INSTITUT FÜR MEHRSPRACHIGKEIT
INSTITUT DE PLURILINGUISME
ISTITUTO DI PLURILINGUISMO
INSTITUT DA PLURILINGUITAD
INSTITUTE OF MULTILINGUALISM

CSP **Center scientific da cumpetenza per la plurilinguitad** Cogniziun Società Formation
CSP **Centro scientifico di competenza per il plurilinguismo** Scuola Arbeit Communauté
CSP **Centre scientifique de compétence sur le plurilinguisme** Migrazione Society Scola
KFM **Wissenschaftliches Kompetenzzentrum für Mehrsprachigkeit** Politique Cognizione
RCM **Research Centre on Multilingualism** Comunità Migration Work Ecole Gemeinschaft

Institute of Multilingualism, Fribourg, Switzerland
Research Centre on Multilingualism, Fribourg, Switzerland

Language aptitude in primary school (LAPS)

TECHNICAL REPORT

compiled by Jan Vanhove
`jan.vanhove@unifr.ch`

Available from https://osf.io/hstv7/
Last update: January 22, 2021

Contents

Part I

Context

Chapter 1

Introduction

1.1 Purpose of this technical report

This report documents how the data collected in the *Language Aptitude* project were coded, transformed, and analysed so that it may serve as a point of reference common to all researchers working on this project and so that it can be referred to in research papers to avoid cramming them with tedious details.

This report is available from https://osf.io/hstv7/, alongside datasets, scripts, and further materials.

1.2 Team

- Raphael Berthele (principal investigator, Fribourg)
- Isabelle Udry (project manager, Fribourg)
- Carina Steiner (scientific collaborator, Zurich University of Teacher Education)
- Hansjakob Schneider (scientific partner, Zurich University of Teacher Education)
- Jan Vanhove (scientific collaborator, Fribourg)

1.3 Acknowledgements

The LAPS project was funded by the Research Centre on Multilingualism at the University of Fribourg and Teacher Training College of Fribourg.

Many people contributed to the successful running of the project. First of all, a panel of experts has guided us with their valuable advice throughout the entire endeavour: Esther Geva, Joachim Grabowski, Susanne Reiterer. We would like to thank Amelia

Lambelet for her contribution to LAPS I, especially in planning and organising the test battery, and Peter Lenz for generously sharing his expertise and assisting us in selecting a suitable English measure. We thank our fieldworkers for their help with data collection and processing. Their dedication has greatly added to the quality of our data: Josef Adler, Thomas Aeppli, Nael Ackermann, Alessandra Dedei, Kinga Dobrowolska, Paola Gagliardi, Noemi Gloor, Alessandra Gregori, Laura Hodel, Rachel Howkins, Patricia Isler, Alexandra Jaszkowski, Jasmin Koch, Luca Krenger, Bente Lowin Kropf, Nina Müller, Heike Reimann, Pauline Robert-Charrue, Maja Schärer, Sarah Singh, Fabio Soares, Laura Sopa, Tanja Zepf, Catarina Zweidler. And last but not least: We thank the teachers and their pupils who remained committed to our project. It could not have happened without them.

Part II

Project design

Chapter 2

Participants and data collections

2.1 Overview

The study's participants were pupils attending school in different municipalties in the Canton of Zurich. At the start of the study in the Autumn of 2017, these pupils were in 4th or 5th grade. The study adopted a longitudinal design with data collections in the autumn of 2017 (= T1), the spring of 2018 (= T2) and the spring of 2019 (= T3). At T2, the pupils who started out in 4th and 5th grade were still in 4th and 5th grade, respectively; at T3, they were in 5th and 6th grade, respectively.

2.2 Participants

The pupils were clustered in 32 classes (Table 2.1).[1] Most classes had exclusively either 4th- or 5th-graders, but four classes were mixed. One class (viz., the class referred to as Class 4) was a special-needs class and contained five pupils.

Table 2.1: Number of participating classes.

