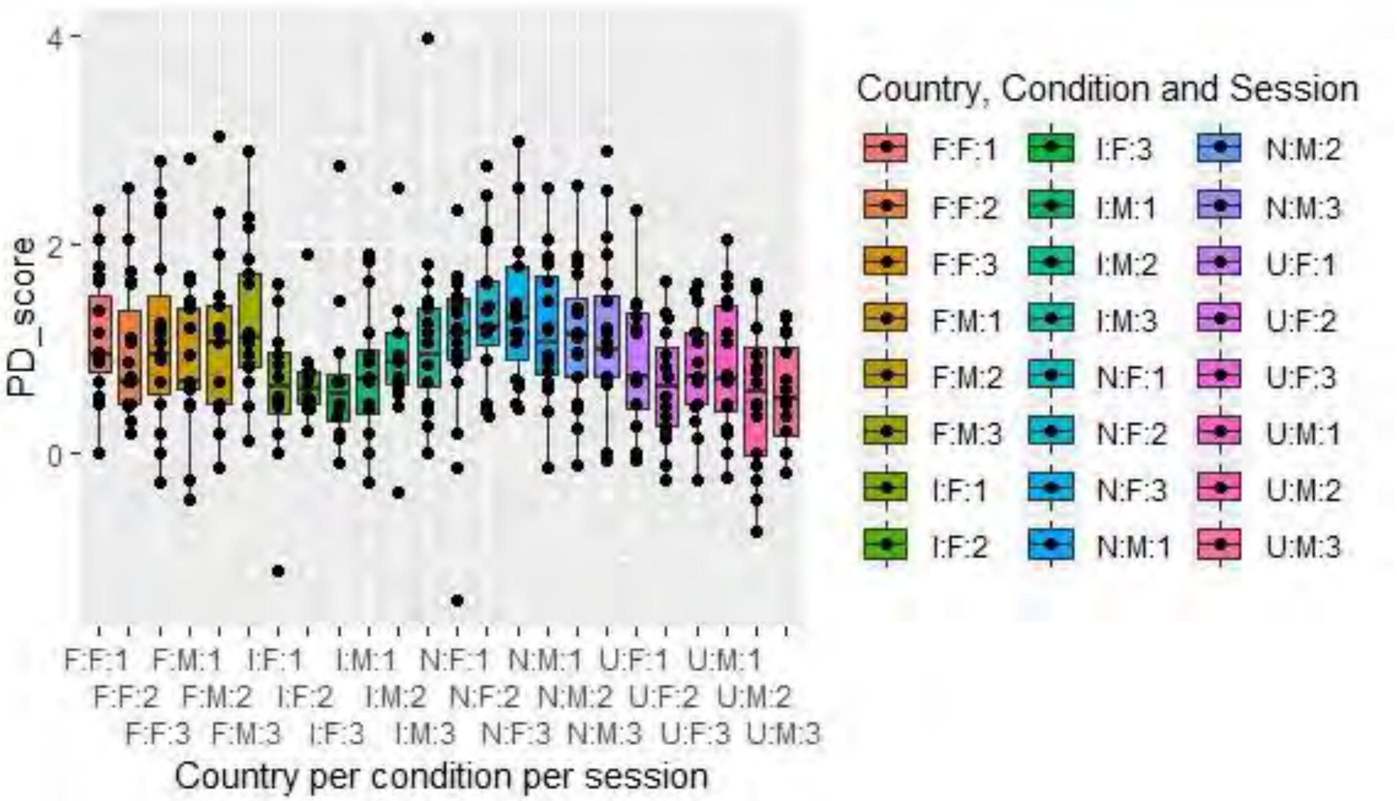
Figure 6.2 Plot of d' scores with their confidence intervals for the M-group of Phoneme Discrimination. All sessions, all countries.



6.2 The effects of time and learning condition in four countries

Figure 6.3 visualizes the results for the four countries involved, with three time points (T0, T3 and T7 and two learning conditions (Form-Focused and Meaning-Based).

Figure 6.3 Phoneme Discrimination. scores across conditions and sessions, all countries except Germany (N= 1655, with 7 missing values)

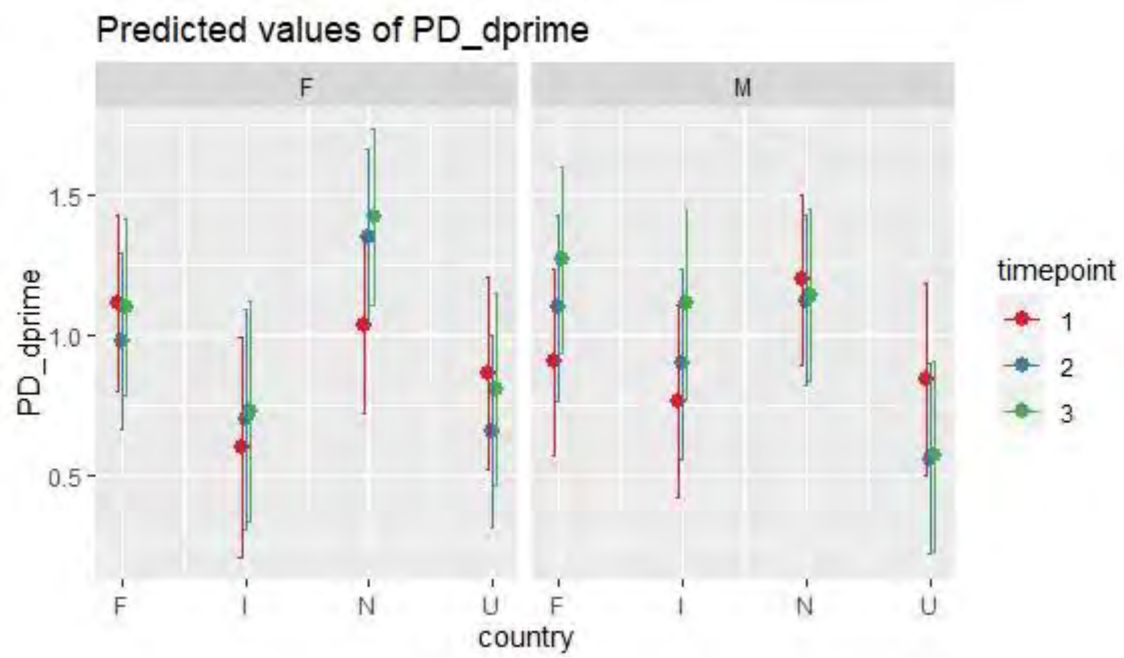


Looking at Figure 6.3, we can see that there is quite a bit of overlap in the results of the various countries throughout the sessions, indicating that countries seem to do equally well on this task. Interestingly, countries don't seem to make a lot of progress over time and some even seem to regress. There seem to be no obvious differences between the learning conditions. Looking at the outliers, some participants seem to be extremely good at this task with scores ranging between 2.5 and 4.0, whereas others perform exceptionally poor with scores ranging between -0.2 and -1.5).

An ANOVA test was run to find out if there were any significant differences related to the countries, the sessions and the learning conditions on Phoneme Discrimination. There were significant effects on country (F(3) = 5.47756; p<0.001)., but none on session, nor were there any significant interactions between country and session, country and learning condition, session and learning condition, or country, session and learning condition.

In the next step, we carried out a pairwise comparison according to the Tukey method, with the Kenward-Roger correction for *dfs* (package *emmeans*), showing that here were in fact no significant effects at all). Appendix K-6 provides more details on progress over time per country. The plot of scores, along with their confidence intervals is shown in Figure 6.4

Figure 6.4 Plot of d' scores and their confidence intervals on Phoneme Discrimination.
Two conditions, all sessions, all countries except Germany.



6.3 Conclusion

Looking at the results on the Phoneme Discrimination, we observe that there were no significant effects at all. Countries seem to score in more or less the same range, made no progress, nor did learning condition play any role. These results support the widespread idea that new phonetic contrasts cannot be learned or after a certain age. It is interesting however that some participants in the VILLA sample seem to be exceptionally skilled at this ability.

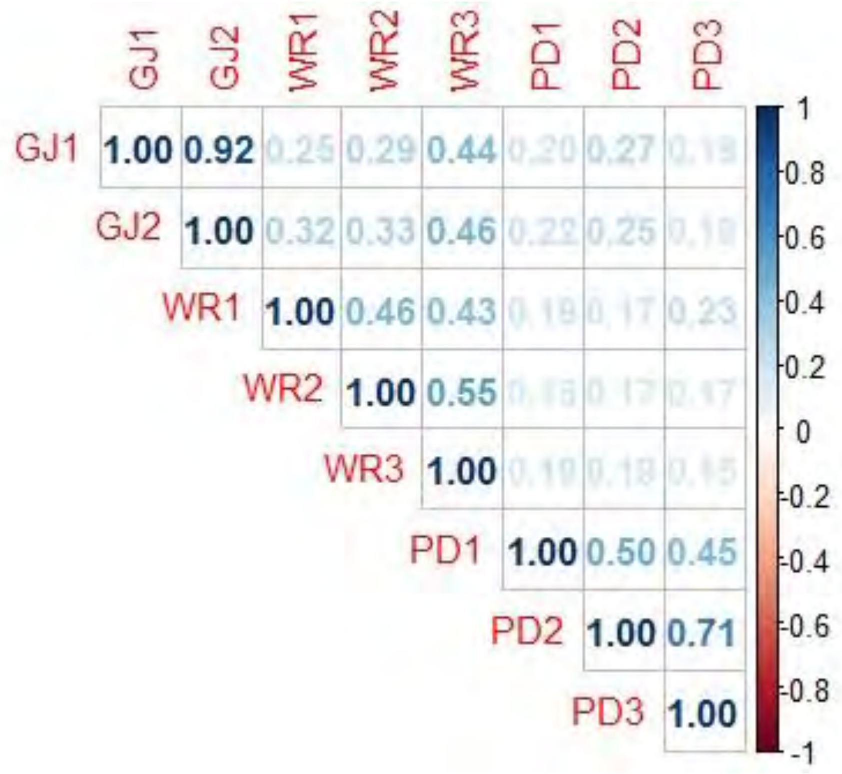
7. The effects of individual learner variables

In this chapter we would like to outline the effect of our selected learner variables on the results of the three language tasks in order to establish their predictive value. In sections 7.1 and 7.2 we have calculated the correlations between the language tasks (7.1) and the psychometric tests (7.2) to find out whether these tests measure unique constructs or whether there is a relationship between the tests. In section 7.3 the effects of a selection of learner variables are analyzed on each of our three selected language tasks. In order to limit the number of missing values, we decided to exclude the background data, the results on the personality tests (Barsch, Neo and ISA), and the results of the Llama B test. We did however include the core variable learning condition in order to analyze its predictive value on the outcomes of the language tasks. The variables we analyzed in this study were: Non-Verbal Intelligence (Raven), Working Memory (Digit Span and LNS), Inhibition Cost (Flanker), Phonological Recognition (Llama D), Grammatical Inferencing (Llama F), Motivation and Learning Condition. Section 7.4 covers our conclusions.

7.1 Correlations between the language tasks

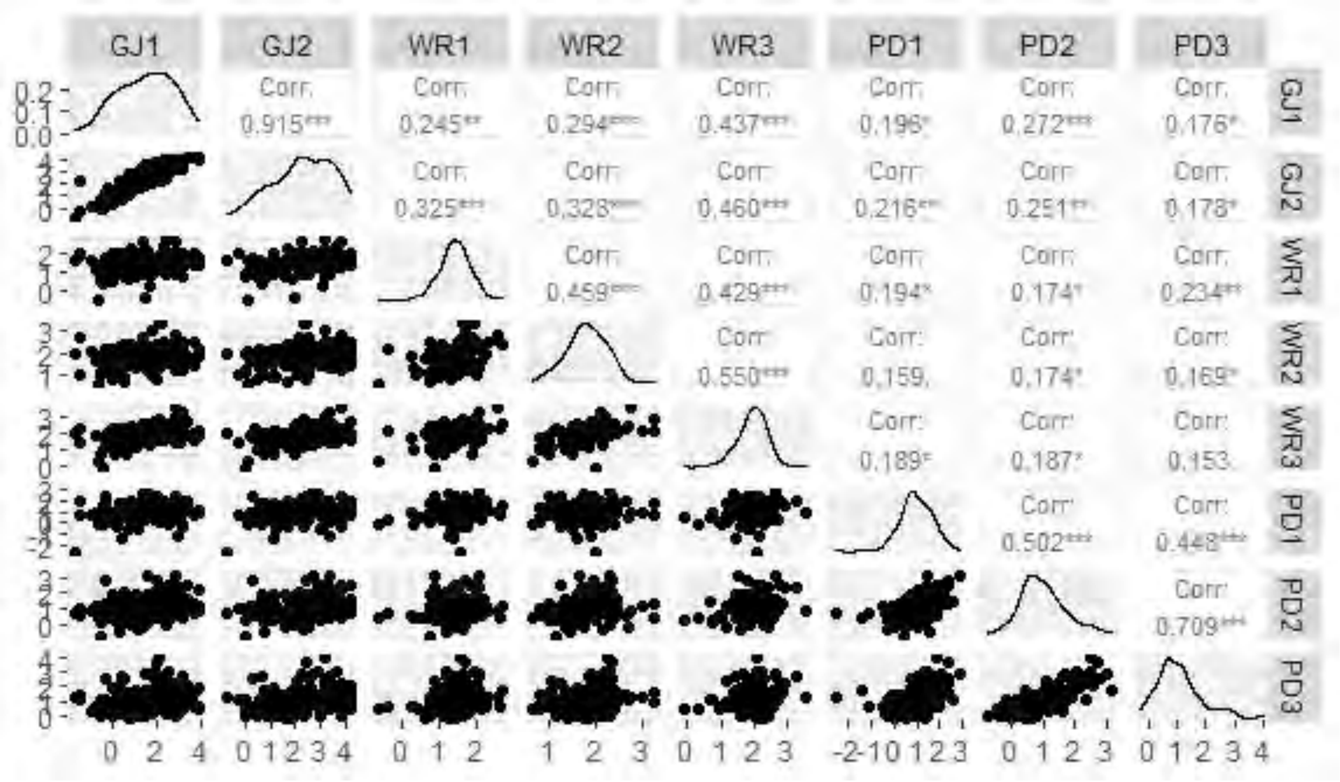
In order to establish whether the tasks used in this study measure different constructs, Pearson r correlations were calculated. The higher the correlation the stronger the relationship, with values of −1 and + 1 indicating a perfect linear relationship between the variables. A positive correlation describes the extent to which variables, in this case the tasks move in the same direction, whereas a negative correlation describes the extent to which variables move in the opposite direction. Figure 7.1 shows a correlation plot or a correlation matrix of the scores for each session of the three language tasks, in which both their correlation strength and their directionality are displayed.

Figure 7.1 Correlation plot of the scores on Grammatical Judgement (GJ), two sessions, Word Recognition (WR), three sessions, and Phoneme Discrimination (PD), three sessions.



Looking at Figure 7.1, we can see that the correlations between the tasks are relatively weak, indicating that the tasks do not measure similar constructs. Correlations found are positive, implying that if a candidate does well on one task, he/she is likely to do well on the other and vice versa, depending on the extent of the correlation between those tasks. A strong positive correlation can be found between the first and the second session of the Grammaticality Judgment task, confirming that both sessions are strongly related to each other and if candidates performed well in the first session, they are likely to perform well in the second session as well. Interestingly, another fairly strong positive correlation can be found, between the second and the third session of the Phoneme Discrimination task, indicating some level of predictability between those sessions. This can be explained by the fact that the first test session was done prior to the course, at zero hours of Polish input, whereas at the third test session candidates had been exposed to 10, 5 hours of Polish input. A similar effect can be found for the Word Recognition task, with a relatively strong positive correlation between the second and the third session. Another way of visualizing the correlations between the tasks is shown in Figure 7.2 with scatterplots displayed on the left side, variable distribution on the diagonal and Pearson correlation drawn on the right. We see no strange distributions of the task scores and no scatterplots that point to a non-linear relationship.

Figure 7.2 Correlation matrix of the scores on Grammaticality Judgment (GJ), two sessions, Word Recognition (WR), three sessions and Phoneme Discrimination (PD), three sessions



7.2 Correlations between the psychometric tests

Pearson correlations were calculated for the psychometric tests employed by the VILLA project and the results can be found in Figure 7.3 and 7.4.

Figure 7.3 Correlation plot of the scores on the 17 Psychometric tests

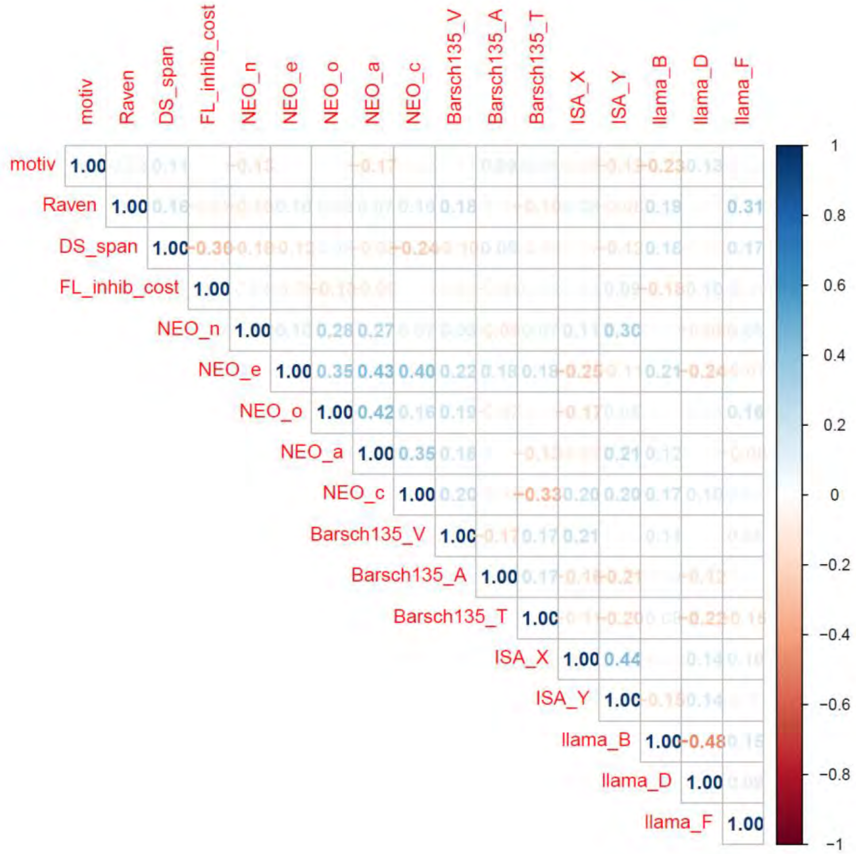
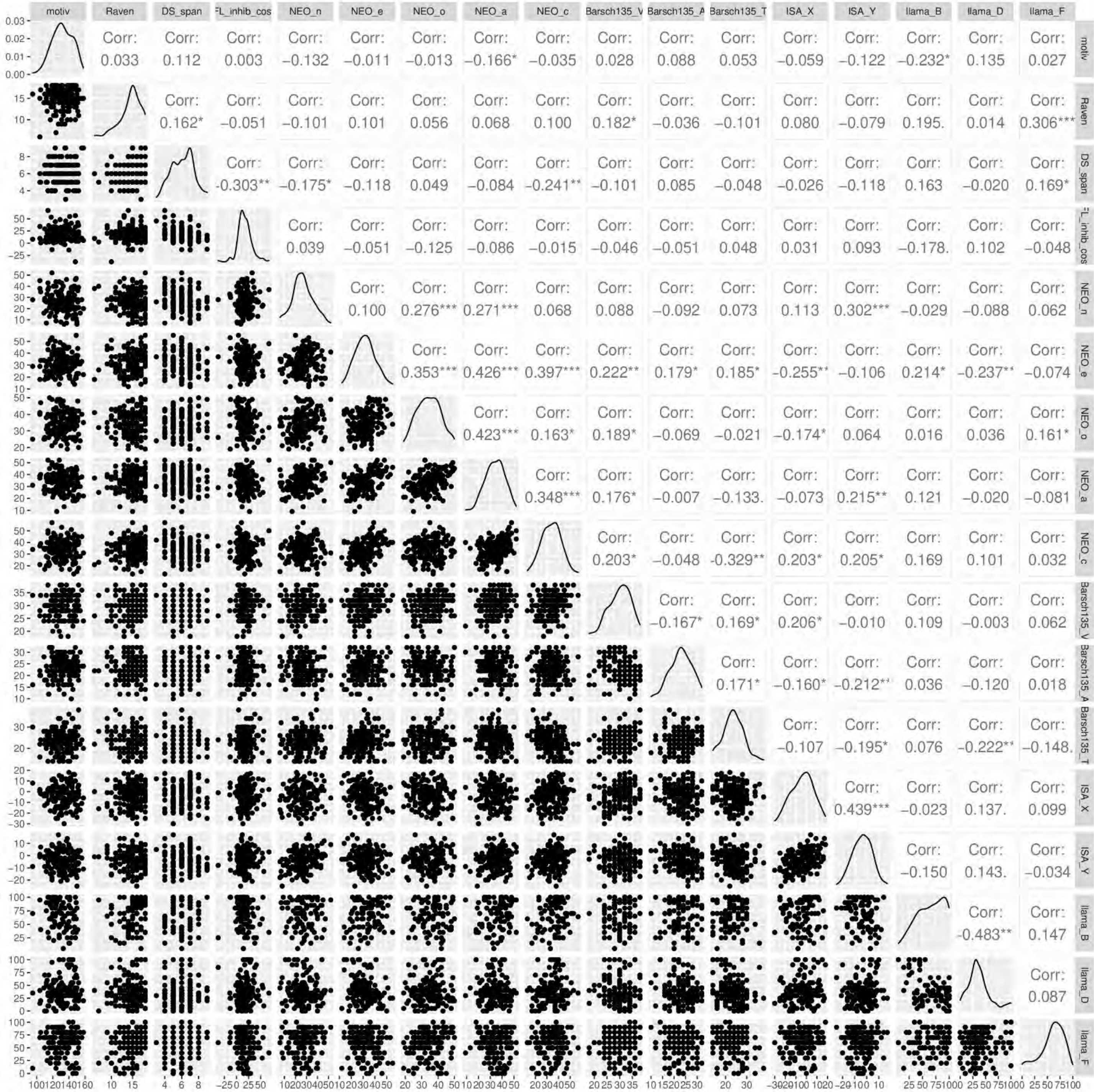


Figure 7.3 shows that most correlations between the psychometric tests are low. This could be a result of the fact that all of the participants in our sample were highly educated, hence limiting the possibility to study the variables in their full range. The highest ones can be found within the clusters belonging to the same set of tests for NEO, ISA and Llama. More details can be found in Figure 7.4. The scatterplots clearly visualize the lack of a clear relation. It also shows that Ll ama B has a remarkably skewed distribution.