Cohort	Number of participating classes
4th grade at T1	13
5th grade at T1	15
Mixed (4th and 5th grade at T1)	4

Due to drop-outs, illness etc., the number of pupils varied between the data collections. Table 2.2 lists the number of pupils who contributed pertinent data at a given data

[1] In the datasets, these classes are labelled with numbers from 1 through 33, but there is no class 26.

collection, viz., who filled out a questionnaire, took part in an aptitude test and/or took part in an English-language test. The mean ages in this table were computed with respect to October 1 for the data collections in the autumn and with respect to May 15 for the data collections in the spring.

Table 2.2: Number of participating pupils and their mean age.

Cohort	Time of data collection (grade)	Mean age	Number of pupils
4th grade at T1	T1, autumn 2017 (4th grade)	9;11	289
	T2, spring 2018 (4th grade)	10;6	274
	T3, spring 2019 (5th grade)	11;6	260
5th grade at T1	T1, autumn 2017 (5th grade)	10;11	326
	T2, spring 2018 (5th grade)	11;7	304
	T3, spring 2019 (6th grade)	12;7	306

Table 2.3 shows how many pupils took part in different data collections as well as the total number of participants who contributed data to the study.

Table 2.3: Number of pupils by participation rate.

Cohort	Participation rate	Number of pupils
4th grade at T1	T1, T2 and T3	241
	T1 and T2 only	24
	T1 and T3 only	10
	T2 and T3 only	9
	T1 only	14
	T2 only	0
	T3 only	0
	Total 4th grade cohort	298
5th grade at T1	T1, T2 and T3	276
	T1 and T2 only	15
	T1 and T3 only	18
	T2 and T3 only	12
	T1 only	17
	T2 only	1
	T3 only	0
	Total 5th grade cohort	339
Grand total		637

2.3 Data collections

Table 2.4 outlines the study's main data sources. Since the pupils' task and questionnaire battery was more extensive at T1, this data collection took place in two sessions. At T2 and T3, the data collections took place in a single session.

Table 2.4: Data sources at the three data collections.

Data source	T1	T2	T3	Comment
Pupils				
Questionnaire English	yes	yes	yes	slight changes T1 vs. T2/T3
Questionnaire German	yes	yes	yes	
Questionnaire French	no	yes	yes	T2: only a subset of the questions for 4th graders
Locus of control	yes	no	no	
Alphabet task	yes	no	no	
CFT (intelligence)	yes	no	no	
Backward digit span	yes	no	no	
Forward digit span	yes	no	no	
Corsi blocks	yes	no	no	
LLama battery	yes	no	no	
ELFE	yes	yes	yes	T1: full test; T2/T3: sentences only
Oxford Placement Test	yes	no	no	abandoned for fear of ceiling effects
English C-tests	no	yes	yes	used in lieu of the Oxford Placement Test
GEFT	yes	no	no	
MLAT-E	yes	yes	yes	
PLAB	yes	yes	yes	
Parents				
Questionnaire	yes	no	no	
Teachers				
Questionnaire	yes	yes	yes	

Part III

Code book

Chapter 3

Organisation

This code book describes the variables and their coding as they appear in the data sets `all_data.csv`, `training_set.csv`, `test_set.csv`, `laps2_full_dataset.csv`, and `construct_scores.csv`. These datasets are available from https://osf.io/hstv7.

`all_data.csv` is the master dataset that contains all information collected in the study. Each row contains all of the data for a single participant. The time at which the data were collected is specified in the column names. For instance, the information in column `CQEng_T1_FB31` was collected at T1, whereas the information in column `PLAB_T2_PLAB4` was collected at T2.

`training_set.csv` and `test_set.csv` are non-overlapping subsets of `all_data.csv`; see Chapter 12. The data in `training_set.csv` served as the study's training data, i.e., the playground for exploratory analyses and model selection. The data in `test_set.csv` served as the study's test or validation data: these data weren't considered during exploratory analyses and model selection, but once a satisfactory model was agreed upon, its predictive power was assessed on the test set.