Figure 7.4 Correlation matrix of the 17 Psychometric tests



Looking at Figure 7.4, we can see that the correlations between the tests are relatively weak, indicating that the tests do not measure similar constructs. Correlations found are both positive and negative. The strongest positive correlations found are between the test NEO-A, and NEO-E (0.43) and ISA-x and ISA-y (0.44)., suggesting a relationship between the personality factors agreeableness and extraversion, and the two axis of experiential preference respectively. For the NEO, a high score on agreeableness is likely to lead to a high score on extraversion and vice versa and for the ISA, a positive score on one axis is likely to lead to a positive score on the other axis and vice versa. The strongest negative correlation found is between LLAMA-B and LLAMA-D (−0.48), reflecting a negative relationship between vocabulary learning (LLAMA-B) and sound recognition (LLAMA-D) in the sense that if a candidate seems to perform well on one test, he /she is likely to get a lower score on the other. We have to be careful with our interpretation though, considering the large number of missing values and the skewed distribution, ands for this reason this learner variable was not included in the forward regression analysis.

7.3 Forward regression: predicting the language tasks scores

In this section we will outline the correlations between the selected learner variables and the results on the three language tasks. We will start with the effects on Word Recognition, proceed with the effects on Grammaticality Judgment and finish with the effects on Phoneme Discrimination.

7.3.1 The effects of learner variables on Word Recognition

Word Recognition was assessed at three timepoints, the first session (T0) was done prior to any Polish input, the second (T5) after 7,5 hours of language input and at last one (T9) after 13,5 hours of Polish. To find out which subset of our selected learner variables correlated the most with this language task, we analyzed them per session using a forward regression procedure to find the best model after which we applied a multilevel regression.

First Session

The results of the regression analysis for the first session of the Word Recognition task can be found in Tables 7.1 and 7.2 below. Table 7.1 presents the results of the forward regression, in which the AIC was used as the selection criterion. As long as the AIC decreases, new learner variables are added.

Table 7.1. Overview of the forward regression results of the first session of the Word Recognition task

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1		NA	NA	127	27.16693	-196.4038
2	+ llama_D	-1	1.8856421	126	25.28129	-203.6116
3	+ llama_F	-1	1.3806600	125	23.90063	-208.8001
4	+ motiv	-1	0.5952505	124	23.30538	-210.0283

Table 7.2 gives the parameters of the final model selected. In this case three learner variables were added. The strength of the correlation is expressed in R (√R²), and the R² outcomes are moderately strong.

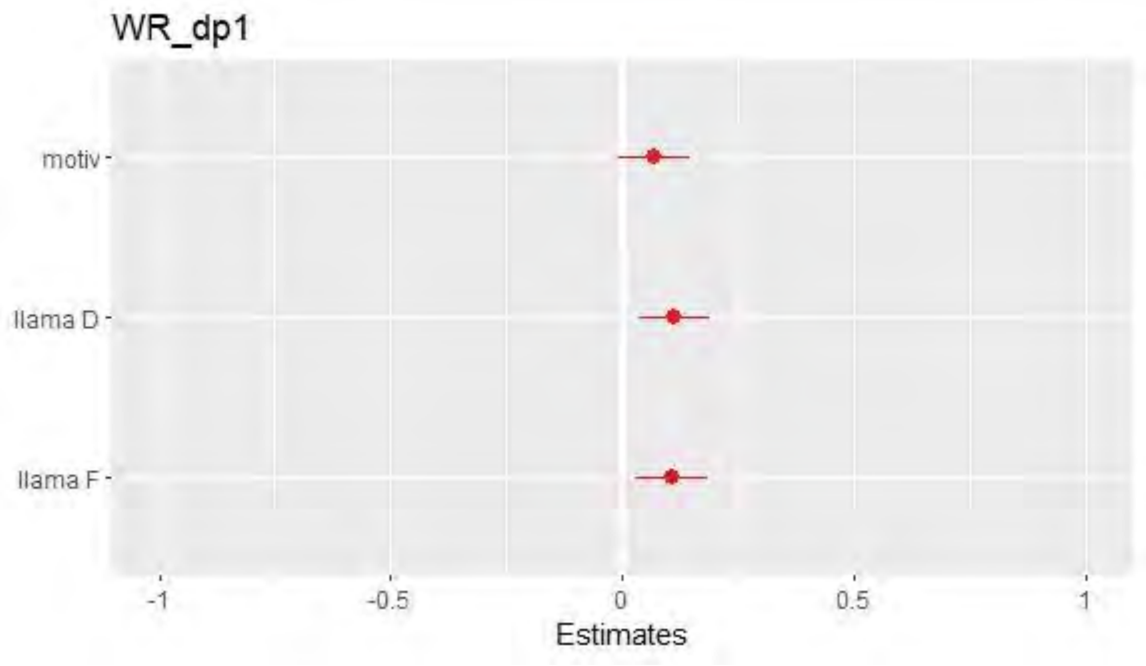
Table 7.2. Overview of the results of the Final Model of the first session of the Word Recognition task

	Word Recognition Sesssion 1		
Predictors	*Estimates*	*Confidence Interval*	*p*
(Intercept)	1.44	1.36-1.51	<**0.001**
Motivation	0.07	-0.01-0.15	0.078
Llama D	0.11	0.04-0.19	**0.004**
Llama F	0.11	0.03-0.18	**0.007**
Observations		128	
R^2/R^2 adjusted		0.142/0.121	

Figure 7.5 visualizes the results of the learner variables displayed in table 7.2 above, showing that in the first session of the Word Recognition task there were significant correlations with the learner variables Llama D (p<0.004) and Llama F (P<0.007), indicating that Phonological Recognition (Llama D) and Grammatical Inferencing (Llama F) were significant predictors for performance on the Word Recognition task in this session. These results are particularly interesting since

none of the participants had been exposed to any Polish input at this stage, so variables that do have predictive value are likely to lie at the heart of their ability to recognize words. Motivation also came up as predictor in our forward regression analysis, but it did not turn out to be significant in the regression analysis indicating that its effect is questionable, but that there is nevertheless an effect.

Figure 7.5. Overview of the results of the Final Model of the first session of the Word Recognition task



Second Session

The results of both regression methods for the first session of the Word Recognition task can be found in Tables 7.3 and 7.4 below. Table 7.3 presents the results of the forward regression, in which the AIC was used as the selection criterion. As long as the AIC decreases, new learner variables are added.

Table 7.3 Overview of the forward regression results of the second session of the Word Recognition task

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1		NA	NA	126	28.25753	-188.8590
2	+ motiv	-1	1.1163374	125	27.14119	-191.9781
3	+ group	-1	0.8870353	124	26.25416	-194.1981
4	+ llama_F	-1	0.6554385	123	25.59872	-195.4089

Table 7.4 below, gives the parameters of the final model selected. In this case three learner variables were added. The strength of the correlation is expressed in R (√R²) and is relatively weak.

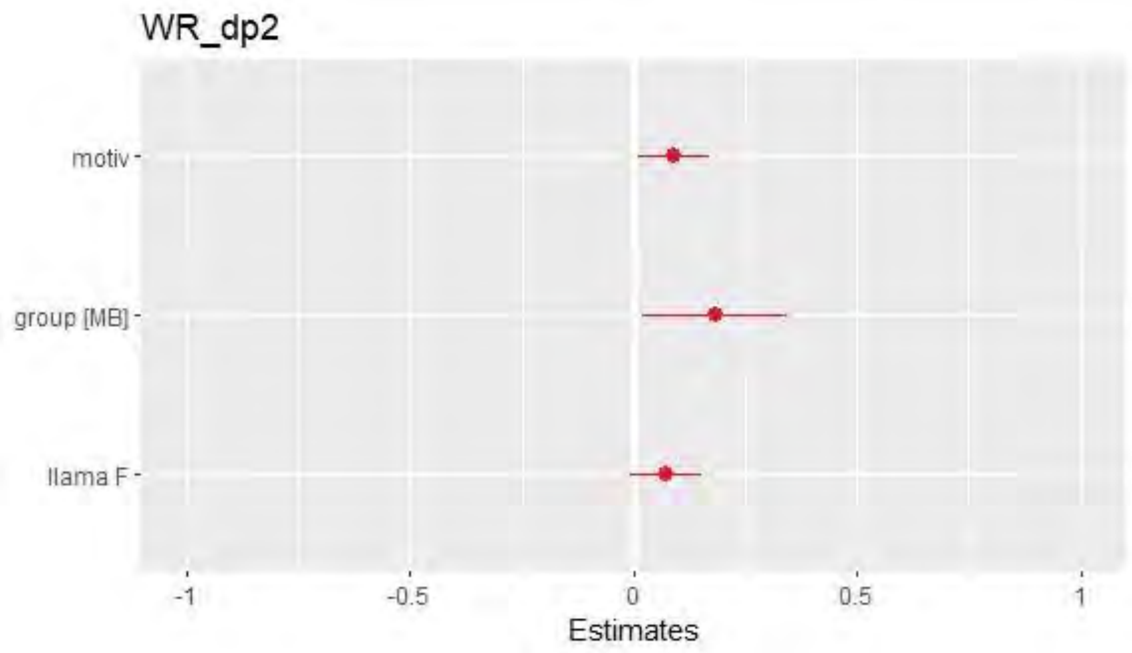
Table 7.4 Overview of the results of the Final Model of the second session of the Word Recognition task

	Word Recognition Sesssion 2		
Predictors	*Estimates*	*Confidence Interval*	*p*
(Intercept)	1.73	1.61-1.85	<**0.001**
Motivation	0.09	0.01-0.17	**0.030**
Group (MB)	0.18	0.02-0.34	**0.029**
Llama F	0.07	-0.01-0.15	0.078
Observations		127	
R^2/R^2 adjusted		0.094/0.078	

Figure 7.6 below, visualizes the results displayed in Table 7.4 above, showing that in the second session of the Word Recognition task there were significant correlations with the learner variables Meaning-Based Learning Condition (p<0.029) and motivation (P<0.030), indicating that these variables were significant predictors for performance on the Word Recognition task in this session. It is interesting to observe that whereas participants in the first session had to rely on existing phonological recognition and grammatical inferencing skills due to the absence of input, the most important predicting variables for the second session after 7,5 hours of input, are the meaning-based learning condition and motivation.

Llama F also showed up in our forward regression model of best fit, but did not turn out to be significant in our regression analysis, indicating that its effect is questionable, but that there is nevertheless an effect. Another effect that can be seen is the difference in the range of the confidence interval. For the meaning-based condition it is much wider than it is for the learner variables motivation and Llama F, and this could be an indication that the sample size was not big enough leading to a confidence range that is less specific about the predicted results.

Figure 7.6 Overview of the results of the Final Model of the first session of the Word Recognition task



Third Session

The results of both regression methods for the third session of the Word Recognition task can be found in Tables 7.5 and 7.6 below. Table 7.5 presents the results of the forward regression, in which the AIC was used as the selection criterion. As long as the AIC decreases, new learner variables are added.

Table 7.5. Overview of the forward regression results of the third session of the Word Recognition task

	Step	Df	Deviance	Resid. Df	. Dev	AIC
1		NA	NA	117	33.01247	-148.3083
2	+ llama_D	-1	2.5181304	116	30.49434	-155.6709
3	+ llama_F	-1	1.2781964	115	29.21614	-158.7237
4	+ motiv	-1	0.5571501	114	28.65899	-158.9956

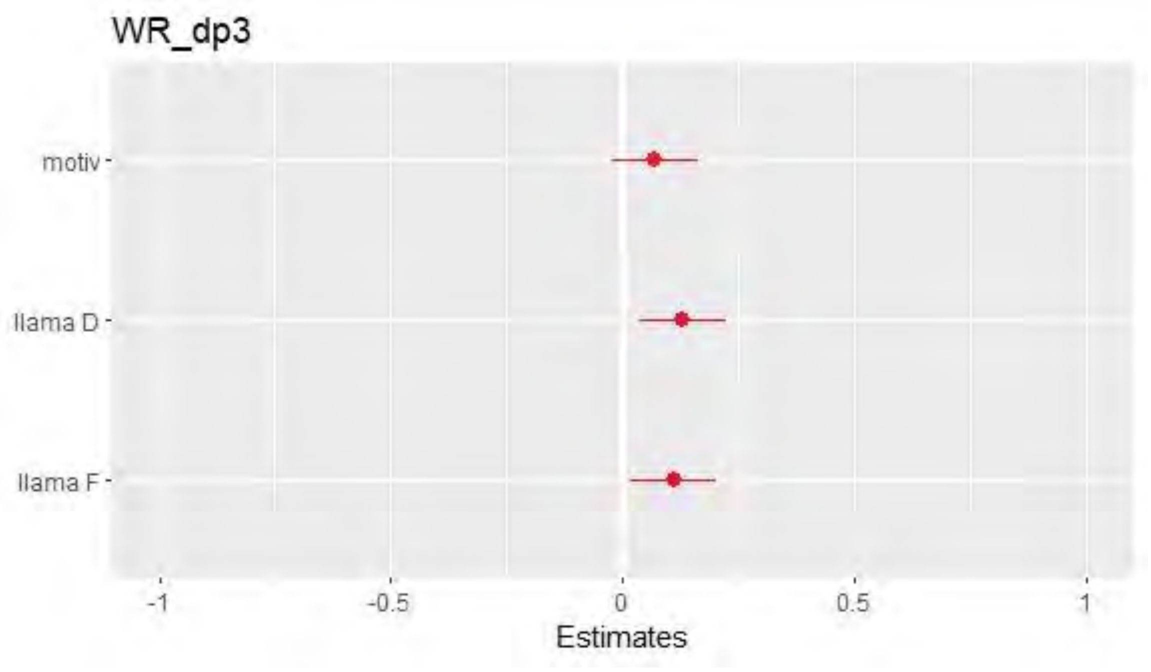
Table 7.6 below, gives the parameters of the final model selected. In this case three learner variables were added. The strength of the correlation is expressed in R ($\sqrt{R^2}$) and is moderately strong.

Table 7.6 Overview of the results of the Final Model of the third session of the Word Recognition task

	Word Recognition Sesssion 3		
Predictors	*Estimates*	*Confidence Interval*	*p*
(Intercept)	1.95	1.86-2.05	<**0.001**
Motivation	0.07	-0.02-0.16	0.139
Llama D	0.13	0.04-0.22	**0.007**
Llama F	0.11	0.02-0.20	**0.021**
Observations		118	
R^2/R^2 adjusted		0.132/0.109	

Figure 7.7 below, visualizes the results displayed in the table 7.6 above, showing that, similar to the first session of the Word Recognition task, there were significant correlations in the third session with the learner variables Llama D ($p<0.007$) and Llama F ($p<0.021$), indicating that these variables were significant predictors for performance on this task in this session. A possible explanation is that participants, after having been exposed to the language for a while where motivation and the meaning-based condition were small to moderate predictors of language success, go back to relying on their phonological recognition and grammatical inferencing skills that they already had prior to the course. It could be that in that second session students form their own hypothesis about the language that are either confirmed or discarded after some time in order to refine their phonological recognition and grammatical inferencing skills and with it their Word Recognition skills. However, the above is also quite speculative, considering the moderate value of the predictors.

Figure 7.7 Overview of the results of the Final Model of the first session of the Word Recognition task



7.3.2 The effects of learner variables on Grammaticality Judgment

Grammaticality Judgment was assessed at two timepoints, the first session (T3) was done after 4,5 hours of language input and at last one (T7) after 10,5 hours of Polish. To find out which of our selected learner variables correlated the most with this language task, we analyzed them per session using a forward regression procedure to find the best model after which we applied a multilevel regression.

First Session

The results of the regression analysis for the first session of the Grammaticality Judgment task can be found in Tables 7.7 and 7.8 below. Table 7.7 presents the results of the forward regression, in which the AIC was used as the selection criterion. As long as the AIC decreases, new learner variables are added.

Table 7.7 Overview of the forward regression results of the first session of the Grammaticality Judgment task

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1		NA	NA	129	205.6341	61.61329
2	+ llama_F	-1	19.641574	128	185.9925	50.56234
3	+ motiv	-1	14.282940	127	171.7096	42.17509
4	+ Raven	-1	6.388575	126	165.3210	39.24608
5.	+ FL_inhib_cost	-1	5.689674	125	159.6313	36.69321
6.	+ llama_D	-1	4.201962	124	155.4293	35.22538

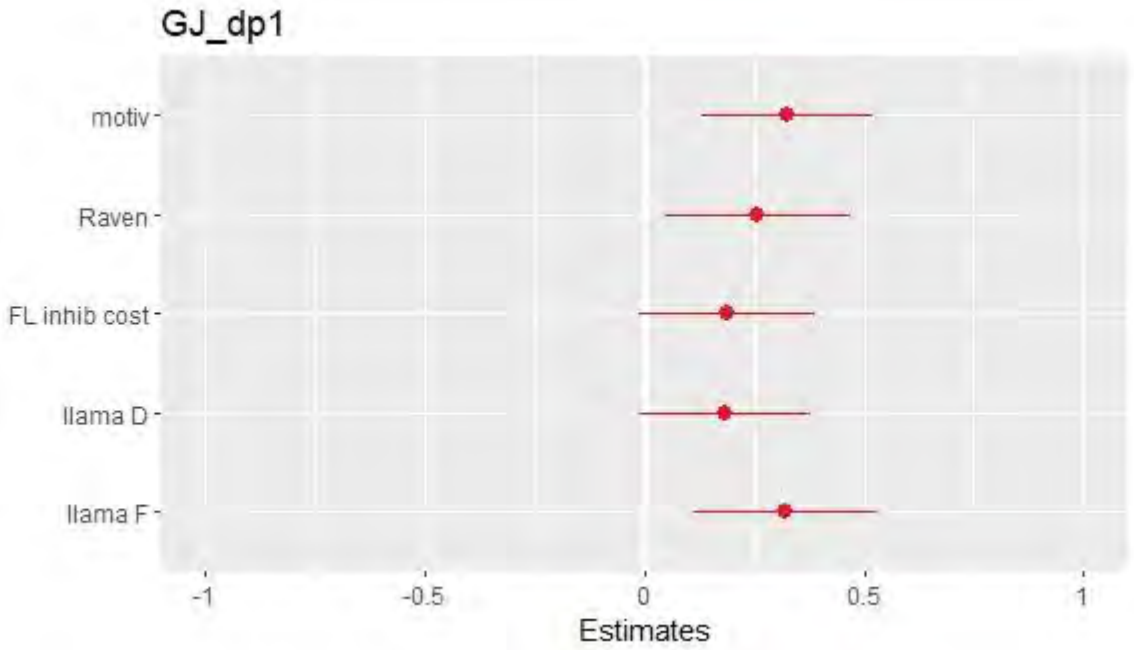
Table 7.8 gives the parameters of the final model selected. In this case three learner variables were added. The strength of the correlation is expressed in R (√R²) and is moderately strong, stronger than the correlations observed for Word Recognition.

Table 7.8 Overview of the results of the Final Model of the first session of the Grammaticality Judgment task

	Grammaticality Judgment Sesssion 1		
Predictors	*Estimates*	*Confidence Interval*	*p*
(Intercept)	1.56	1.36-1.75	<**0.001**
Motivation	0.32	0.13-0.52	**0.001**
Raven	0.26	0.05-0.47	**0.017**
Flanker Inhibition Cost	0.19	-0.01-0.39	0.065
Llama D	0.18	-0.01-0.38	0.070
Llama F	0.32	0.11-0.53	**0.003**
Observations		130	
R^2/R^2 adjusted		0.244/0.214	

Figure 7.8 below, visualizes the results displayed in the tables above, showing tat there are five learner variables that ha ve an
effect on predicting language learning success on this task for this session, of which there are three that have a significa nt effect (Motivation, Raven and Llama F). Flanker Inhibition cost and Llama D have confidence intervals including zero, making their effect debatable, but there is an effect, nonetheless. With estimates of 0.32, Motivation and Llama F have a stronger predictive value than Raven in this first session. Another effect that can be seen is the range of the confidence intervals of the variables. They all have identical confidence intervals, but they do seem on the wide side which makes them less specific about the predicted results. In sum, it seems that motivated students with a high level of Non-Verbal Intelligence and great Grammatical Inferencing skills are likely to do well on the Grammaticality Judgment task.

Figure 7.8 Overview of the results of the Final Model of the first session of the Grammaticality Judgment task



Second Session

The results of the regression analysis for the first session of the Grammaticality Judgment task can be found in Tables 7.9 and 7.10 below. Table 7.9 presents the results of the forward regression, in which the AIC was used as the selection criterion. As long as the AIC decreases, new learner variables are added.

Table 7.9 Overview of the forward regression results/second session of the Grammaticality Judgment task

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1		NA	NA	129	178.7185	43.37608
2	+ llama_F	-1	13.795486	128	164.9230	34.93277
3	+ motiv	-1	13.193853	127	151.7292	26.09315
4	+ llama_D	-1	7.121444	126	144.6077	21.84374
5	+ Raven	-1	4.693822	125	139.9139	19.55406

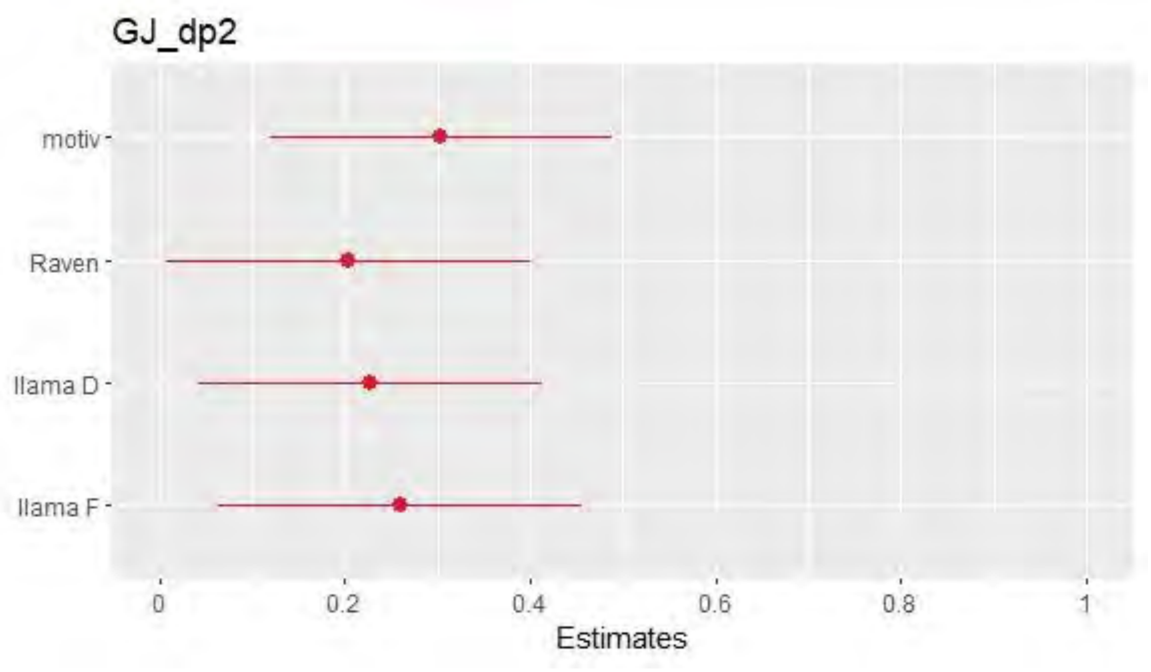
Table 7.10 below, gives the parameters of the final model selected. In this case three learner variables were added. The strength of the correlation is expressed in R (√R²) and is moderately strong.

Table 7.10 Overview of the results of the Final Model of the second session of the Grammaticality Judgment task

	Grammaticality Judgment Sesssion 2		
Predictors	*Estimates*	*Confidence Interval*	*p*
(Intercept)	2.40	2.21-2.58	<**0.001**
Motivation	0.30	0.12-0.49	**0.001**
Raven	0.20	0.01-0.40	**0.043**
Llama D	0.23	0.04-0.41	**0.017**
Llama F	0.26	0.06-0.46	**0.010**
Observations		130	
R²/R² adjusted		0.217/0.192	

Figure 7.9 below, visualizes the results displayed in the tables above, showing that of the five predictors that were relevant in the first session, only four remain relevant in this session. However, all of them are significant this time. Looking at the estimates, similar to the first session, Motivation and Llama F have the strongest predictive values, followed by Llama D and Raven. This could be an indication that Motivation and Grammatical Inferencing skills lie at the heart of the ability to do well on the Grammaticality Judgment task, considering the fact that these variables came out as moderately strong predictors for both sessions. It is interesting to observe that Non-Verbal Intelligence as a predictor seems less important in the second session and that the ability to recognize phonemes instead seems to have a stronger predictive value. Perhaps as students are exposed to the language longer, they finetune their existing linguistic abilities, such as phonological recognition and learn that phonological elements can contain clues to the grammar of the language, they need to rely less on non-verbal intelligence. However, the above is also quite speculative, considering the moderate value of the predictors.

Figure 7.9 Overview of the results of the Final Model of the second session of the Grammaticality Judgment task



7.3.3 The effects of learner variables on Phoneme Discrimination

Phoneme Discrimination was assessed at three timepoints, the first session (T0) was done prior to any Polish input, the second session (T3) was done after 4,5 hours of language input and at last one (T7) after 10,5 hours of Polish. To find out which of our selected learner variables correlated the most with this language task, we analyzed them per session using a forward regression procedure to find the best model after which we applied a multilevel regression.

First Session

The forward regression did not yield any results for the first session, indicating that none of the selected learner variables had any predictive effect on the results of this task. The results of the forward regression can be found in Table 7.11 below.

Table 7.11 Overview of the forward regression results of the first session of the Phoneme Discrimination task

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1	NA	NA	126	70.50205	72.74526

Second Session

Just like in the first session, the forward regression did not yield any results for the second session either, indicating that none of the selected learner variables had any predictive effect on the results of this task. The results of the forward regression can be found in Table 7.12 below.

Table 7.12 Overview of the forward regression results / second session of the Phoneme Discrimination task

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1	NA	NA	126	66.48414	-80.19741

Third Session

In line with the previous two session, there were no results from the forward regression for the third session, indicating that none of the selected learner variables had any predictive effect on the results of this task. The results of the forward regression can be found in Table 7.13 below.

Table 7.13 Overview of the forward regression results of the third session of the Phoneme Discrimination task

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1	NA	NA	126	84.42244	-48.36544

Perhaps these results signify that discrimination of unknown phonological distinctions is a skill on its own, which is not facilitated by any of our selected variables. And perhaps, keeping in mind that none of the participants in our sample really made any significant progress throughout the sessions, these results support the idea that the ability to discriminate between contrasts that were not acquired earlier is not something that can be learned automatically after a certain age. Nevertheless, several participants have high discrimination scores indicating that they can hear the distinction. That means that clear individual differences do exist between the participants, but this ability was not measured by the psychometric tests applied in VILLA.

7.4 Conclusion

Relatively weak correlations were found with both subsets of language tasks and psychometric tests, indicating that the three language tasks measure unique skills and that the learner variables are fairly unique as well. The correlations found between the language tasks were all positive, whereas both positive and negative correlations were found in the psychometric tests.

The strongest correlations in the language tasks were found between the sessions of the same language task, between the first and the last session of the Word Recognition task and the Phoneme Discrimination task. Grammaticality Judgment only had two sessions, but had a fairly strong correlation between the first and the last session.
In the psychometric tests the strongest positive correlations were found between NEO-A and NEO-E, and ISA-x and ISA-y respectiely, suggesting a relationship between the personality factors agreeableness and extraversion, and the two axis of experiential preference. The strongest negative correlation found was between Llama B and Llama D, reflecting a relationship between vocabulary learning and sound recognition.

What about the correlations between the language tasks and the individual learner variables?
Of the eight variables we included in our analysis (Raven, Digit Span, LNS, Flanker Inhibition Cost, Llama F, Llama D, motivation and learning condition, six of them turned out to have predictive value on Word Recognition and Grammaticality Judgment. None of the variables had any predictive effect on Phoneme Discrimination.
Table 7.14 gives an overview of the predictors and their effect on the various sessions of the language tasks.

Table 7.14 Overview of the predictors and their predictive effect on the language tasks per session

Predictors	Predictive effect on the language tasks per session: positive (+), positive significant (++) negative (-) no effect (0)							
	Word Recognition			**Grammaticality Judgment**		**Phoneme Discrimination**		
	WR1	WR2	WR3	GJ1	GJ2	PD1	PD2	PD2
Flanker (Inhibition Cost)	0	0	0	+	0	0	0	0
Llama D (Phonological Recognition)	++	0	++	+	++	0	0	0
Llama F (Grammatical Inferencing)	++	+	++	++	++	0	0	0
MB-Condition (Meaning Based)	0	++	0	0	0	0	0	0
Motivation	+	++	+	++	++	0	0	0
Raven (Non-verbal Intelligence)	0	0	0	++	++	0	0	0

We already speculated about the reasons for the absence of any significant predictors for Phoneme Discrimination and observed that whereas most participants were unable to hear new phonetic distinctions, some participants seemed to have a natural ability to do this, a skill that apparently was not measured by any of the psychometric tests employed by VILLA. We saw that the best prediction results (R2 values) were found for Grammaticality Judgment.The predictive values of the individual learner variables on the Word Recognition task were lower, possibly due to the fact that higher cognitive skills could also lead to a high score on the Word Recognition task at the very beginning (T0). The learner variables Llama D, Llama F and motivation were the most consistent predictors, predicting language learning success for many of the sessions of Word Recognition and Grammaticality Judgment. The one occurrence of condition seems a coincidence. We found some coincidental significances as well in chapter 4 to 6. They were country- and test- specific and not consistent.

8. Conclusion

It has been a long and interesting journey to finally get to the point where we can present the results of this study. We first had to make an inventory of the collected learner variables and the administered language tasks and we would not have been able to do this without the help of the VILLA researchers who were simultaneously working on finishing the VILLA Field Manual (2022). When we finally had collected and documented the necessary data we could proceed with the main aim of our study: analyzing the data in search of answers for our research questions.

In this thesis we used the VILLA data to study the effect of individual learner variables on language acquisition in its initial stages and drew up three research questions for this purpose which we will try to answer in this chapter.

1) Are there any differences between the countries in the way learner variables are represented?
2) Does the distinction between meaning- and form-based learning conditions play a role in predicting language learning success?
3) Can individual learner differences predict language learning success?

In chapter 2, we analyzed the differences between the countries. The VILLA researchers had carefully designed a set of selection criteria to ensure a homogenous sample of language learners for their experiment and had collected background data (sex, type of study, background languages), data on cognition subdivided in executive function (Digit Span, Letter Number Sequencing, Flanker) and Perceptual Preference (Barsch), Language Aptitude (Llama B (*Word Learning sills),* Llama D (*Phonological Recogntion*) and Llama F (*Grammatical Inferencing*)), personality (NEO FFI-3 and ISALEM-97) and motivation (*adapted version of AMTB*). Despite this careful selection, some significant differences were found between the countries in the way the learner variables were represented, signifying that the sample was less homogenous than the researcher had hoped. The countries differed in the variables age, type of study, Barsch, Llama B, Llama D, NEO and motivation.

To answer research questions 2 and 3, we had to investigate which language tasks' outcomes were available to define language learning success. Chapter 3 gives an overview of the language tasks administered. In selecting language tasks for our subset, we looked at tasks that had been administered at a minimum of two time intervals, tasks that covered different language domains, and tasks that did not have a large number of missing values. We ended up with Word Recognition, Grammaticality Judgment and Phoneme Discrimination.

In chapters 4 to 6 we analyzed the results on the three selected language tasks to see if there were any differences between the countries in the way they performed over time and whether the learning condition played any role.
Looking at the results of the Word Recognition task in the meaning-based group, we observed that something must have gone wrong in the UK, during the administration of the third session, as all of the data for this session was missing. In the subset we made for the meaning-based group including Germany, significant progress was made by Germany, Italy and the Netherlands between the first (T0) and the second session (T5) and between the first and the last session (T9). None of the countries made any significant progress between the second and the last session. Interestingly, the French participants had a significantly better default score at the first session (T0) compared to their German and British counterparts, which is odd given the fact the French did not show any significant progress throughout the sessions.
In the form-focused group with all countries except Germany, significant progress was made by France and the UK between the first and the last session, after 13,5 hours of Polish input. We observed no significant effect of learning condition. The main overall effect on the Grammaticality Judgment task was that all countries made significant progress. In the meaning-based group, including Germany, both the Dutch and the Italian participants scored significantly higher than the British respondents. No significant effects were found on learning condition. In the last task, Phoneme Discrimination, we found that there were no significant effects at all. None over time, none between the countries and none concerning the learning condition.

We conclude that countries overall, do make progress on Word Recognition and Grammaticality Judgment, so learning takes place, but that, remarkably, the results on the three tasks show that there were no systematic differences between the countries in terms of learning condition. A reason for this could be that due to the fact that VILLA had opted to employ just one teacher, the differences between the two conditions might have been insufficiently operationalized for it to have an effect on highly educated learners, skilled at learning. Interestingly, the meaning-based learning condition did seem to play a significant role in predicting language learning success on the second session of the Word Recognition task. In sum, we have to conclude that the learning condition does not seem to play a significant role in predicting language learning success and that we have to consider the one occurrence where the variable did seem to play a significant role to be a coincidence. However, we do point out that we only analyzed the effects of this variable on three language tasks, and we expect further analyses done on the entire range of language tasks to provide more conclusive results.

Looking at the selection of variables that were included in our regression analysis we found that only Llama D, Llama F, Raven, motivation, and the meaning-based learning condition played a significant role in predicting language success for two of the three language tasks we studied in this thesis. None of these variables had any influence on predicting the results on the Phoneme Discrimination task. Predictors for the Word Recognition task were Llama D, Llama F, the MB-learning condition and motivation. Interestingly their predictive value seemed to depend on the testing session. In the first testing session at T0, prior to any Polish instruction, participants with great phonological recognition skills (Llama D) and or grammatical inferencing abilities (Llama F) also had higher scores on the Word Recognition task. For the second testing session, the meaning-based learning condition and motivation correlated with results on the Word Recognition task. In the last session, predictors of success on the Word Recognition task were Llama D and Llama F once again, hinting that these learner variables seem to lie at the heart of successful Word Recognition skills and that a meaning-based learning condition and motivation might temporarily boost this process. Predictors for the Grammaticality Judgment task were motivation, Llama F, Raven and Llama D with motivation and Llama F being consistent factors for both testing sessions. Raven, a measure of non-verbal intelligence was the third significant predictor in the first session and the last in second session, where Llama D suddenly seemed to play a role in predicting language learning success on this task. This latter phenomenon could perhaps be explained by the fact that after more hours of exposure to the language, students start to become more aware and thus rely more on morphological aspects of the language such as suffixes for grammatical gender or case. Returning to our third research question if individual difference can predict language learning success, the answer is a qualified “yes”! There were several individual learner variables that turned out to be significant in predicting language learning success on the Word Recognition task and the Grammaticality Judgment task.

Returning to the differences in learner variables found between the countries, can any of them be linked to performance on the three language tasks? Looking at the variables that resulted in a significant difference, Llama D and motivation were the only ones that were also included in subset used in the regression analysis and they did turn out to be significant predictors for the language learning success on two of our language tasks: Word Recognition (in the first and the last session) and Grammaticality Judgment (in the second session). Llama D, assessing Phonological Recognition, yielded a significant effect between Italy and France, where the Italian participants were significantly more skilled than the French were, so we would expect to find significant differences between the Italians and the French in the way they performed on these tasks, but our results do not confirm this. Motivation yielded a significant effect between the French and the Dutch participants, where the French turned out to be more motivated to learn the Polish language than the Dutch were. The variable was a significant predictor for the second session of the Word Recognition task and the second session of the Grammaticality Judgment task. Looking at the scores on both tasks we would expect to find significant differences in performance between the French and the Dutch respondents, but unfortunately, our results do not confirm this. It seems that although learner variables can have predictive value for language learning success, they are not solely responsible for this success.

9. Discussion

In this chapter we will first address the flaws in the way the data was collected and/or processed, the limitations of the instruments used and the limitations of the VILLA learner sample. This section is then followed by a paragraph in which we will discuss the huge potential of this data. We will conclude with some recommendations for future research.

Flaws in data collection and or processing

For unknown reasons there were many missing values in the data collected. This phenomenon can be seen across all the methods employed to retrieve the data: in the background data, in the psychometric tests and in the language tests. It is not clear if this is the result of failing computer systems, which could be an explanation for the fact that the data for an entire session of the Word Recognition task went missing in the UK, or because participants were given the opportunity to not complete the tests they had started, or possibly due to human error as many tests were scored manually. Another interesting observation is the fact that, again for unknown reasons, the LLAMA B test was not administered in France and the UK, resulting in 72 missing values on that test. Unfortunately, we had to exclude this test in our subset for the regression analysis. Heather Hilton, one of the researchers involved in the VILLA project noted in her description of the psychometric tests in the VILLA Field Manual (2022) that in some countries there was no supervision whilst the participants were taking the Talkbank tests (Flanker, Digit Span and Letter Number Sequencing), designed by Brian McWhinney, making the results on those tests highly questionable. An overview of how and in which countries the tests and tasks had been administered would certainly shed some light on this uncertainty but could not be obtained at the time of writing this thesis.

Limitations of the instruments used

The VILLA Field Manual (2022) also contains reflections on the fact that the VILLA team had employed instruments that had not yet been standardized. That applies to the entire Llama Suite and the adapted version of Gardner's aptitude and motivation battery).

Limitations of the VILLA learner sample

The VILLA sample consisted of university students only, which could have resulted in a restriction of range effect by which language learning processes and learner variables could not be studied to their full extent. It also raises some questions about the ecological validity of the results on the motivation questionnaire (VILLA Field Manual 2022). Motivation of VILLA learners that participated on a voluntary basis and were renumerated, would have been generally higher than the motivation found with students in an institutional setting where the study of a foreign language is imposed.