`laps2_full_dataset.csv` is the final compilation of both the training and test set. It contains all item-level information and construct scores based on them (e.g., the participants' performance on the MLAT at the second data collection). `construct_scores.csv` contains a subset of the variables in `laps2_full_dataset.csv`, viz., the construct scores but not item-level information.

Chapter 4

Structural information

The variables in this chapter pertain mainly to how, where and when the data collections were carried out.

`StudentID`: A unique identifier for each student in the format `0.12.34`. The first digit specifies the grade the student was in at T1 (either `4` or `5`). The next two digits specify the class the student is in (from `01` to `33`, excluding 26). The last two digits identify the student within the class. Example: `4.12.07` means that the student in question was in grade 4 at the first data collection and is the 7th student of class 12. Students retained their identifier for all three data collections, i.e., Student `4.12.07` is also referred to as Student `4.12.07` even at the third data collection when (s)he was in 5th grade.

`Grade`: The grade the student was in *at the first data collection* (either `4` or `5`).

`Class`: A numeric ID of the student's class (from `01` to `33`, excluding 26).

`Class_Gemeindetyp`: How is the municipality where the school is located classified? (`1` = regional centre, `2` = suburban municipality with concentrated urban development, `3` = high-income municipality, `4` = periurban municipalty without concentrated urban development)

`Class_StudentsTotalT1`, `Class_StudentsTotalT2`, `Class_StudentsTotalT3`: The number of students in the class at T1, T2 and T3, respectively.

`Class_StudentsStudyT1`, `Class_StudentsStudyT2`, `Class_StudentsStudyT3`: The number of study participants in the class at T1, T2 and T3, respectively.

`Class_DateT1S1`: Date of the first session of the first data collection (T1). Format: `dd.mm.yyyy`.

`Class_DateT1S2`: Date of the second session of the first data collection (T1). Format: `dd.mm.yyyy`.

`Class_DateT2`: Date of the second data collection (T2). Format: `dd.mm.yyyy`.

`Class_DateT3`: Date of the third data collection (T3). Format: `dd.mm.yyyy`.

`Class_T1_Session1`: When during the day did the first session of T1 take place? (`1` = morning, 45' of testing, 15'-20' break, 90' of testing; `2` = morning, 90' of testing, 15-20' break, 45' testing; `3` = morning, 135' of testing with smaller breaks; `4` = afternoon, 45' of testing, 15'-20' break, 90' of testing; `5` = afternoon, 90' of testing, 15'-20' break, 45' of testing; `6` = afternoon, 135' of testing with smaller breaks)

`Class_T1_Session2`: When during the day did the second session of T1 take place? (`1` = morning, before the break; `2` = morning, after the break; `3` = morning, with one part before and one part after the break; `4` = afternoon)

`Class_T2_Session1`: When during the day did T2 take place? (`7` = morning, before the break; `8` = morning, after the break; `9` = afternoon)

`Class_T3_Session1`: When during the day did T3 take place? (`7` = morning, before the break; `8` = morning, after the break; `9` = afternoon)

`Class_T1_Comment`, `Class_T2_Comment`, `Class_T3_Comment`: Text fields with comments about the first, second and third data collections, respectively.

`Class_ElfeSchonGelöst`: Text field with comments about whether the ELFE test battery had already been used in class prior to the study.

`Class_GeneralComments`: Text field with comments about the class.

`Class_FrenchTeacher`: Is the French teacher also the `class teacher` or another person (`other`)?

`Class_EnglishTeacher`: Is the English teacher also the `class teacher` or another person (`other`)?

`Class_SameTeacher_en_fr`: Are the French and English teacher the same person? (`yes`; `no`)

`AddInfo_QuestENComments`: Participants' comments about the English questionnaire.

`AddInfo_QuestFRComments`: Participants' comments about the French questionnaire.

`AddInfo_DispEN`: Exempted from English lessons? (`ja` = yes or `NA` (= no))

`AddInfo_DispFR`: Exempted from French lessons? (`ja` = yes or `NA` (= no))

`AddInfo_ElfeSchonGelöst`: Text field with comments about whether the participant had already completed the ELFE prior to the study.