The VILLA potential

Despite its limitations, we were able to find some interesting results and we feel that further research is needed to exploit the full potential of the VILLA data to investigate the effects of individual learner differences. In this study we opted to analyze the predictive effect of a selection of learner variables on a subset of language tasks, excluding many learner variables and many language tasks for which a proper analysis is still needed. It would be interesting to see for example if the participants' background languages exert any effect on performance on any of the language tasks. Considering our finding that the Dutch participants scored significantly higher on any of personality traits assessed by the NEO-FFI-3 test, it would be worth to explore if this variable has any predictive value on the results of the language tasks. One of the things we found in this study was that the variable learning condition did not have any effect on the outcomes on the language tasks, however, this conclusion seems a bit premature since we haven't studied the effect of this variable on the other language tasks included in the VILLA experiment. Future research might come to more conclusive results.

Future research opportunities

Researchers interested in the field of individual learner differences could use the VILLA data to focus on the internal linguistic variables of the language tasks which we did not include in our study. Furthermore, future research could look at the interactions between the predictors. The VILLA Field Manual (2022) lists a couple of interesting research opportunities for those interested in the effects of individual learner variables on language learning success. One of them is to perform multivariate analyses “to help uncover more complex interactions between factors and behaviors and outcomes in the complex VILLA dataset”. However, this idea comes with a note of warning as this would add to the complexity of the analyses that need to be done and the question arises if the sample is large enough to handle this type of complexity. The VILLA Field Manual (2022) also suggests carrying out cluster analyses “to look for possible shared profiles among the learners and related learning behaviors or outcomes”. Another idea worth investigating is to look at those participants that scored exceptionally high (outliers) on either learner variables or language tasks and study their progress throughout the entire video- and audio recorded course in a qualitative, case study fashion, in an attempt to find interactions between learner profiles, behavior in response to the two learning conditions and the outcomes of the language tasks.

References

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

Barsch, J. (1991). *Barsch Learning Style Inventory*. Novato, CA: Academic Therapy.

Best, C. T. (1993). Emergence of Language-Specific Constraints in Perception of Non-Native Speech: A window on early phonological development. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289-304). Dordrecht: Springer Netherlands.

Best, C. T. (1994). Learning to perceive the sound pattern of English. In C. RoveeCollier & L. P. Lipsitt (Eds.), *Advances in infancy research* (Vol. 9, pp. 217- 304). Norwood, NJ: Ablex.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in crosslanguage research* (pp. 171-204). Baltimore: York Press.

Bohlmann, N. L., Maier, M. F., & Palacios, N. (2015). Bidirectionality in self-regulation and expressive vocabulary: comparisons between monolingual and dual language learners in preschool. *Child Development*, 86(4), 1094–1111.

Brief descriptions of the most commonly used measures/testing procedures. (z.d.). NCBI Geraadpleegd op 18 juni 2021, van https://www.ncbi.nlm.nih.gov/books/NBK285344/

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences*, 29(6), 1057–1068.

Cahay, R., Honorez, M., Monfort, B., Remy, F., & Therer, J. (1997). *ISALEM 97*. *Université de Liège.*

Cameron, L. 2002. “Measuring vocabulary size in English as an additional language”. *Language Teaching Research* 6 (2): 145–173.

Carlson, S.M., Davis, A.C., & Leach, J.G. (2005). Less is more: Executive function and symbolic representation in preschool children. *Psychological Science*, 16, 609–616

Cassidy, S., & Eachus, P. (2000). Learning style, academic belief systems, self-report student proficiency and academic achievement in higher education. *Educational Psychology*, 20, 307–322.

Chamorro-Premuzic, T., Furnham, A., & Lewis, M. (2007). Personality and approaches to learning predict preference for different teaching methods. *Learning and Individual Differences*, 17(3), 241–250. https://doi-org.ru.idm.oclc.org/10.1016/j.lindif.2006.12.001

Cenoz, J., Hufeisen, B., and Jessner, U. (2001). *Cross-Linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives.* Clevedon: Multilingual Matters.

Costa, A., Hernández M., Costa-Faidella, J., & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: now you see it, now you don't. *Cognition*,113(2), 135–149. https://doi.org/10.1016/j.cognition.2009.08.001

De Angelis, G. & Selinker, L. (2001). Interlanguage transfer and competing linguistic systems in the multilingual mind. In J. Cenoz, B. Hufeisen & U. Jessner (eds.), *Cross-linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives* (pp. 42-58). Clevedon: Multilingual Matters.

De Angelis, G. (2005). Interlanguage transfer of function words. *Language Learning*, 55(3), 379–414.

De Angelis, G. (2007). *Third or Additional Language Acquisition*. Clevedon: Multilingual Matters.

Dewaele, J-M. (1998). Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics*, 19, 471-490.

Dewaele, J-M. (2001). Activation or inhibition? The interaction of L1, L2 and L3 on the language mode continuum. In J. Cenoz, B. Hufeisen & U. Jessner (eds.), *Cross-linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives* (pp. 69-89). Clevedon: Multilingual Matters.

Diamond, A. (2013). Executive functions. Annual Review of Psychology, 64, 135– 168. https://doi.org/10.1146/annurev-psych-113011-143750

Dimroth, C. 2012. "Videoclips zur Elizitation von Erzahlungen: Methodische Uberlegungen und einige Ergebnisse am Beispiel der 'Finite Story'" In *Einblicke in die Zweitspracherwerbsforschung und ihre methodischen Verfahren*, Bernt Ahrenholz (ed), 77‑98. Berlin: de Gruyter.

Dimroth, C., Rast, R., Starren, M., & Watorek, M. (2013). Methods for studying the acquisition of a new language under controlled input conditions: the VILLA project. *Eurosla Yearbook, 13*(1), 109–138. https://doi-org.ru.idm.oclc.org/10.1075/eurosla.13.07dim

Dörnyei & Skehan in: Doughty, C. J., & Long, M. H. (2003). *The handbook of second language acquisition* (Ser. Blackwell handbooks in linguistics, 14). Blackwell Publishing.

Dörnyei, Z. (2005). The psychology of the language learner: Individual differences in second language acquisition. Mahwah NJ: Lawrence Erlbaum.

Dufva M., Niemi P., & Voeten M. J. M. (2001). The role of phonological memory, word recognition, and comprehension skills in reading development: From preschool to grade 2. *Reading and Writing*, 14, 91–117. [Google Scholar]

Dunn, R. (1983). Learning style and its relationship to exceptionality at both ends of the continuum. *Exceptional Children*, 49, 496–506.

Dunn, R. (1993). Learning styles of the multiculturally diverse. *Emergency Librarian*, 20(4), 24–32.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: correlates of language learning success. *The Modern Language Journal*, 79(1), 67–89.

Ehrman, M. E., Leaver, B. L., & Oxford, R. L. (2003). A brief overview of individual differences in second language learning. *System,* 31(3), 313–330.

Eriksen, B.A., Eriksen, C.W. (1974) Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* 16, 143–149. https://doi.org/10.3758/BF03203267

Eriksen, C. W. (1995). The flankers task and response competition: a useful tool for investigating a variety of cognitive problems. Visual Cognition, 2(2-3), 101–118. https://doi.org/10.1080/13506289508401726

Eyckmans, J. 2004. *Measuring receptive vocabulary size.* Utrecht: LOT

Falk, Y., & Bardel, C. (2010). The study of the role of the background languages in third language acquisition. the state of the art. *International Review of Applied Linguistics in Language Teaching (Iral)*, 48(2-3), 185–219.

Faust, M. E., Balota, D. A., Spieler, D. H. and Ferraro, F. R. 1999. "Individual differences in information-processing rate and amount: Implications for group differences in response latency". *Psychological Bulletin* 125 (6): 777–799.

Foster-Cohen, S. H., Medved Krajnovic, M., Mihaljević Djigunović Jelena, & Harrington, M. (2006). The lexical decision task as a measure of l2 lexical proficiency: volume 6 (2006). *Eurosla Yearbook*, 6, 147–168. https://doi-org.ru.idm.oclc.org/10.1075/eurosla.6.10har

Friend, M., & Bates, R. P. (2014). The union of narrative and executive function: Different but complementary. *Frontiers in Psychology, 5,* Article 469.

Fukkink, R. G., Hulstijn, J. and Simis, A. 2005. "Does training in second-language word recognition affect reading comprehension? An experimental study". *Modern Language Journal* 89 (1): 54–75.

Furnham, A. (1992). Personality and learning style: a study of three instruments. *Personality and Individual Differences, 13(4), 429–438.* https://doi-org.ru.idm.oclc.org/10.1016/0191-8869(92)90071-V

Furnham, A., Jackson, C. J. & Miller, T. (1999). Personality, learning style and work performance. Personality and Individual Differences, 27(6), 1113-1122.

Fernandez-Castillo, A., Gutiérrez-Rojas, M.E. (2009). Selective attention, anxiety, depressive symptomatology and academic performance in adolescents. *Electronic Journal of Research in Educational Psychology*, 7, 49-76.

Gathercole, S.E., Alloway, T.P., Kirkwood, H. J, Elliot, J.G., Holmes, J., & Hilton, K.A. (2008). Attentional and executive function behaviours in children with poor working memory. *Learning and Individual Differences*, 18, 214-223.

Gsanger, K., Wa, S., Homack, S., Siekierski, B., & Riccio, C. (2002). The relation of memory and attention to academic achievement in children. *Archives of Clinical Neuropsychology*, 17, 790.

Godfroid, A., Loewen, S., Jung, S., Park, J.-H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: evidence from eye-movement patterns. *Studies in Second Language Acquisition*,37(2), 269–297.

Gooch, D., Hulme, C., Nash, H.M., & Snowling, M.J. (2013). Comorbidities in preschool children at family risk of dyslexia. *Journal of Child Psychology and Psychiatry*, 55, 237–246.

Gooch, D., Thompson, P., Nash, H. M., Snowling, M. J., & Hulme, C. (2016). The development of executive function and language skills in the early school years. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 57(2), 180–187.

Goriot, C. M. M. (2019). *Early-English education works no miracles: Cognitive and linguistic development in mainstream, early-English, and bilingual primary-school pupils in the Netherlands* (Doctoral dissertation, [Sl: sn]).

Granena, G. (2011). Cognitive aptitudes for L2 learning and the LLAMA aptitude test: What aptitude does the LLAMA test measure? *Presentation at the European Second Language Acquisition (EUROSLA) Conference, (Stockholm)*

Green DM, Swets JA. 1966. *Signal detection theory and psychophysics*: Wiley New York.

Grundy, J. G., Chung-Fat-Yim, A., Friesen, D. C., Mak, L., & Bialystok, E. (2017). Sequential congruency effects reveal differences in disengagement of attention for monolingual and bilingual young adults. *Cognition*, 163, 42–55. https://doi.org/10.1016/j.cognition.2017.02.010

Haatveit, B. C., Sundet, K., Hugdahl, K., Ueland, T., Melle, I., & Andreassen, O. A. (2010). The validity of d prime as a working memory index: results from the "bergen n-back task". *Journal of Clinical and Experimental Neuropsychology*, 32(8), 871–80. https://doi-org.ru.idm.oclc.org/10.1080/13803391003596421

Hammarberg, B. (2001). Roles of L1 and L2 in L3 production and acquisition. In J. Cenoz, B. Hufeisen & U. Jessner (eds.), *Cross-linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives* (pp. 21-41). Clevedon: Multilingual Matters.

Harley, T. 2001. *The Psychology of Language. From Data to Theory*. Hove: Psychology Press.

Harrison, G., Andrews, J., & Saklofske, D. (2003). Current perspectives on cognitive learning styles. Education Canada, 43(2), 44–47.

Ibrahimoglu, N., Unaldi, I., Samancioglu, M., & Baglibel, M. (2013). The relationship between personality traits and learning styles: a cluster analysis. *Asian journal of management sciences and education* 2013 2 (3), 93–108. http://ajmse.leena-luna.co.jp/AJMSEPDFs/Vol.2(3)/AJMSE2013(2.3-10).pdf

Ionin, T., & Zyzik, E. (2014). Judgment and interpretation tasks in second language research. *Annual Review of Applied Linguistics*, 34, 37–64. https://doi-org.ru.idm.oclc.org/10.1017/S0267190514000026

Jackson, C., & Lawty-Jones, M. (1996). Explaining the overlap between personality and learning style. *Personality and Individual Differences*, 20(3), 293–300. https://doi-org.ru.idm.oclc.org/10.1016/0191-8869(95)00174-3

Kail, M., Lemaire, P., & Lecacheur, M. (2012). Online grammaticality judgments in french young and older adults. *Experimental Aging Research*, 38(2), 186–207. https://doi-org.ru.idm.oclc.org/10.1080/0361073X.2012.660031

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131, 66–71.

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass and L. Selinker (eds.), *Language Transfer in Language Learning* (pp. 112-34). Rowley, MA: Newbury House.

Klein, W. & Perdue, C. 1997. “The Basic Variety. Or: Couldn’t Natural Languages be much Simpler?” *Second Language Research* 13: 301‐347.

Kolb, D. (1976). *Learning Style Inventory.* Boston: McBet.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development.* Englewood Cliffs: Prentice Hall.

Lalonde, R. N., & Gardner, R. C. (1984). Investigating a causal model of second language acquisition: where does personality fit? *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, 16(3), 224–237.

Lambert, W. E., Gardner, R. C., Barik, H. C., & Tunstal, K. (1963). Attitudinal and cognitive aspects of intensive study of a second language. *Journal of Abnormal and Social Psychology, 66*(4), 358-368.

Lamers, M. J. M.& Roelofs, A (June 2011). "Attentional control adjustments in Eriksen and Stroop task performance can be independent of response conflict". *The Quarterly Journal of Experimental Psychology.* 64 (6): 1056–1081. doi:10.1080/17470218.2010.523792. PMID 21113864. S2CID 1762898

Laufer, B. and Nation, I.S.P. 2001. “Passive vocabulary size and speed of meaning recognition: Are they related?”. In *EU ROSLA Yearbook*, Volume 1, S. Foster-Cohen and A. Nizegorodcew (eds), 7–28. Amsterdam: John Benjamins.

Loewen, S., Grammaticality Judgment Tests and the Measurement of Implicit and Explicit L2 Knowledge.
In: Ellis, R., Loewen, S., Elder, R., Erlam, R., Philp, J., Reinders, H.(Eds.) (2009),
Implicit and explicit knowledge in second language learning, testing and teaching.
` *Multilingual Matters*, Tonawanda, NY, pp. 94-112

Luce, R. D. 1986. *Response Times*. New York: Oxford University Press.

Macintyre, P, & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology*, 15(1), 3–26. https://doi-org.ru.idm.oclc.org/10.1177/0261927X960151001

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21-57). Hillsdale, NJ: Erlbaum

Martin R. C. (2005). Components of short-term memory and their relation to language processing: Evidence from neuropsychology and neuroimaging. *Current Directions in Psychological Science*, 14, 204–208. [Google Scholar]

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*,52(1), 81–90

Mc Dermott, J.M., Pérez-Edgar, K. & Fox, N. A. (2007). Variations of the flanker paradigm: assessing selective attention in young children. *Behavior Research Methods*, 39(1), 62–70. https://doi.org/10.3758/BF03192844

Meara, P. 1996. "The dimensions of lexical competence". In *Performance and Competence in Second Language Acquisition*, G. Brown, K. Malmkjaer and J. Williams (eds), 35–53. Cambridge: Cambridge University Press.

Meara, P. M. & Milton, J. L. 2002, *X_Lex: The Swansea Vocabulary Levels Test.* Newbury: Express.

Meara, P. M. (2005). *Llama language aptitude tests*. Swansea: Lognostics.

Meara, P. and Buxton, B. 1987. "An alternative multiple choice vocabulary tests." *Language Testing* 4 (2): 142–145.

Mielicki, M. K., Koppel, R. H., Valencia, G., & Wiley, J. (2018). Measuring working memory capacity with the letter-number sequencing task: advantages of visual administration. *Applied Cognitive Psychology*, 32(6), 805–814. https://doi.org/10.1002/acp.3468

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology*, 41(1), 49–100.

Mochida, A. and Harrington, M. 2006. "The Yes–No test as a measure of receptive vocabulary knowledge". *Language Testing* 26 (1) 73–98.

Müller, U., Jacques, S., Brocki, K., & Zelazo, P.D. (2009). The executive functions of language in preschool children. In A. Winsler, C. Fernyhough & I. Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-regulation.* New York: Cambridge University Press

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development, 38*(1-2).

Nelson, K. (1981). Individual differences in language development: Implications for development and language. *Developmental Psychology, 17,* 170-187.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. (2005). Working memory and intelligence – Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131, 61–65.

Paradis, J. (2011). Individual differences in child second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism, 1*(3), 213-237.

Raven, J. (2000). The raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*, 41(1), 1–48.

Riccio, C.A., Lee, D., Romine, C. Cash, D. & Davis, B. (2002). Relation of memory and attention to academic achievement in adults. *Archives of Clinical Neuropsychology*, 18, 755-756.

Redick, T. S., Broadway, J. M., Kuriakose, P. S., Engle, R. W., Meier, M. E., Kane, M. J., & Unsworth, N. (2012). Measuring working memory capacity with automated complex span tasks*. European Journal of Psychological Assessment*, 28(3), 164–171. https://doi.org/10.1027/1015-5759/a000123

Ringbom, H. (1986). Crosslinguistic influence and the foreign language learning process. In El. Kellerman & Shardwood Smith (eds.), *Crosslinguistic Influence in Second Language Acquisition* (pp. 150-62). New York: Pergamon Press.

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

Schepens, J. J., der Slik, F., & Hout, R. (2016). L1 and l2 distance effects in learning l3 dutch. *Language Learning*, 66(1), 224–256. https://doi-org.ru.idm.oclc.org/10.1111/lang.12150

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A. Snelling, P. and Stevenson, M. 2003. "First and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge". *Language Learning*, 53 (2): 165–202.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable digit span: a systematic review and cross-validation study. *Assessment*, 19(1), 21–30. https://doi-org.ru.idm.oclc.org/10.1177/107319111142876

Segalowitz, N. and Frenkiel-Fishman, N. 2005. "Attention control and ability in complex cognitive skill: Attention-shifting and second language proficiency". *Memory and Cognition* 10 (3):

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

Shelton, J. T., Elliott, E. M., Hill, B. D., Calamia, M. R., & Gouvier, W. D. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence*, 37, 283–293.

Shillaw, J. 1996. "*The application of Rasch modelling to Yes/No vocabulary tests*". Vocabulary Acquisition Research Group, University of Wales Swansea. Retrieved November 5, 2004 from the World Wide Web: http://www.swan.ac.uk/cals/calsres/vlibrary/js96a.htm

Shiu, L.-J., Spada, N., & Yalcin. (2018). Exploring second language learners' grammaticality judgment performance in relation to task design features. *System*, 72, 215–225. https://doi-org.ru.idm.oclc.org/10.1016/j.system.2017.12.004

Shoemaker, E., & Rast, R. (2013). Extracting words from the speech stream at first exposure. *Second Language Research*, 29(2), 165–183.

Segalowitz, N. & Hulstijn, J. 2005. "Automaticity in bilingualism and second language learning". In *Handbook of Bilingualism: Psycholinguistics Approaches.* F. F. Kroll and A.M.B. De Groot, (eds), 179–201. Oxford: Oxford University Press.

Slabakova, R. (2017). The scalpel model of third language acquisition. *International Journal of Bilingualism*, 21(6), 651–665. https://doi-org.ru.idm.oclc.org/10.1177/1367006916655413

Slik, F., Schepens, J., Bongaerts, T., & Hout, R. (2021). Critical period claim revisited: reanalysis of hartshorne, tenenbaum, and pinker (2018) suggests steady decline and learner-type differences. Language Learning, (20210907). https://doi-org.ru.idm.oclc.org/10.1111/lang.12470

Speciale, G., Ellis, N., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied PsychoLinguistics*, 25,293–321.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. https://doi-org.ru.idm.oclc.org/10.3758/BF03207704

Sternberg, S. 1998. "Inferring mental operations from reaction time data: How we compare objects". In *An Invitation to Cognitive Science*, Volume 4: Methods, Models, and Conceptual Issues, D. Scarborough and S. Sternberg (eds), 436–440. Cambridge, MA: MIT Press.

Swets, J., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, 68, 301–340.

Therer, J. (1998). Styles d'enseignement, styles d'apprentisage et pédagogie différenciée en sciences. *Informations Pédagogiques*, 1998(40), 1–22. https://reseauconceptuel.umontreal.ca/rid=1150300837751_494810271_14247/Styles%20

Tonér S., Kallioinen, P., & Lacerda, F. (2021). Selective auditory attention associated with language skills but not with executive functions in swedish preschoolers. *Frontiers in Psychology*, 12, 664501–664501. https://doi.org/10.3389/fpsyg.2021.664501

Tremblay, M. (2006). Cross-linguistic influence in third language acquisition: The role of L2 proficiency and L2 exposure. *CLO/OPL*, 34, 109-119.