`TrainingSet`: Was the participant part of the training set or of the test set? (`1` = training set; `0` = test set; only in `laps2_full_dataset.csv`)

Chapter 5

Pupil questionnaire

The variables in this chapter were reported by means of a questionnaire filled out by the students themselves. A questionnaire was administered to the pupils at all three data collections, though some of the questions differed (see below). All questions were asked in German.

5.1 General information

`CQEng_T1_DateOfBirth`: The student's date of birth. Format: `yyyy-mm-dd`, or `NA`. For one student (`4.01.10`), only the year and month of birth are known; for computation purposes, we set the day at the 15th of the month of birth.

`CQEng_T1_Sex`: The student's sex (`boy`, `girl`; `NA`).

5.2 English and German

One part of the questionnaires filled out by the pupils concerned their motivation with respect to English and German (only self-concept for the latter). Table 5.1 lists which questionnaire items were assumed to tap into which affective dimensions and motivational constructs.

All variables listed in this section could take the values `1`, `2`, `3`, `4`, and `NA` (not available). The value `4` indicates maximum agreement and the value `1` indicates maximum disagreement.[1]

[1] When the student assistants entered the data at T1, the coding was switched. In preparing the datasets, however, we made sure that the value `4` meant maximum agreement even at T1.

Table 5.1: Questionnaire items and the constructs they are assumed to reflect.

Construct	Items	Comment
Extrinsic motivation: school	4, 5, 8	
Extrinsic motivation: leisure	9, 10, 11	
Intrinsic motivation	1, 6, 13, 14	Q13 and Q14 were not asked at T2 and T3. Q13 should be recoded when computing construct scores.
Usefulness as lingua franca	2, 3, 7, 12	
Foreign-language anxiety	19, 20, 21, 22, 23	
Self-concept (English/French)	24, 25, 26	
Self-concept (German)	27, 28, 29, 30	
Parental encouragement	36, 37, 38, 39, 40	Q37 and Q39 were not asked at T2 and T3.
Teacher motivation	31, 32, 33, 34, 35	Q35 was not asked at T2 and T3.
Dedication	15, 16, 17, 18	
Future self	41, 42, 43	Not collected at T1.

5.2.1 "I learn English ..."

`CQEng_T1_FB01`: "... because I like to speak English."

`CQEng_T1_FB02`: "... because many people in the world speak English."

`CQEng_T1_FB03`: "... because later I'll be able to communicate with people from all over the world."

`CQEng_T1_FB04`: "... because I'd like to get good grades."

`CQEng_T1_FB05`: "... so that I'll be as good as the others in class."

`CQEng_T1_FB06`: "... because I like to listen to English."

`CQEng_T1_FB07`: "... because later I'll be able to get to know people from various countries."

`CQEng_T1_FB08`: "... so that I'll be good in school."

`CQEng_T1_FB09`: "... to understand the text of my favourite music."

`CQEng_T1_FB10`: "... to understand what I read on the Internet."

`CQEng_T1_FB11`: "... to understand my computer game."

CQEng_T1_FB12: "… to talk with English-speaking people when on vacation."

CQEng_T1_FB13: "I only learn English because I have to."

CQEng_T2_FB01, CQEng_T2_FB02, CQEng_T2_FB03, CQEng_T2_FB04, CQEng_T2_FB05, CQEng_T2_FB06, CQEng_T2_FB07, CQEng_T2_FB08, CQEng_T2_FB09, CQEng_T2_FB10, CQEng_T2_FB11, CQEng_T2_FB12: Responses to the same questions as above, but at T2. Question 13 was not asked at T2.

CQEng_T3_FB01, CQEng_T3_FB02, CQEng_T3_FB03, CQEng_T3_FB04, CQEng_T3_FB05, CQEng_T3_FB06, CQEng_T3_FB07, CQEng_T3_FB08, CQEng_T3_FB09, CQEng_T3_FB10, CQEng_T3_FB11, CQEng_T3_FB12: Responses to the same questions as above, but at T3. Question 13 was not asked at T3.

5.2.2 English classes

CQEng_T1_FB14: "I enjoy English class."

CQEng_T1_FB15: "I cooperate during English class."

CQEng_T1_FB16: "I learn a lot for English."

CQEng_T1_FB17: "I make an effort during English class."

CQEng_T1_FB18: "I do my utmost to learn English."

CQEng_T1_FB19: "I'm afraid to make errors during English class."

CQEng_T1_FB20: "I often feel stressed during English class, because everything is so difficult."

CQEng_T1_FB21: "I get nervous when I have talk during English class."

CQEng_T1_FB22: "I'd rather not raise my hand during English class to avoid given wrong answers."

CQEng_T1_FB23: "I'm always glad when I don't have to say anything during English class."

CQEng_T2_FB15, CQEng_T2_FB16, CQEng_T2_FB17, CQEng_T2_FB18, CQEng_T2_FB19, CQEng_T2_FB20, CQEng_T2_FB21, CQEng_T2_FB22, CQEng_T2_FB23: Responses to the same questions as above, but at T2. Question 14 was not asked at T2.

CQEng_T3_FB15, CQEng_T3_FB16, CQEng_T3_FB17, CQEng_T3_FB18, CQEng_T3_FB19, CQEng_T3_FB20, CQEng_T3_FB21, CQEng_T3_FB22, CQEng_T3_FB23: Responses to the same questions as above, but at T3. Question 14 was not asked at T3.

5.2.3 English and German in school

CQEng_T1_FB24: “I already understand lots of English words.”

CQEng_T1_FB25: “I think English is pretty easy.”

CQEng_T1_FB26: “I already know English fairly well.”

CQEng_T1_FB27: “I’m good in German (as a school subject).”

CQEng_T1_FB28: “I think that German (as a school subject) is easy.”

CQEng_T1_FB29: “German (as a school subject) doesn’t give me trouble.”

CQEng_T1_FB30: “I’m able to follow well in German class.”

CQEng_T2_FB24, CQEng_T2_FB25, CQEng_T2_FB26, CQEng_T2_FB27, CQEng_T2_FB28, CQEng_T2_FB29, CQEng_T2_FB30: Responses to the same questions as above, but at T2.

The following three questions were added to the questionnaire at T2:

CQEng_T2_FB41: “I can imagine that one day I’ll be able to speak English very well.”

CQEng_T2_FB42: “I can imagine that I’ll use English a lot in the future.”

CQEng_T2_FB43: “I can imagine that later I’ll talk with English-speaking people.”

CQEng_T3_FB24, CQEng_T3_FB25, CQEng_T3_FB26, CQEng_T3_FB27, CQEng_T3_FB28, CQEng_T3_FB29, CQEng_T3_FB30, CQEng_T3_FB41, CQEng_T3_FB42, CQEng_T3_FB43: Responses to the same questions as above, but at T3.

5.2.4 Teacher and parents

CQEng_T1_FB31: “My English teacher can get me interested in English (as a school subject).”

CQEng_T1_FB32: “My English teacher shows me that English is a beautiful language.”

CQEng_T1_FB33: “I like how my English teacher teaches English.”

CQEng_T1_FB34: “My English teacher shows me that it’s important to learn English.”

CQEng_T1_FB35: “My English teacher enjoys teaching English.”

CQEng_T1_FB36: “My parents encourage me to learn English.”

CQEng_T1_FB37: “My parents are glad when I’m good at English.”

CQEng_T1_FB38: “My parents encourage me to practice English as often as possible.”

CQEng_T1_FB39: “My performance in English is important to my parents.”

CQEng_T1_FB40: “My parents encourage me to learn English in my spare-time.”

CQEng_T2_FB31, CQEng_T2_FB32, CQEng_T2_FB33, CQEng_T2_FB34, CQEng_T2_FB36, CQEng_T2_FB38, CQEng_T2_FB40: Responses to the same questions as above, but at T2. Questions 35, 37 and 39 were not asked at T2.