Tyler, M.,2019. PAM-L2 and phonological category acquisition in the foreign language classroom. In *A Sound Approach to Language Matters—In Honor of Ocke-Schwen Bohn*. Edited by Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen and Mette Hjortshøj Sørensen. Denmark: Department of English, School of Com munication & Culture, Aarhus University, pp. 607–30.

Unsworth, N., & Engle, R.W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & Cognition*, 33, 213-220.

van der Schuit M., Segers E., van Balkom H., & Verhoeven L., (2011). How cognitive factors affect language development in children with intellectual disabilities. *Research in Developmental Disabilities*, 32, 1884–1894.

van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn J., Simis, A., Snelling, P. and Stevenson, M. 2004. “Linguistic knowledge, processing speed and metacognitive knowledge in first and second language reading comprehension: A componential analysis”. *Journal of Educational Psychology* 96 (1): 19–30.

Vermeiren, A., & Cleeremans, A. (2012). The validity of d' measures. Plos One, 7(2), 31595. https://doi-org.ru.idm.oclc.org/10.1371/journal.pone.0031595

VILLA Field Manual (2022), team villa researchers. Wissenschaftliche Schriften der WWU Münster, to be published.

von Wittich, D., & Antonakis, J. (2011). The kai cognitive style inventory: was it personality all along? *Personality and Individual Differences*, 50(7), 1044–1049. https://doi-org.ru.idm.oclc.org/10.1016/j.paid.2011.01.022

Williams, S & Hammarberg, B. (1998). Language switches in L3 production: Implications for a polyglot speaking model. *Applied Linguistics* 19, 295 – 333.

Zapalska, A. M., & Dabb, H. (2002). Learning styles. *Journal of Teaching in International Business*, 13, 77–97.

Zhang, L.-fang. (2003). Does the big five predict learning approaches? Personality and Individual Differences, 34(8), 1431–1446. https://doi-org.ru.idm.oclc.org/10.1016/S0191-8869(02)00125-3

Appendices

Barsch Learning Styles Inventory

This inventory examines and reports your preferences in terms of the three primary senses you use to take in information: visual, auditory, and tactile/kinesthetic (touch and movement). Check the appropriate line for each.

often sometimes seldom

____ ____ ____ 1. Can remember more about a subject through listening than reading,

____ ____ ____ 2. Follow written directions better than oral directions.

____ ____ ____ 3. Like to write things down or take notes for visual review

____ ____ ____ 4. Bear down extremely hard with a pen or pencil when writing.

____ ____ ____ 5. Require explanations of diagrams, graphs or visual directions.

____ ____ ____ 6. Enjoy working with tools.

____ ____ ____ 7. Are skillful with and enjoy developing and making graphs and charts.

____ ____ ____ 8. Can tell if sounds match when presented with pairs of sounds.

____ ____ ____ 9. Remember best when I write things down several times.

____ ____ ____ 10. Can understand and follow directions on maps.

____ ____ ____ 11. Do better at academic subjects by listening to lectures and tapes.

____ ____ ____ 12. Play with coins or keys in pocket.

____ ____ ____ 13. Learn to spell better by repeating the letters aloud than by writing the the word on paper.

____ ____ ____ 14. Can better understand a news article by reading about it in the paper than by listening to it on the radio.

____ ____ ____ 15. Chew gum, smoke, or snack during studies.

____ ____ ____ 16. Feel the best way to remember is to picture it in your head.

____ ____ ____ 17. Learning to spell by “finger spelling” the words.

____ ____ ____ 18. Would rather listen to a good lecture or speech than read about the same material in a book.

____ ____ ____ 19. Are good at solving and working on jigsaw puzzles and mazes.

____ ____ ____ 20. Prefer to be shown rather than told.

____ ____ ____ 21. Prefer listening to the news on the radio rather than reading about it in the paper.

____ ____ ____ 22. Obtain information on an interesting subject by reading relevant materials.

____ ____ ____ 23. Feel very comfortable touching others, hugging, handshaking, etc.

____ ____ ____ 24. Follow oral directions better than written ones.

Scoring

OFTEN = 5 points SOMETIMES = 3 points SELDOM = 1 point

Place a point value on the line next to its corresponding item number. Next, add the points to obtain the preference scores under each heading.

Visual Preference		Auditory Preference		Tactile/Kinesthetic	
No.	Points	No.	Points	No.	Points
2.	_____	1.	_____	4.	_____
3.	_____	5.	_____	6.	_____
7.	_____	8.	_____	9.	_____
10.	_____	11.	_____	12.	_____
14.	_____	13.	_____	15.	_____
16.	_____	18.	_____	17.	_____
20.	_____	21.	_____	19.	_____
22.	_____	24.	_____	23.	_____
Total Visual	_____	Total Auditory	_____	Total Tactile/ Kinesthetic	_____

If the scores in each modality (i.e. visual, auditory, tactile) are within a few points of each other, you probably use all your modes equally. On the other hand, the inventory suggests you have a preference if one score is significantly higher than the others.

Profiles

Visual Learners…

Like to see words in writing or have concepts presented pictorially. They remember what they *see*. They are attuned to physical elements in a classroom. They like illustrations, diagrams, charts, etc. Visual Learners benefit from overhead transparencies, handouts, charts, diagrams and board work. They take lots of notes and are able to recall information by reviewing them.

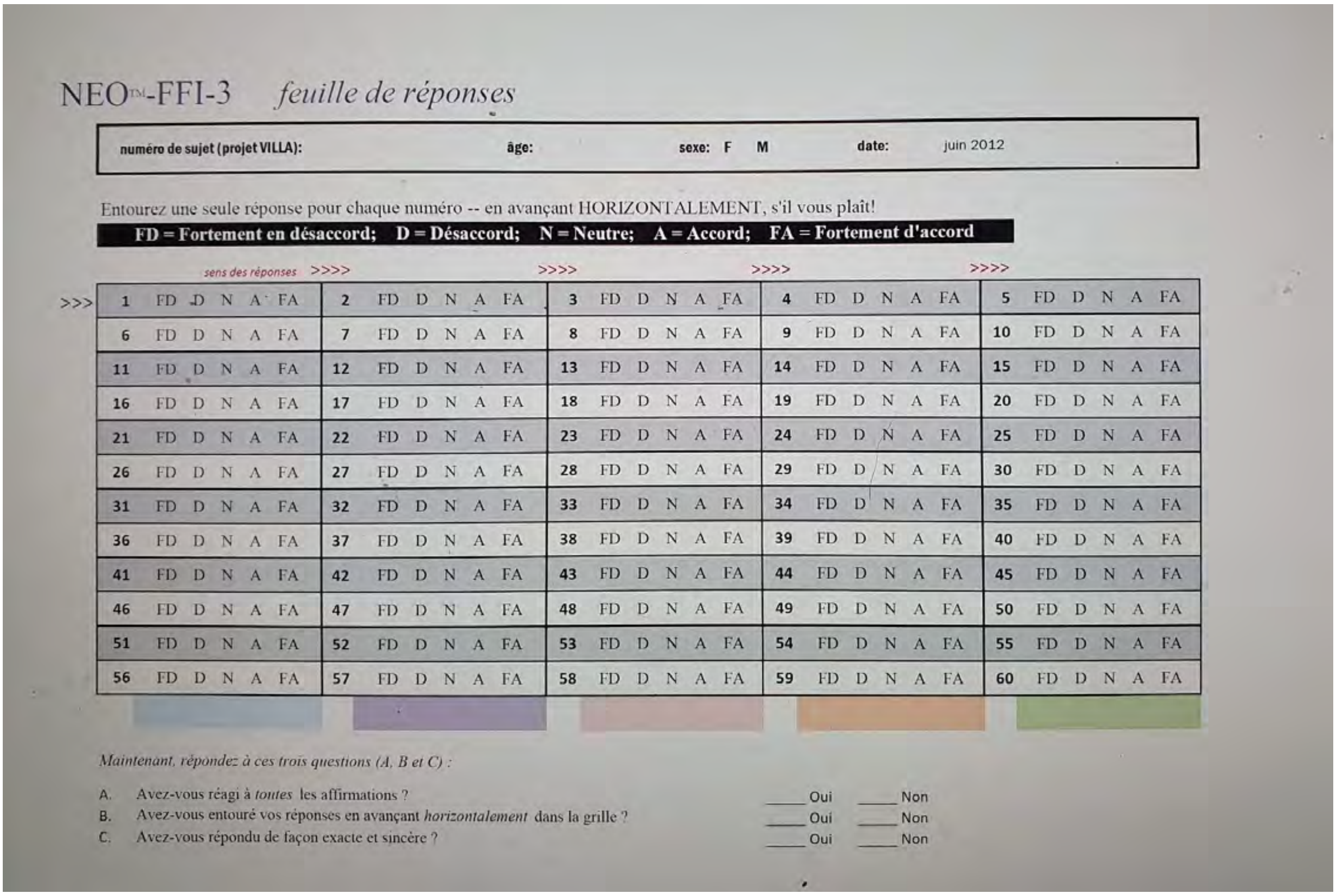
Auditory Learners…

Use their voices and ears as the primary modes for learning. They remember what they *hear*. They express themselves verbally. They understand things by talking them through. Auditory learners love class discussion and are not as likely to take notes. They often “vocalize” what they read. Auditory learners often benefit when they can obtain information from audio tapes or lectures.

Tactile/Kinesthetic Learners…

Learn better when they touch and are physically involved in what they study. They want to handle material, make products, do projects, etc. They understand and remember what they *do*. They learn best by trying things out, experimenting and practicing. Tactile/kinesthetic learners benefit from taking notes because it is something they can do in the learning experience, but they may never -- and never need to -- reread them. Fidgeting and doodling may help them think clearly. A tactile/kinesthetic learner does best when subject matter can be applied to real-life situations.

Answering sheet



NEO™-FFI-3 *feuille de réponses*

numéro de sujet (projet VILLA): âge: sexe: F M date: juin 2012

Entourez une seule réponse pour chaque numéro -- en avançant HORIZONTALEMENT, s'il vous plaît!

FD = Fortement en désaccord; D = Désaccord; N = Neutre; A = Accord; FA = Fortement d'accord

sens des réponses >>>> >>>> >>>> >>>>

>>> 1 FD D N A FA	2 FD D N A FA	3 FD D N A FA	4 FD D N A FA	5 FD D N A FA
6 FD D N A FA	7 FD D N A FA	8 FD D N A FA	9 FD D N A FA	10 FD D N A FA
11 FD D N A FA	12 FD D N A FA	13 FD D N A FA	14 FD D N A FA	15 FD D N A FA
16 FD D N A FA	17 FD D N A FA	18 FD D N A FA	19 FD D N A FA	20 FD D N A FA
21 FD D N A FA	22 FD D N A FA	23 FD D N A FA	24 FD D N A FA	25 FD D N A FA
26 FD D N A FA	27 FD D N A FA	28 FD D N A FA	29 FD D N A FA	30 FD D N A FA
31 FD D N A FA	32 FD D N A FA	33 FD D N A FA	34 FD D N A FA	35 FD D N A FA
36 FD D N A FA	37 FD D N A FA	38 FD D N A FA	39 FD D N A FA	40 FD D N A FA
41 FD D N A FA	42 FD D N A FA	43 FD D N A FA	44 FD D N A FA	45 FD D N A FA
46 FD D N A FA	47 FD D N A FA	48 FD D N A FA	49 FD D N A FA	50 FD D N A FA
51 FD D N A FA	52 FD D N A FA	53 FD D N A FA	54 FD D N A FA	55 FD D N A FA
56 FD D N A FA	57 FD D N A FA	58 FD D N A FA	59 FD D N A FA	60 FD D N A FA
🟦	🟪	🟥	🟧	🟩

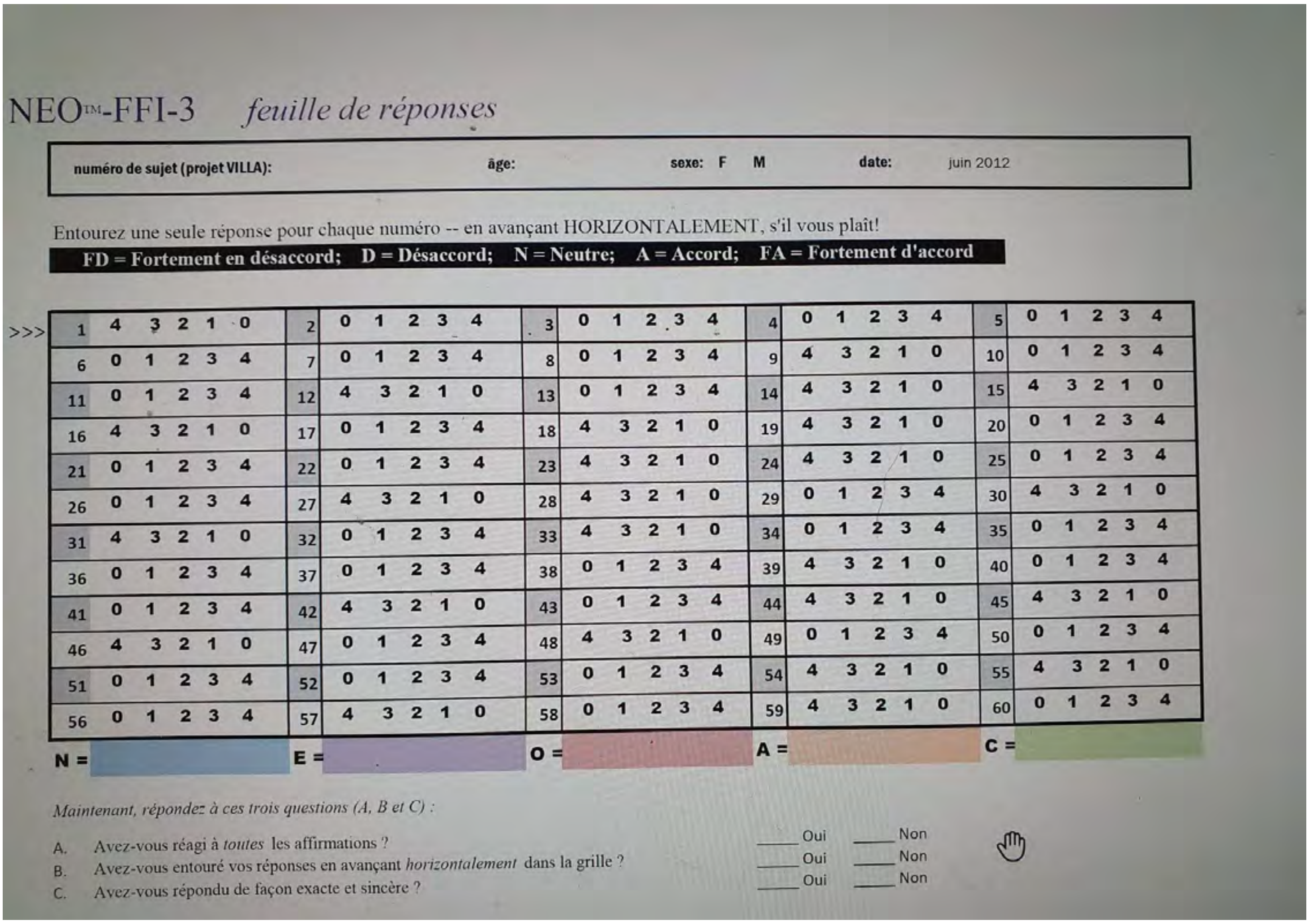
Maintenant, répondez à ces trois questions (A, B et C) :

A. Avez-vous réagi à *toutes* les affirmations ? _____ Oui _____ Non

B. Avez-vous entouré vos réponses en avançant *horizontalement* dans la grille ? _____ Oui _____ Non

C. Avez-vous répondu de façon exacte et sincère ? _____ Oui _____ Non

Scoring key:



NEO™-FFI-3 *feuille de réponses*

numéro de sujet (projet VILLA): âge: sexe: F M date: juin 2012

Entourez une seule réponse pour chaque numéro -- en avançant HORIZONTALEMENT, s'il vous plaît!

FD = Fortement en désaccord; D = Désaccord; N = Neutre; A = Accord; FA = Fortement d'accord

>>> 1	4 3 2 1 0	2	0 1 2 3 4	3	0 1 2 3 4	4	0 1 2 3 4	5	0 1 2 3 4
6	0 1 2 3 4	7	0 1 2 3 4	8	0 1 2 3 4	9	4 3 2 1 0	10	0 1 2 3 4
11	0 1 2 3 4	12	4 3 2 1 0	13	0 1 2 3 4	14	4 3 2 1 0	15	4 3 2 1 0
16	4 3 2 1 0	17	0 1 2 3 4	18	4 3 2 1 0	19	4 3 2 1 0	20	0 1 2 3 4
21	0 1 2 3 4	22	0 1 2 3 4	23	4 3 2 1 0	24	4 3 2 1 0	25	0 1 2 3 4
26	0 1 2 3 4	27	4 3 2 1 0	28	4 3 2 1 0	29	0 1 2 3 4	30	4 3 2 1 0
31	4 3 2 1 0	32	0 1 2 3 4	33	4 3 2 1 0	34	0 1 2 3 4	35	0 1 2 3 4
36	0 1 2 3 4	37	0 1 2 3 4	38	0 1 2 3 4	39	4 3 2 1 0	40	0 1 2 3 4
41	0 1 2 3 4	42	4 3 2 1 0	43	0 1 2 3 4	44	4 3 2 1 0	45	4 3 2 1 0
46	4 3 2 1 0	47	0 1 2 3 4	48	4 3 2 1 0	49	0 1 2 3 4	50	0 1 2 3 4
51	0 1 2 3 4	52	0 1 2 3 4	53	0 1 2 3 4	54	4 3 2 1 0	55	4 3 2 1 0
56	0 1 2 3 4	57	4 3 2 1 0	58	0 1 2 3 4	59	4 3 2 1 0	60	0 1 2 3 4
N =		E =		O =		A =		C =	

Maintenant, répondez à ces trois questions (A, B et C) :

A. Avez-vous réagi à *toutes* les affirmations ? _____ Oui _____ Non

B. Avez-vous entouré vos réponses en avançant *horizontalement* dans la grille ? _____ Oui _____ Non

C. Avez-vous répondu de façon exacte et sincère ? _____ Oui _____ Non

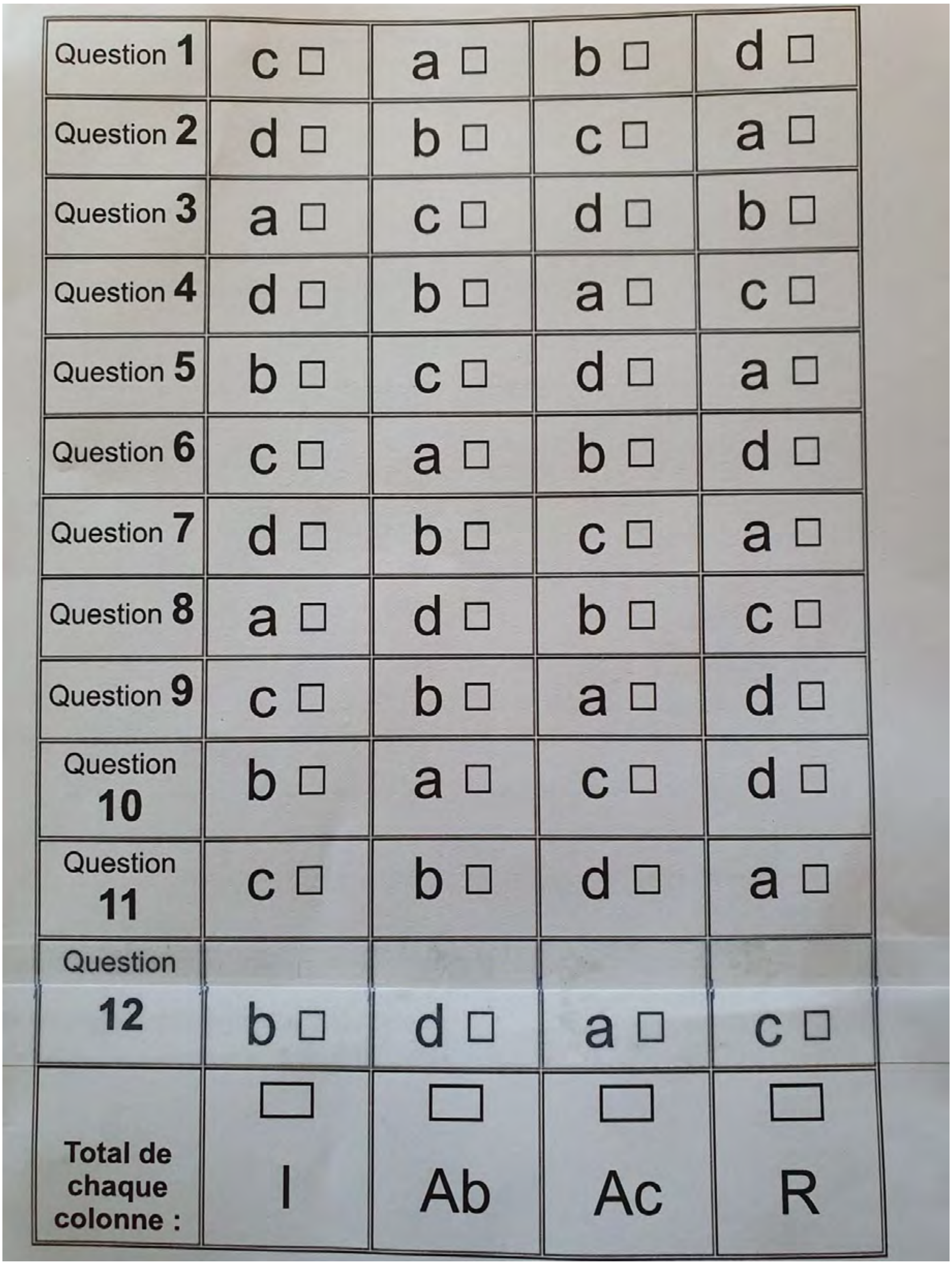
ISALEM-97 Scoring Sheet

Université de Liège

LABORATOIRE D'ENSEIGNEMENT MULTIMEDIA

GRILLE DE DÉCODAGE DU QUESTIONNAIRE ISALEM-97

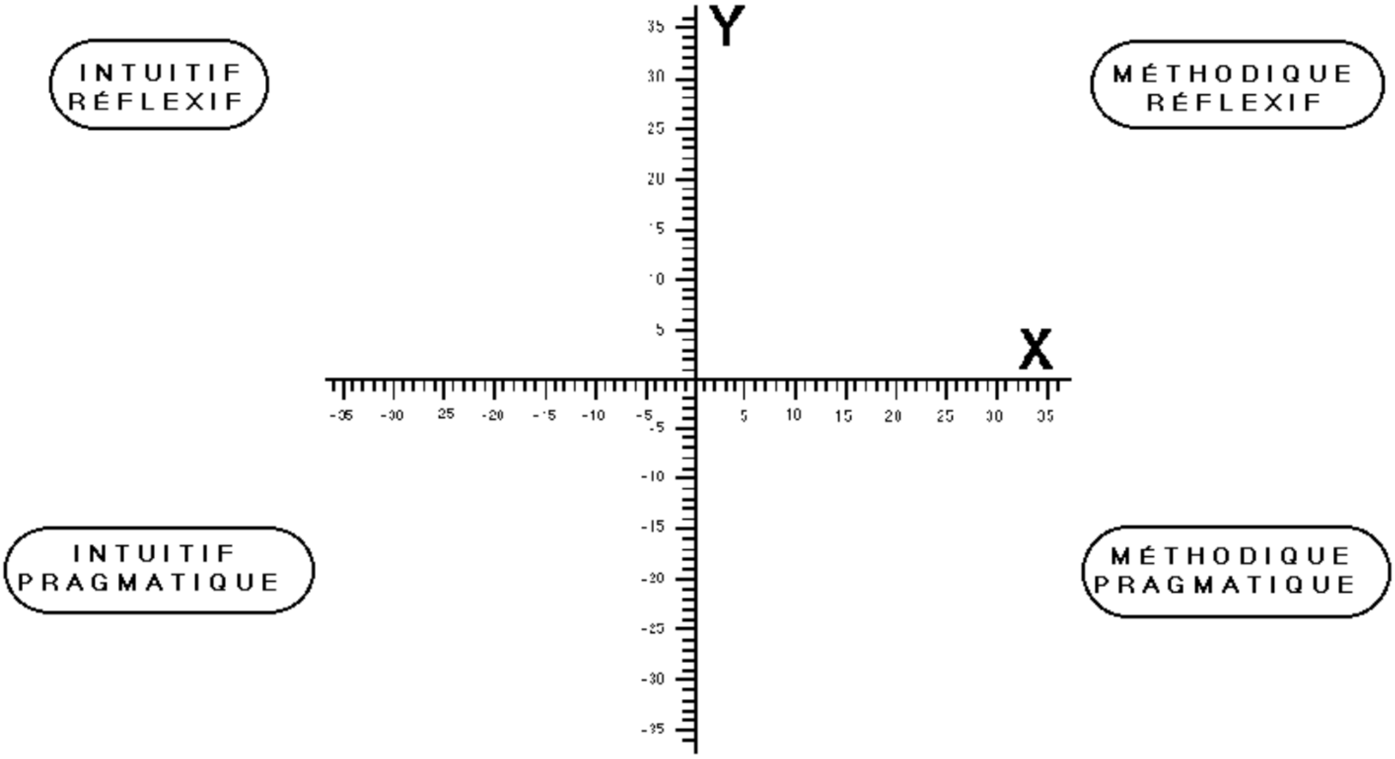
Pour chacune des douze questions, reportez, dans le tableau suivant, le chiffre que vous avez attribué à chacune des quatre propositions.



Question 1	c □	a □	b □	d □
Question 2	d □	b □	c □	a □
Question 3	a □	c □	d □	b □
Question 4	d □	b □	a □	c □
Question 5	b □	c □	d □	a □
Question 6	c □	a □	b □	d □
Question 7	d □	b □	c □	a □
Question 8	a □	d □	b □	c □
Question 9	c □	b □	a □	d □
Question 10	b □	a □	c □	d □
Question 11	c □	b □	d □	a □
Question 12	b □	d □	a □	c □
Total de chaque colonne :	□ I	□ Ab	□ Ac	□ R

Valeurs à reporter sur les axes		
Si vous avez ***moins*** de 18 ans(enseignement secondaire)	**I - Ab - 2 = X**	**Ac - R - 2 = Y**
Si vous avez plus de 18 ans (ens. supérieur et adultes)	**I - Ab - 8 = X**	**Ac - R - 5 = Y**

ISALEM-97 Scoring Grid



ISALEM-97 Learning Profiles

Description des styles d'apprentissage

La découverte de votre style d'apprentissage **préférentiel** (avec vos points forts et vos points faibles) vous aidera à optimiser vos propres apprentissages et à mieux percevoir la diversité et la complémentarité des réactions des autres face à un problème.

Si vous êtes plutôt **intuitif réflexif**	Si vous êtes plutôt **méthodique réflexif**
Vous excellez à considérer une situation sous des angles très variés. Votre réaction initiale est plutôt d'observer que d'agir. Vous appréciez les situations qui nécessitent un foisonnement d'idées comme, par exemple, lors d'un "brainstorming". Vous avez des intérêts culturels très larges et vous aimez rassembler des informations avec éclectisme. **Vos points forts** Vous êtes particulièrement doué pour : • imaginer; • comprendre les gens; • identifier les problèmes. **Vos points faibles** Vous auriez tendance à : • hésiter dans vos choix; • retarder vos décisions.	Vous excellez à synthétiser un vaste registre d'informations de manière logique et concise. Vous vous centrez plus sur l'analyse des idées et des problèmes que sur les personnes comme telles. Vous êtes surtout intéressé par la rigueur et la validité des théories. **Vos points forts** Vous êtes particulièrement doué pour : • planifier; • créer des "modèles scientifiques"; • définir des problèmes; • développer des théories. **Vos points faibles** Vous auriez tendance à : • "construire des châteaux en Espagne"; • méconnaître les applications pratiques d'une théorie.

Si vous êtes plutôt **intuitif pragmatique**	Si vous êtes plutôt **méthodique pragmatique**
Vous aimez apprendre en mettant la "main à la pâte". Vous prenez plaisir à mettre en oeuvre des projets et à vous impliquer personnellement dans de nouvelles expériences que vous percevez comme des défis. Vous réagissez davantage par instinct qu'en fonction d'une analyse purement logique. Lors de la résolution d'un problème, vous aimez vous informer auprès des autres avant de procéder à vos propres investigations. **Vos points forts** Vous êtes particulièrement doué pour : • réaliser des projets; • diriger; • prendre des risques. **Vos points faibles** Vous auriez tendance à : • agir pour agir; • vous disperser.	Vous excellez à mettre en pratique les idées et les théories. Vous êtes capable de résoudre des problèmes et de prendre des décisions sans tergiverser et en sélectionnant la solution optimale. Vous préférez vous occuper de sciences appliquées ou de technologies plutôt que de questions purement sociales ou relationnelles. **Vos points forts** Vous êtes particulièrement doué pour : • définir et résoudre les problèmes; • prendre des décisions; • raisonner par déduction. **Vos points faibles** Vous auriez tendance à : • prendre des décisions précipitées; • vous attaquer à de faux problèmes.

Adapté de D. Kolb, Learning-Style Inventory, Self-scoring inventory and interpretation Booklet, Revised Edition, 1985

- **Le style d'apprentissage dominant, c'est la manière préférentielle de résoudre un problème.**
- Nous utilisons les quatre styles de base, mais 75% des gens ont un style dominant.
- On ne peut pas hiérarchiser les styles d'apprentissage : leur efficacité spécifique varie en fonction des circonstances ...
- Il n'y a donc pas UNE bonne façon d'apprendre ou de résoudre un problème ...En conséquence, nous sommes différents, mais complémentaires.

VILLA Motivation Questionnaire Instructions, Item codes and Scoring Key

Instructions:

Vous verrez ci-après des déclarations, avec lesquelles on peut être ou ne pas être d'accord.
Pour chaque déclaration, veuillez cocher la réaction qui exprime votre degré d'accord ou de désaccord. Voici un exemple:

	PAS DU TOUT D'ACCORD	PAS D'ACCORD	PLUTOT PAS D'ACCORD	PLUTOT D'ACCORD	D'ACCORD	TOUT A FAIT D'ACCORD
Le polonais est une langue qui me plaît						
Le polonais est une langue facile à prononcer						
C'est inutile d'étudier le polonais						

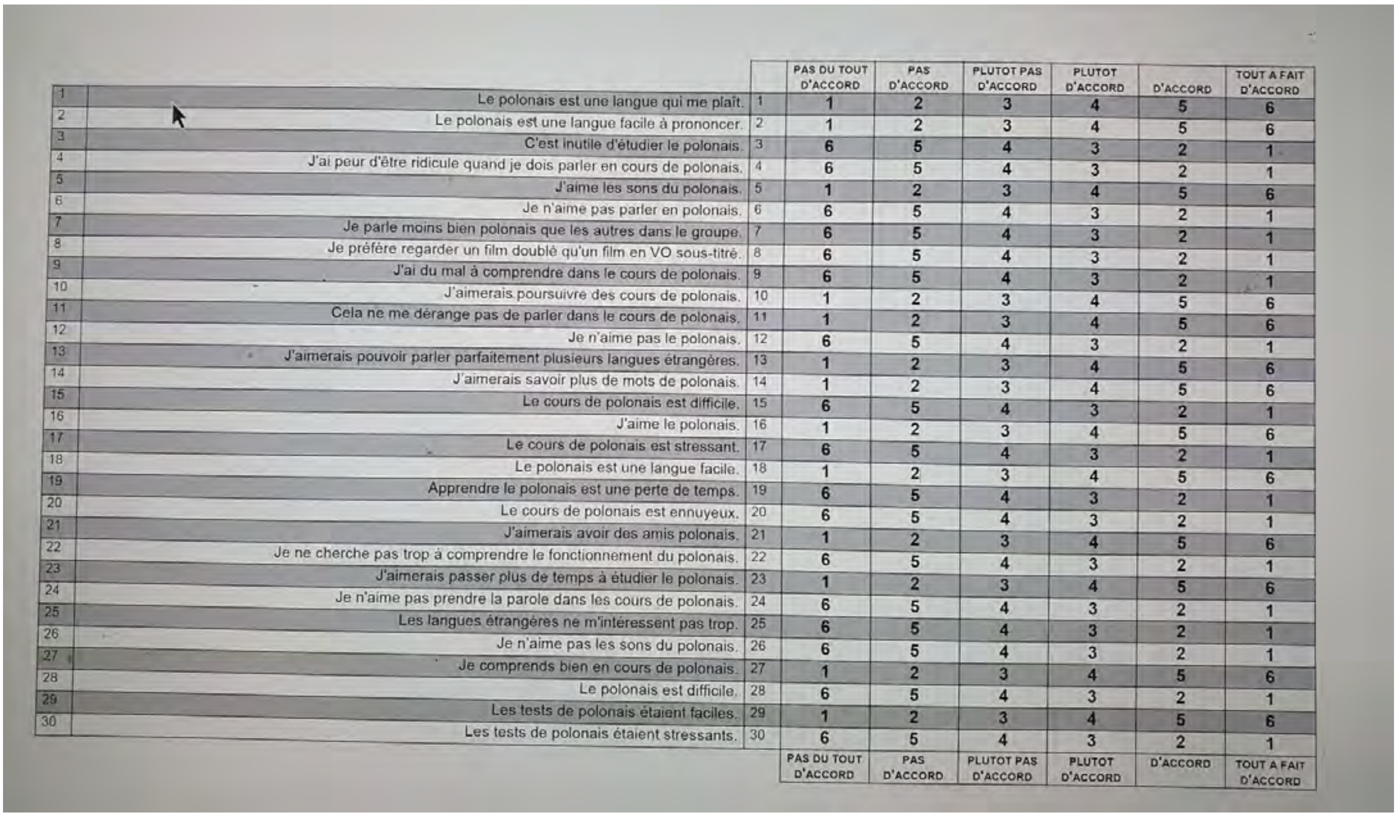
En répondant, vous devez cocher l'une des six réactions possibles. Certains cocheront la case sous ***tout à fait d'accord***, d'autres ***pas du tout d'accord***, d'autres une case entre ces deux extrêmes.
La réponse que vous choisissez doit refléter votre propre opinion. Il n'y pas de « réponse juste » ou de « réponse fausse ».
C'est votre opinion, **en toute honnêteté**, qui nous intéresse.

Item Codes :

		ITEM CODE
1	Le polonais est une langue qui me plaît.	AP
2	Le polonais est une langue facile à prononcer.	AP
3	C'est inutile d'étudier le polonais.	INS
4	J'ai peur d'être ridicule quand je dois parler en cours de polonais.	PCA
5	J’aime les sons du polonais.	AP
6	Je n'aime pas parler en polonais.	PUA
7	Je parle moins bien polonais que les autres dans le groupe.	PCA
8	Je préfère regarder un film doublé qu'un film en VO sous-titré.	IFL
9	J'ai du mal à comprendre dans le cours de polonais.	PUA
10	J’aimerais poursuivre des cours de polonais.	ALP
11	Cela ne me dérange pas de parler dans le cours de polonais.	PUA
12	Je n’aime pas le polonais.	AP
13	J'aimerais pouvoir parler parfaitement plusieurs langues étrangères.	IFL
14	J’aimerais savoir plus de mots de polonais.	ALP
15	Le cours de polonais est difficile.	CE
16	J'aime le polonais.	AP
17	Le cours de polonais est stressant.	PCA
18	Le polonais est une langue facile.	AP
19	Apprendre le polonais est une perte de temps.	ALP
20	Le cours de polonais est ennuyeux.	CE
21	J'aimerais avoir des amis polonais.	INT
22	Je ne cherche pas trop à comprendre le fonctionnement du polonais.	MI
23	J'aimerais passer plus de temps à étudier le polonais.	MI
24	Je n'aime pas prendre la parole dans les cours de polonais.	PCA
25	Les langues étrangères ne m'intéressent pas trop.	IFL
26	Je n’aime pas les sons du polonais.	AP
27	Je comprends bien en cours de polonais.	PUA
28	Le polonais est difficile.	AP
29	Les tests de polonais étaient faciles.	ATP
30	Les tests de polonais étaient stressants.	ATP

ALP	attitude towards learning Polish
AP	attitude towards the Polish language
ATP	attitude towards the Polish tests
CE	class evaluation
IFL	interest for foreign languages
INS	instrumental orientation
INT	integrative orientation
MI	motivational intensity
PCA	Polish class anxiety
PUA	Polish use anxiety

Scoring Key:



			PAS DU TOUT D'ACCORD	PAS D'ACCORD	PLUTOT PAS D'ACCORD	PLUTOT D'ACCORD	D'ACCORD	TOUT A FAIT D'ACCORD
1	Le polonais est une langue qui me plaît.	1	1	2	3	4	5	6
2	Le polonais est une langue facile à prononcer.	2	1	2	3	4	5	6
3	C'est inutile d'étudier le polonais.	3	6	5	4	3	2	1
4	J'ai peur d'être ridicule quand je dois parler en cours de polonais.	4	6	5	4	3	2	1
5	J'aime les sons du polonais.	5	1	2	3	4	5	6
6	Je n'aime pas parler en polonais.	6	6	5	4	3	2	1
7	Je parle moins bien polonais que les autres dans le groupe.	7	6	5	4	3	2	1
8	Je préfère regarder un film doublé qu'un film en VO sous-titré.	8	6	5	4	3	2	1
9	J'ai du mal à comprendre dans le cours de polonais.	9	6	5	4	3	2	1
10	J'aimerais poursuivre des cours de polonais.	10	1	2	3	4	5	6
11	Cela ne me dérange pas de parler dans le cours de polonais.	11	1	2	3	4	5	6
12	Je n'aime pas le polonais.	12	6	5	4	3	2	1
13	J'aimerais pouvoir parler parfaitement plusieurs langues étrangères.	13	1	2	3	4	5	6
14	J'aimerais savoir plus de mots de polonais.	14	1	2	3	4	5	6
15	Le cours de polonais est difficile.	15	6	5	4	3	2	1
16	J'aime le polonais.	16	1	2	3	4	5	6
17	Le cours de polonais est stressant.	17	6	5	4	3	2	1
18	Le polonais est une langue facile.	18	1	2	3	4	5	6
19	Apprendre le polonais est une perte de temps.	19	6	5	4	3	2	1
20	Le cours de polonais est ennuyeux.	20	6	5	4	3	2	1
21	J'aimerais avoir des amis polonais.	21	1	2	3	4	5	6
22	Je ne cherche pas trop à comprendre le fonctionnement du polonais.	22	6	5	4	3	2	1
23	J'aimerais passer plus de temps à étudier le polonais.	23	1	2	3	4	5	6
24	Je n'aime pas prendre la parole dans les cours de polonais.	24	6	5	4	3	2	1
25	Les langues étrangères ne m'intéressent pas trop.	25	6	5	4	3	2	1
26	Je n'aime pas les sons du polonais.	26	6	5	4	3	2	1
27	Je comprends bien en cours de polonais.	27	1	2	3	4	5	6
28	Le polonais est difficile.	28	6	5	4	3	2	1
29	Les tests de polonais étaient faciles.	29	1	2	3	4	5	6
30	Les tests de polonais étaient stressants.	30	6	5	4	3	2	1
			PAS DU TOUT D'ACCORD	PAS D'ACCORD	PLUTOT PAS D'ACCORD	PLUTOT D'ACCORD	D'ACCORD	TOUT A FAIT D'ACCORD

Packages used in R- studio, version 4.1.1

Package	**Description**
Emmeans	The emmeans package enables users to easily obtain least-squares means for many linear, generalized linear, and mixed models as well as compute contrasts or linear functions of least-squares means, and comparisons of slopes
lmer4	lme4 provides functions for fitting and analyzing mixed models: linear (lmer), generalized linear (glmer) and nonlinear (nlmer.)
lmerTest	The lmerTest package provides p-values in type I, II or III anova and summary tables for linear mixed models (lmer model fits cf. lme4) via Satterthwaite's degrees of freedom method.
sjPlot	sjPlot is a Collection of plotting and table output functions for data visualization.
sjstats	Sjstats is a collection of convenient functions for common statistical computations, which are not directly provided by R's base or stats packages.
sjmisc	Sjmisc is a collection of miscellaneous utility functions, supporting data transformation tasks like recoding, dichotomizing or grouping variables, setting and replacing missing values.
ggplot2	ggplot2' is a plotting system based on the grammar of graphics.
GGgally	GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data. Some of these functions include a pairwise plot matrix, a two group pairwise plot matrix, a parallel coordinates plot, a survival plot, and several functions to plot networks.
haven	Haven enables R to read and write various data formats used by other statistical packages. It currently supports: SAS, SPSS and Stata.
Rcpp	The Rcpp package helps to integrate R and C++ via R functions and a (header-only) C++ library.

Transparency	**Frequency**	**Target**	**AudStim (Test sentence)**
HIGH TRANSPARENCY **HT** N=24	LOW FREQUENCY **LF** N=12	kontrola	Konieczna będzie zatem codzienna kontrola szlaków i usuwanie śmieci.
		jogurt	Sąsiadka mi dała zamrożony jogurt z boskimi owocami lasu.
		kultura	Mimo krępujących więzów rodzima kultura rozwijała się wykazując prężność.
		program	Godzinny sobotni popularny program powstał za prezesury szefa.
		kontynent	Przybyliście na nasz w pełni pokojowy kontynent w zamiarach dywersyjnych.
		plastik	Późniejsze samochody to już wyłącznie plastic projektowany przez stylistów.
		lampa	W rogu stała smukła miedziana lampa z jarzeniówką o odcieniu jasnego złota.
		element	Niestety bezrobocie to trwały element współczesnego życia społecznego.
		format	Rada przegłosowała nowy format rozpraw i wydawnictw naukowych.
		grupa	Na rynku pracy rozszerza się grupa zawodowa pojętnych czeladników.
		dokument	Groźne skutki przemocy pokazuje document papieskiej rady do spraw przekazu.
		spectakl	Zmiana naszych wspólnych działań w spektakl polityczny służy partyjnym celom.
	HIGH FREQUENCY **HF** N=12	teatr	We Lwowie radny zorganizował teatr kukiełkowy ludowego twórcy.
		profesor	Moje jak najszczersze słowa professor poczytał mi za okropne bluźnierstwo.
		adres	Każdy może z łatwością wyszukać adres niedrogiego noclegu w okolicy.
		muzyk	Od czasu do czasu dorabia jako muzyk w pobliskiej wiejskiej restauracji.
		studentka	Wyglądała jak dwudziestoletnia studentka przystępująca do egzaminu.
		inżynier	O takim kontrakcie marzył każdy inżynier zaraz po wojnie na Bliskim Wschodzie.
		dialog	Kłótnie zostają zastąpione przez dialog przyjacielski pomiędzy mówcami.
		telefon	Nieznajomy wyjął szybkim ruchem telefon z kieszeni zdziwionego przechodnia.
		francuz	Tegoroczną biesiadę otworzy Francuz zakamuflowany jako hrabia.

		fotograf	Być może utalentowany fotograf wykonałby te zdjęcia znacznie lepiej.
		artyska	Młoda utalentowana artystka przystraja się w gipiurową suknię z trenem.
		norweg	W ostatnim konkursie potężnie zbudowany Norweg zdobył złoty medal.
LOW TRANSPARENCY **LT** N=24	LOW FREQUENCY **LF** N=12	lodówka	Gigantyczna mrożąco chłodząca lodówka należy do tych fundamentów.
		źrebak	Nie zapędzony do stajni na noc źrebak ganiał po błotnistym podwórku.
		wałek	Trzeba było też włożyć specjalny wałek rozrządu z innymi krzywkami.
		kula	Niemal na pewno celna kula udzieliła mu pouczającej odpowiedzi.
		garłacz	Szybkim ruchem wyciągnął garłacz z szuflady i zaczął strzelać dookoła siebie.
		trzepak	Energicznie zarzucił go na odrapany trzepak stojący na trawniku.
		fala	Ale najgorszą konsekwencją jest rosnąca fala ubóstwa i bezdomność.
		pszczoła	Zdaje się poruszać jak pszczoła and kwiatem na górskiej łące między przełęczami.
		czajnik	Masywny przerdzewiały czajnik na wodę szumi leniwie na piecyku.
		kubeł	Lekko uniósł okropnie śmierdzący kubeł wypełniony krowimi plackami.
		naleśnik	Bujał się jak cienki sprasowany naleśnik na zawieszonym hamaku.
		gruzinka	Starsza o bardzo nobliwym wyglądzie gruzinka opowiada swoje dzieje.
	HIGH FREQUENCY **HF** N=12	mieszka	Sama samiusieńka jedna mieszka w ponurej, ogromnej pracowni na strychu.
		imię	Tamtejszy mędrzec nadał mu imię biegnącego z wiatrem szarego wilka.
		listonosz	Rząd zachował się jak swoisty listonosz doręczający Sejmowi project.
		włoszka	Wszystkich zebranych urzeka piękna włoszka swoją nienaganną sylwetką.
		lubisz	Obywatelowi w oczy to ty patrzeć nie lubisz jak każdy zawadiaka.
		strażak	W uzasadnionych okolicznościach strażak kierujący ma prawo zarządzenia.

		język	Przełożył bowiem między innymi na język obcy ziemię obiecaną.
		nazywa	A dalej poczciwy mistrz Maciej nazywa ją pośredniczką zagubionych ziemian.
		lekarz	Wykonujący aktywnie zawód lekarz ma prawo do wystawiania recept.
		niemiec	Sudecki kandydat na kanclerza niemiec wygra najbliższe jesienne wybory.
		dobrze	Jeszcze nie zdążyły się dobrze zabliźnić rany i wyschnąć łzy po wojnie.
		kucharka	Kolejną ofiarą śmiertelną jest kucharka dowództwa marynarki wojennej.

HIGH TRANSPARENCY
HT
N=16
	LOW FREQUENCY
	LF
	N=8
		MALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				EDWARD JEST INFORMATYKIEM
				gj_test prof01.wav[24]
				SEBASTIAN JEST ARCHITEKTEM
				gj_test prof02.wav

			INCORRECT= NOMINATIVE
			N=2
				STEFAN JEST INFORMATYK
				gj_test prof03.wav
				LUDWIG JEST ARCHITEKT
				gj_test prof04.wav

		FEMALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				BARBARA JEST SEKRETARKĄ
				gj_test prof05.wav
				KAROLINA JEST ASYSTENTKĄ
				gj_test prof06.wav

			INCORRECT= NOMINATIVE
			N=2
				EMMA JEST SEKRETARKA
				gj_test prof07.wav
				LIDIA JEST ASYSTENTKA
				gj_test prof08.wav

	HIGH FREQUENCY
	HF
	N=8
		MALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				ADAM JEST INŻYNIEREM
				gj_test prof09.wav
				ALBERT JEST FOTOGRAFEM
				gj_test prof10.wav

			INCORRECT= NOMINATIVE
			N=2
				WiKTOR JESTINŻYNIER
				gj_test prof11.wav
				TOMASZ JEST FOTOGRAF
				gj_test prof12.wav

		FEMALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				HELENA JEST STUDENTKĄ
				gj_test prof13.wav
				IZABELA JEST ARTYSTKĄ
				gj_test prof14.wav

			INCORRECT= NOMINATIVE
			N=2
				KRYSTYNA JEST STUDENTKA
				gj_test prof15.wav
				LIZA JEST ARTYSTKA
				gj_test prof16.wav

LOW TRANSPARENCY
LT
N=16
	LOW FREQUENCY
	LF
	N=8
		MALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				FILIP JEST ROLNIKIEM
				gj_test prof17.wav

[24] Each.wav file name refers to a column labeled ‘Target’ in E-prime, Excel and Text files.

				ROBERT JEST DZIENNIKARZEM gj_test prof18.wav
			INCORRECT= NOMINATIVE N=2	LEONARD JEST ROLNIK gj_test prof19.wav ERYK JEST DZIENNIKARZ gj_test prof20.wav
		FEMALE N=4	CORRECT= INSTRUMENTAL N=2	ANNA JEST KRAWCOWĄ gj_test prof21.wav MARIA JEST PIOSENKARKĄ gj_test prof22.wav
			INCORRECT= NOMINATIVE N=2	EWA JEST KRAWCOWA gj_test prof23.wav KARINA JEST PIOSENKARKA gj_test prof24.wav
	HIGH FREQUENCY **HF** N=8	MALE N= 4	CORRECT= INSTRUMENTAL N=2	PATRYK JEST LEKARZEM gj_test prof25.wav DANIEL JEST STRAżAKIEM gj_test prof26.wav
			INCORRECT= NOMINATIVE N=2	DAWID JEST LEKARZ gj_test prof27.wav STEFAN JEST STRAżAK gj_test prof28.wav
		FEMALE N=4	CORRECT= INSTRUMENTAL N=2	AGATA JEST NAUCZYCIELKĄ gj_test prof29.wav WERONIKA JEST TŁUMACZKĄ gj_test prof30.wav
			INCORRECT= NOMINATIVE N=2	NATALIA JEST NAUCZYCIELKA gj_test prof31.wav MARTA JEST TŁUMACZKA gj_test prof32.wav

HIGH TRANSPARENCY
HT
N=16
	LOW FREQUENCY
	LF
	N=8
		MALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				PATRYK JEST GREKIEM**Fout! Bladwijzer niet gedefinieerd.**
				ANTONI JEST AUSTRALIJCZYKIEM
				gj_testnat02.wav

			INCORRECT= NOMINATIVE
			N=2
				EDMUND JEST GREK
				gj_test nat03.wav
				TOMASZ JEST AUSTRALIJCZYK
				gj_test nat04.wav

		FEMALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				NORA JEST IRANKĄ
				gj_test nat05.wav
				MAGDA JEST WIETNAMKĄ
				gj_test nat06.wav

			INCORRECT= NOMINATIVE
			N=2
				OLGA JEST IRANKA
				gj_test nat07.wav
				SABINA JEST WIETNAMKA
				gj_test nat08.wav

	HIGH FREQUENCY
	HF
	N=8
		MALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				DAWID JEST NORWEGIEM
				gj_test nat09.wav
				DANIEL JEST FRANCUZEM
				gj_test nat10.wav

			INCORRECT= NOMINATIVE
			N=2
				EDWARD JEST NORWEG
				gj_test nat11.wav
				SEBASTIAN JEST FRANCUZ
				gj_test nat12.wav

		FEMALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				IZABELA JESTPORTUGALKĄ
				gj_test nat13.wav
				EMMA JEST BRAZYLIJKĄ
				gj_test nat14.wav

			INCORRECT= NOMINATIVE
			N=2
				MARIA JEST PORTUGALKA
				gj_test nat15.wav
				PATRYCJA JEST BRAZYLIJKA
				gj_test nat16.wav

LOW TRANSPARENCY
LT
N=16
	LOW FREQUENCY
	LF
	N=8
		MALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				LEONARD JEST LITWINEM
				gj_test nat17.wav
				FILIP JEST WALIJCZYKIEM
				gj_test nat18.wav

			INCORRECT= NOMINATIVE
			N=2
				ALBERT JEST LITWIN
				gj_test nat19.wav
				ADAM JEST WALIJCZYK
				gj_test nat20.wav

		FEMALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				JOANNA JEST GRUZINKĄ
				gj_test nat21.wav
				KLARA JEST DUNKĄ
				gj_test nat22.wav

			INCORRECT= NOMINATIVE
			N=2
				TATIANA JEST GRUZINKA
				gj_test nat23.wav
				HELENA JEST DUNKA
				gj_test nat24.wav

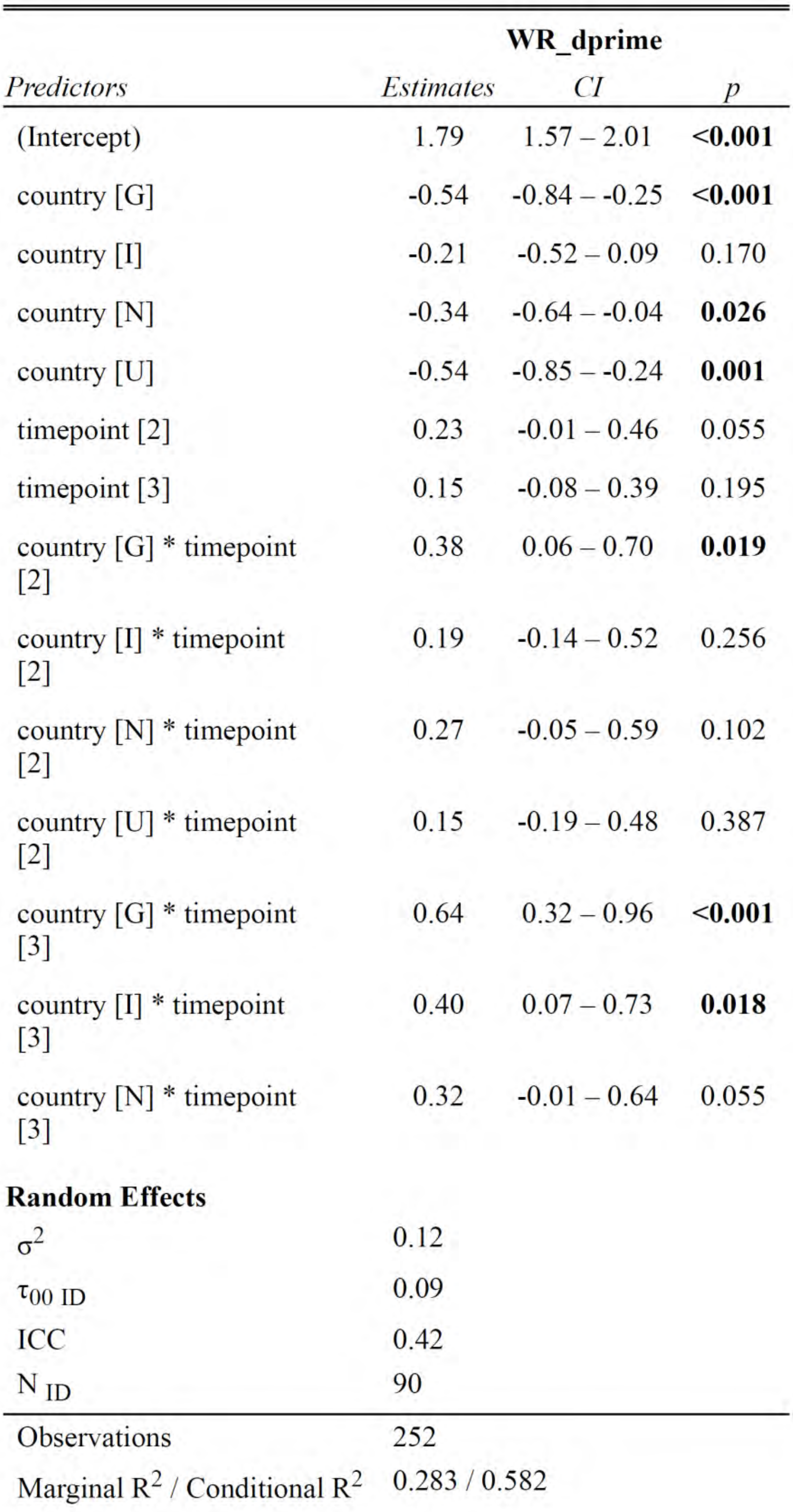
	HIGH FREQUENCY
	HF
	N=8
		MALE
		N= 4
			CORRECT= INSTRUMENTAL
			N=2
				ROBERT JEST CHORWATEM
				gj_test nat25.wav
				WIKTOR JEST CHIŃCZYKIEM
				gj_test nat26.wav

			INCORRECT= NOMINATIVE
			N=2
				LUDWIG JEST CHORWAT
				gj_test nat27.wav
				ERYK JEST CHIŃCZYK
				gj_test nat28.wav

		FEMALE
		N=4
			CORRECT= INSTRUMENTAL
			N=2
				NADIA JEST NIEMKĄ
				gj_test nat29.wav
				ELIZA JEST WŁOSZKĄ
				gj_test nat30.wav

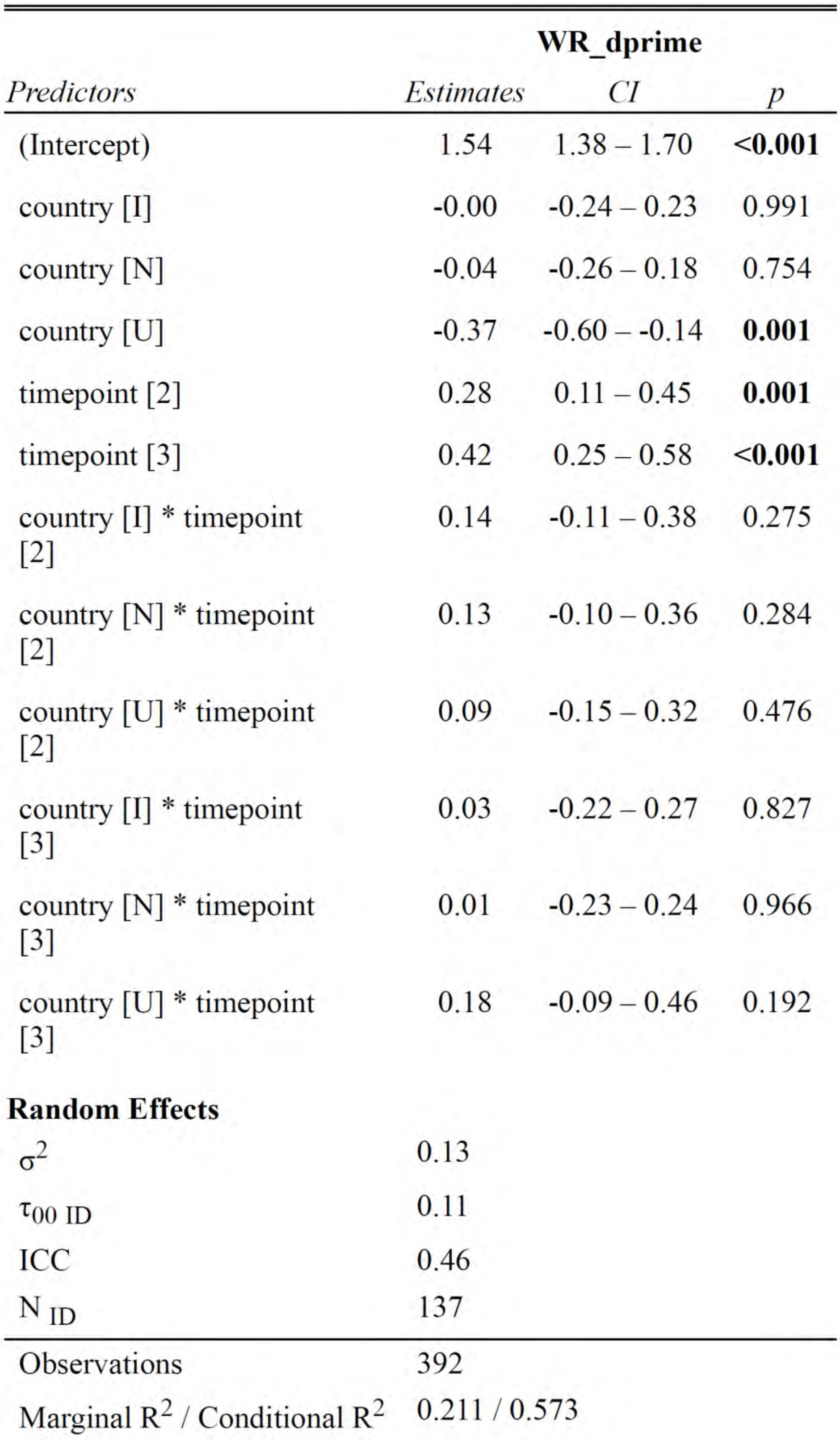
			INCORRECT= NOMINATIVE
			N=2
				LAURA JEST NIEMKA
				gj_test nat31.wav
				SANDRA JEST WŁOSZKA
				gj_test nat32.wav

Table 1 Overview progress over time Word Recognition M- group



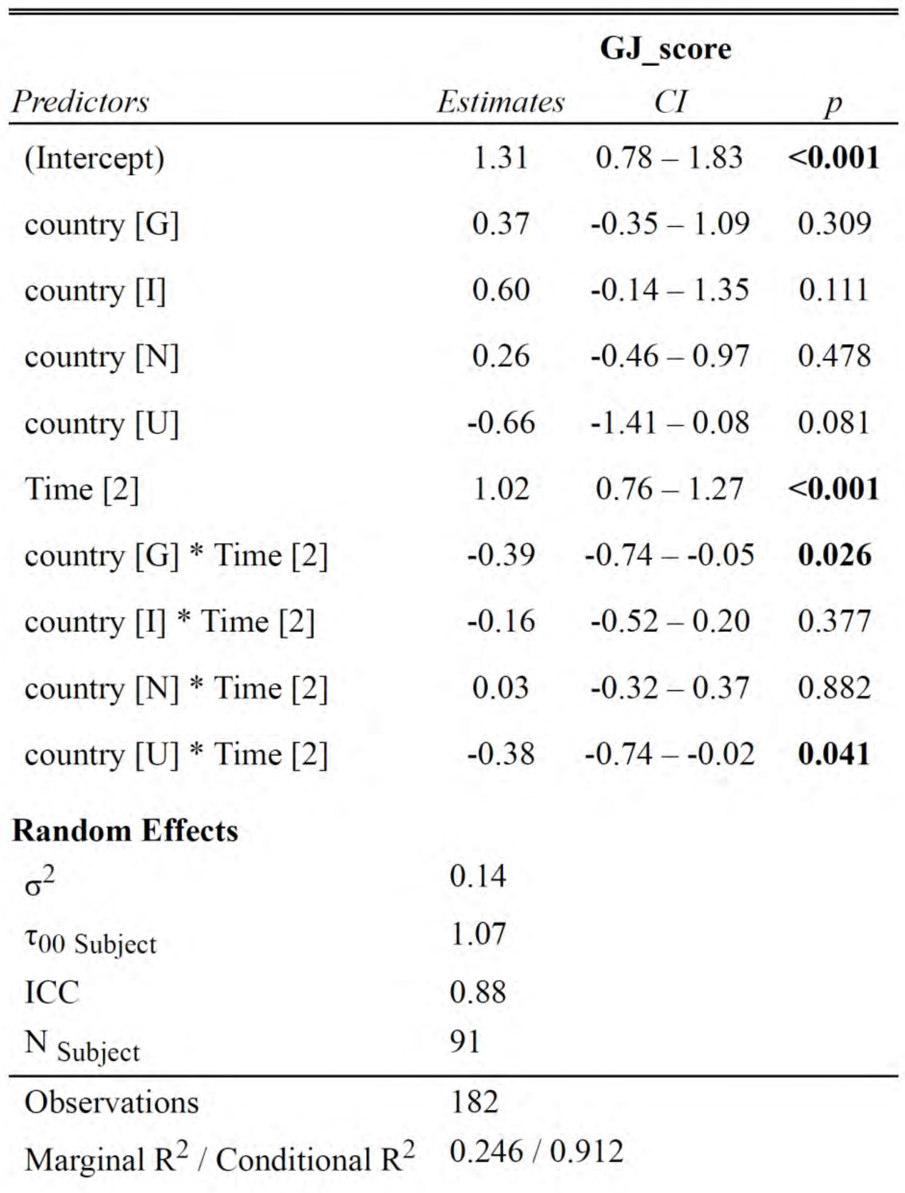
	WR_dprime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.79	1.57 – 2.01	**<0.001**
country [G]	-0.54	-0.84 – -0.25	**<0.001**
country [I]	-0.21	-0.52 – 0.09	0.170
country [N]	-0.34	-0.64 – -0.04	**0.026**
country [U]	-0.54	-0.85 – -0.24	**0.001**
timepoint [2]	0.23	-0.01 – 0.46	0.055
timepoint [3]	0.15	-0.08 – 0.39	0.195
country [G] * timepoint [2]	0.38	0.06 – 0.70	**0.019**
country [I] * timepoint [2]	0.19	-0.14 – 0.52	0.256
country [N] * timepoint [2]	0.27	-0.05 – 0.59	0.102
country [U] * timepoint [2]	0.15	-0.19 – 0.48	0.387
country [G] * timepoint [3]	0.64	0.32 – 0.96	**<0.001**
country [I] * timepoint [3]	0.40	0.07 – 0.73	**0.018**
country [N] * timepoint [3]	0.32	-0.01 – 0.64	0.055
Random Effects			
σ^2	0.12		
$\tau_{00\ ID}$	0.09		
ICC	0.42		
N_{ID}	90		
Observations	252		
Marginal R^2 / Conditional R^2	0.283 / 0.582		

Table 2 Overview progress over time Word Recognition, all conditions



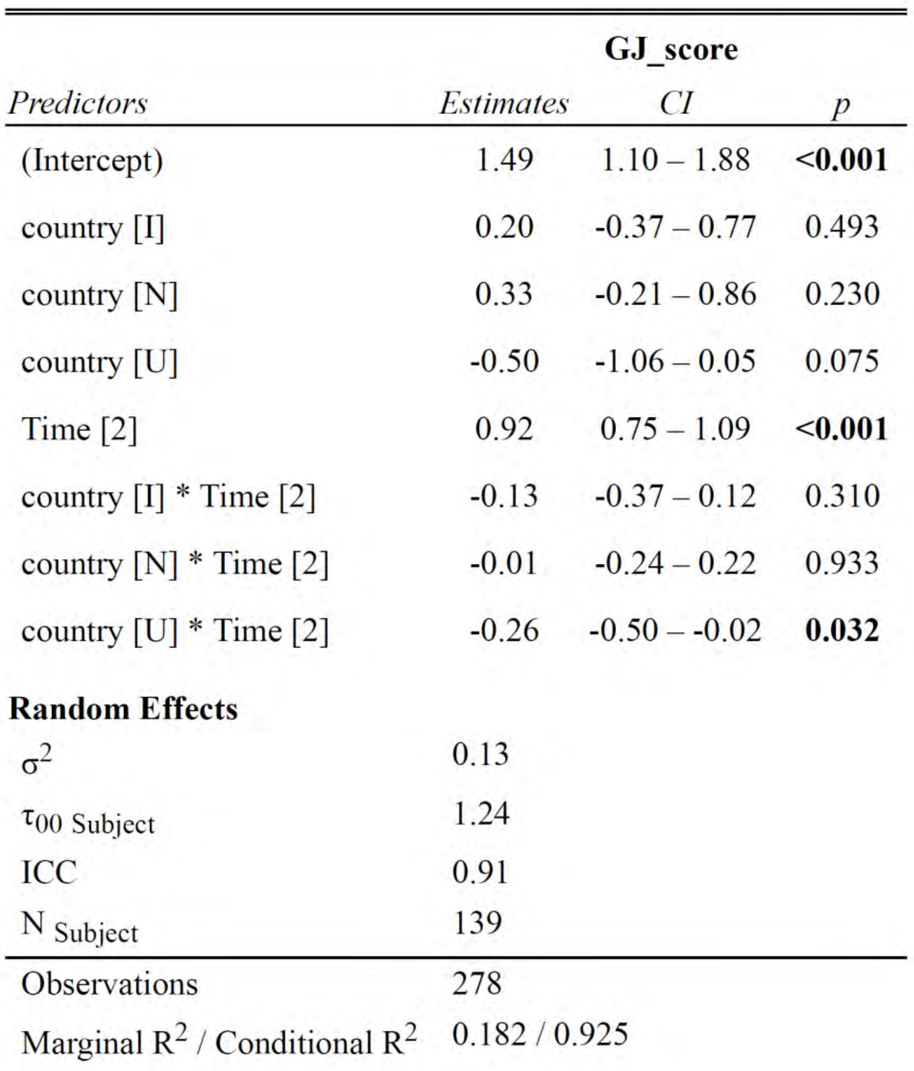
	WR_dprime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.54	1.38 – 1.70	**<0.001**
country [I]	-0.00	-0.24 – 0.23	0.991
country [N]	-0.04	-0.26 – 0.18	0.754
country [U]	-0.37	-0.60 – -0.14	**0.001**
timepoint [2]	0.28	0.11 – 0.45	**0.001**
timepoint [3]	0.42	0.25 – 0.58	**<0.001**
country [I] * timepoint [2]	0.14	-0.11 – 0.38	0.275
country [N] * timepoint [2]	0.13	-0.10 – 0.36	0.284
country [U] * timepoint [2]	0.09	-0.15 – 0.32	0.476
country [I] * timepoint [3]	0.03	-0.22 – 0.27	0.827
country [N] * timepoint [3]	0.01	-0.23 – 0.24	0.966
country [U] * timepoint [3]	0.18	-0.09 – 0.46	0.192
Random Effects			
σ^2	0.13		
$\tau_{00\ \mathrm{ID}}$	0.11		
ICC	0.46		
N_{ID}	137		
Observations	392		
Marginal R^2 / Conditional R^2	0.211 / 0.573		

Table 3 **Overview progress over time Grammaticality Judgment M-group**



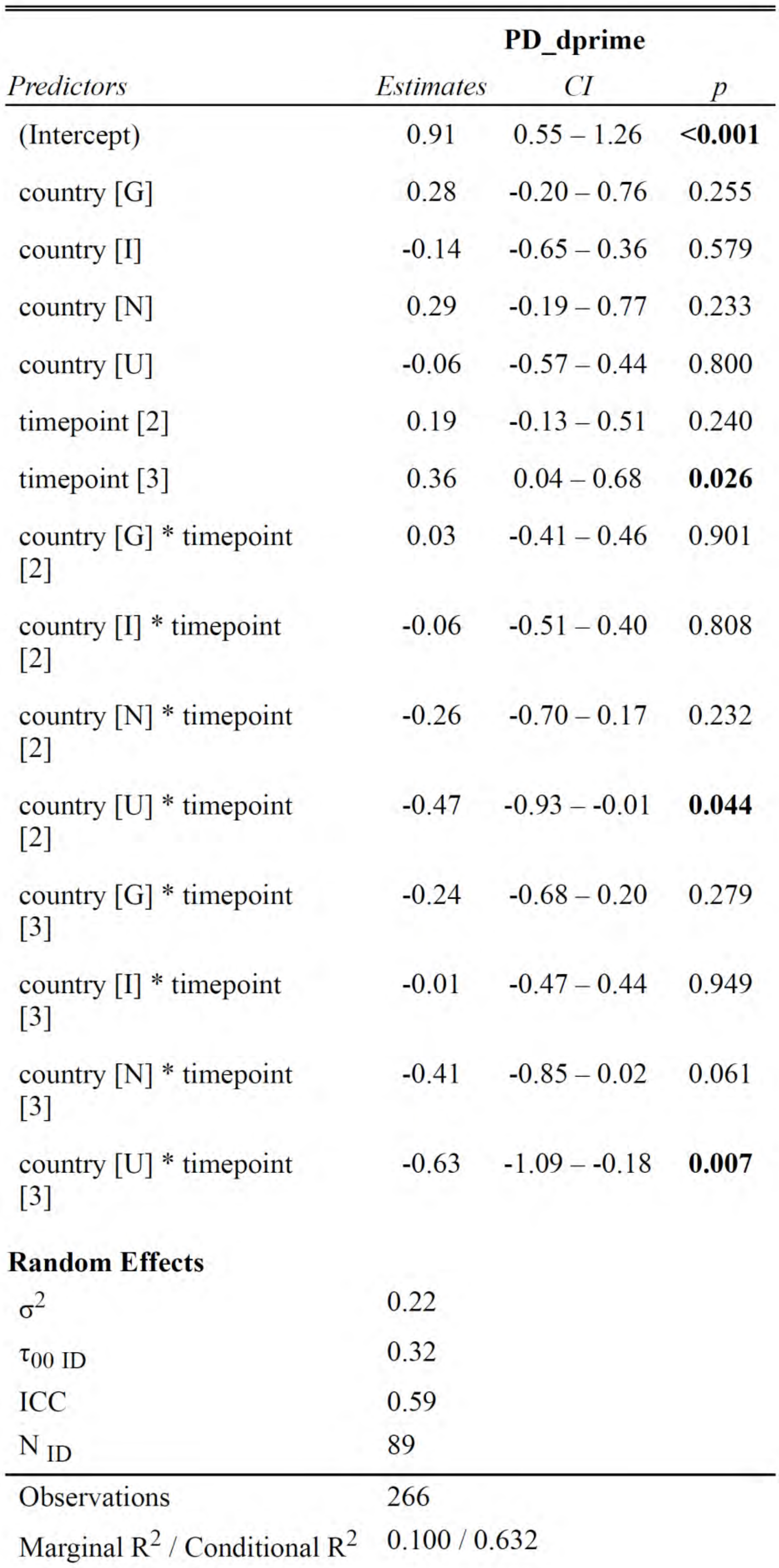
	GJ_score		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.31	0.78 – 1.83	**<0.001**
country [G]	0.37	-0.35 – 1.09	0.309
country [I]	0.60	-0.14 – 1.35	0.111
country [N]	0.26	-0.46 – 0.97	0.478
country [U]	-0.66	-1.41 – 0.08	0.081
Time [2]	1.02	0.76 – 1.27	**<0.001**
country [G] * Time [2]	-0.39	-0.74 – -0.05	**0.026**
country [I] * Time [2]	-0.16	-0.52 – 0.20	0.377
country [N] * Time [2]	0.03	-0.32 – 0.37	0.882
country [U] * Time [2]	-0.38	-0.74 – -0.02	**0.041**
Random Effects			
σ^2	0.14		
$\tau_{00\ \mathrm{Subject}}$	1.07		
ICC	0.88		
N_{Subject}	91		
Observations	182		
Marginal R^2 / Conditional R^2	0.246 / 0.912		

Table 4 Overview progress over time Grammaticality Judgment, all conditions



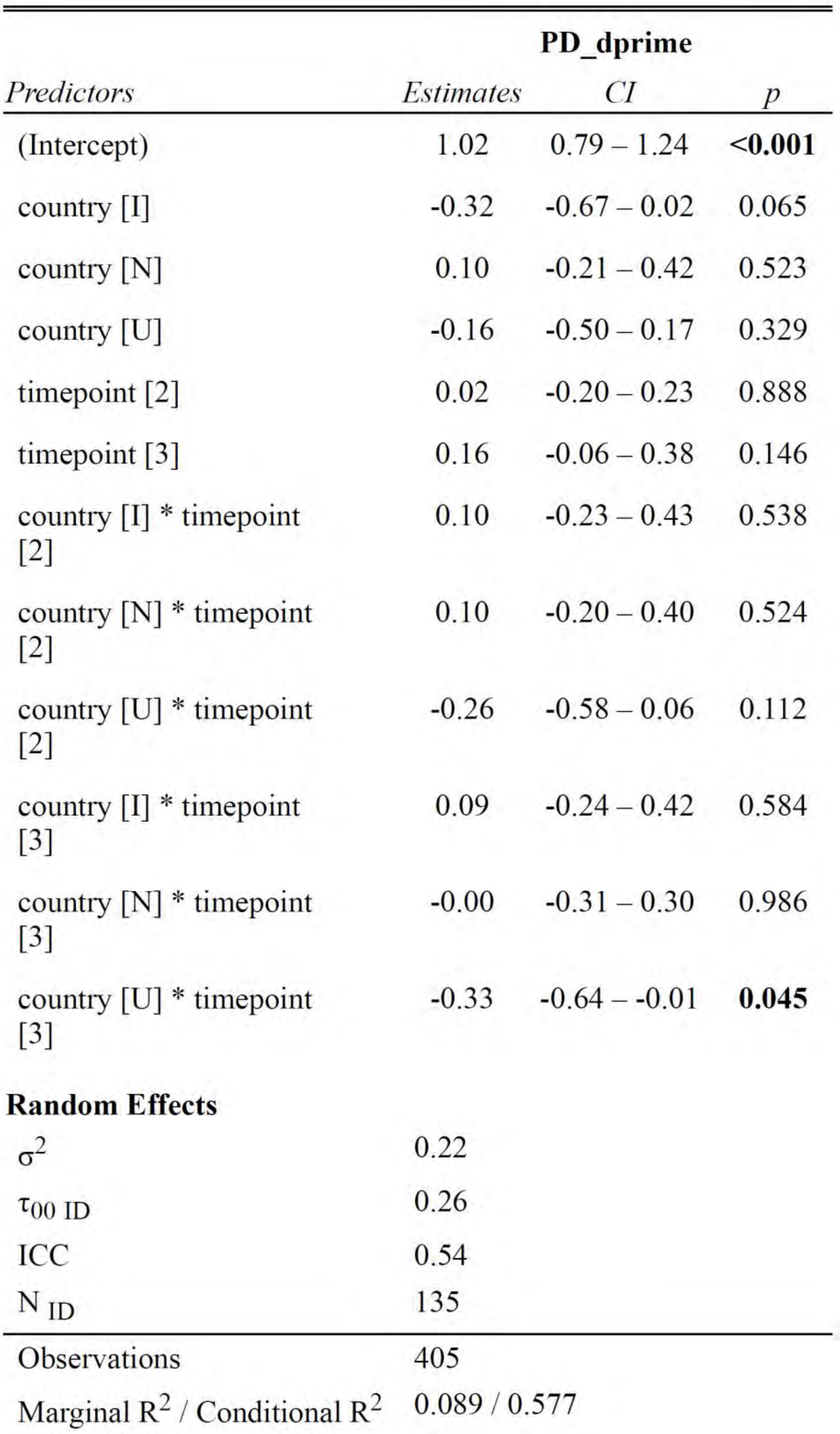
		GJ_score	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.49	1.10 – 1.88	**<0.001**
country [I]	0.20	-0.37 – 0.77	0.493
country [N]	0.33	-0.21 – 0.86	0.230
country [U]	-0.50	-1.06 – 0.05	0.075
Time [2]	0.92	0.75 – 1.09	**<0.001**
country [I] * Time [2]	-0.13	-0.37 – 0.12	0.310
country [N] * Time [2]	-0.01	-0.24 – 0.22	0.933
country [U] * Time [2]	-0.26	-0.50 – -0.02	**0.032**
Random Effects			
σ^2	0.13		
τ_{00} Subject	1.24		
ICC	0.91		
N Subject	139		
Observations	278		
Marginal R^2 / Conditional R^2	0.182 / 0.925		

Table 5 Overview progress over time Phoneme Discrimination M-group



	PD_dprime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.91	0.55 – 1.26	**<0.001**
country [G]	0.28	-0.20 – 0.76	0.255
country [I]	-0.14	-0.65 – 0.36	0.579
country [N]	0.29	-0.19 – 0.77	0.233
country [U]	-0.06	-0.57 – 0.44	0.800
timepoint [2]	0.19	-0.13 – 0.51	0.240
timepoint [3]	0.36	0.04 – 0.68	**0.026**
country [G] * timepoint [2]	0.03	-0.41 – 0.46	0.901
country [I] * timepoint [2]	-0.06	-0.51 – 0.40	0.808
country [N] * timepoint [2]	-0.26	-0.70 – 0.17	0.232
country [U] * timepoint [2]	-0.47	-0.93 – -0.01	**0.044**
country [G] * timepoint [3]	-0.24	-0.68 – 0.20	0.279
country [I] * timepoint [3]	-0.01	-0.47 – 0.44	0.949
country [N] * timepoint [3]	-0.41	-0.85 – 0.02	0.061
country [U] * timepoint [3]	-0.63	-1.09 – -0.18	**0.007**
Random Effects			
σ^2	0.22		
$\tau_{00\ ID}$	0.32		
ICC	0.59		
N_{ID}	89		
Observations	266		
Marginal R^2 / Conditional R^2	0.100 / 0.632		

Table 6 Overview progress over time Phoneme Discrimination, all conditions



	PD_dprime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.02	0.79 – 1.24	**<0.001**
country [I]	-0.32	-0.67 – 0.02	0.065
country [N]	0.10	-0.21 – 0.42	0.523
country [U]	-0.16	-0.50 – 0.17	0.329
timepoint [2]	0.02	-0.20 – 0.23	0.888
timepoint [3]	0.16	-0.06 – 0.38	0.146
country [I] * timepoint [2]	0.10	-0.23 – 0.43	0.538
country [N] * timepoint [2]	0.10	-0.20 – 0.40	0.524
country [U] * timepoint [2]	-0.26	-0.58 – 0.06	0.112
country [I] * timepoint [3]	0.09	-0.24 – 0.42	0.584
country [N] * timepoint [3]	-0.00	-0.31 – 0.30	0.986
country [U] * timepoint [3]	-0.33	-0.64 – -0.01	**0.045**
Random Effects			
σ^2	0.22		
$\tau_{00\ ID}$	0.26		
ICC	0.54		
N_{ID}	135		
Observations	405		
Marginal R^2 / Conditional R^2	0.089 / 0.577		