CQEng_T3_FB31, CQEng_T3_FB32, CQEng_T3_FB33, CQEng_T3_FB34, CQEng_T3_FB36, CQEng_T3_FB38, CQEng_T3_FB40: Responses to the same questions as above, but at T3. Questions 35, 37 and 39 were not asked at T3.

5.3 French

At T2 and T3, the questions listed above were also asked but with respect to French.

CQFr_T2_FB01, CQFr_T2_FB02, CQFr_T2_FB03, CQFr_T2_FB04, CQFr_T2_FB05, CQFr_T2_FB06, CQFr_T2_FB07, CQFr_T2_FB08, CQFr_T2_FB09, CQFr_T2_FB10, CQFr_T2_FB11, CQFr_T2_FB12, CQFr_T2_FB15, CQFr_T2_FB16, CQFr_T2_FB17, CQFr_T2_FB18, CQFr_T2_FB19, CQFr_T2_FB20, CQFr_T2_FB21, CQFr_T2_FB22, CQFr_T2_FB23, CQFr_T2_FB24, CQFr_T2_FB25, CQFr_T2_FB26, CQFr_T2_FB31, CQFr_T2_FB32, CQFr_T2_FB33, CQFr_T2_FB34, CQFr_T2_FB36, CQFr_T2_FB38, CQFr_T2_FB40, CQFr_T2_FB41, CQFr_T2_FB42, CQFr_T2_FB43: Responses to the same questions as in Section 5.2, but with respect to French at T2.

CQFr_T3_FB01, CQFr_T3_FB02, CQFr_T3_FB03, CQFr_T3_FB04, CQFr_T3_FB05, CQFr_T3_FB06, CQFr_T3_FB07, CQFr_T3_FB08, CQFr_T3_FB09, CQFr_T3_FB10, CQFr_T3_FB11, CQFr_T3_FB12, CQFr_T3_FB15, CQFr_T3_FB16, CQFr_T3_FB17, CQFr_T3_FB18, CQFr_T3_FB19, CQFr_T3_FB20, CQFr_T3_FB21, CQFr_T3_FB22, CQFr_T3_FB23, CQFr_T3_FB24, CQFr_T3_FB25, CQFr_T3_FB26, CQFr_T3_FB31, CQFr_T3_FB32, CQFr_T3_FB33, CQFr_T3_FB34, CQFr_T3_FB36, CQFr_T3_FB38, CQFr_T3_FB40, CQFr_T3_FB41, CQFr_T3_FB42, CQFr_T3_FB43: Responses to the same questions as in Section 5.2, but with respect to French at T3.

At T2, 4th graders were only asked questions 1, 2, 3, 6, 7, 9, 10, 11, 12, 24, 25, 26, 41, 42 and 43 since the other questions did not yet apply to them. At T3, all pupils were asked all questions.

5.4 Locus of control

Questions related to the children's locus of control were included in the questionnaire at T1 only.

Agreement with these statements was coded as 1, disagreement as 0. (NA = not available)

Agreement to items 3, 13, and 19 is assumed to reflect an **internal** locus of control; agreement to the other items is assumed to reflect an **external** locus of control.

CQEng_T1_LOC01: “Some children are naturally lucky.”

CQEng_T1_LOC02: “Making an effort isn’t usually worth it because in the end nothing works out.”

CQEng_T1_LOC03: “Parents pay heed to what they children say.”

CQEng_T1_LOC04: “If I wish for something, it can happen.”

CQEng_T1_LOC05: “It’s nearly impossible to change my parents’ opinion.”

CQEng_T1_LOC06: “When I’ve done something wrong, there’s little I can do to make it right again.”

CQEng_T1_LOC07: “Most sporty children are naturally sporty.”

CQEng_T1_LOC08: “Most children of my age are stronger than me.”

CQEng_T1_LOC09: “The best way to solve a problem is to not think about it.”

CQEng_T1_LOC10: “Four-leaved clovers bring me luck.”

CQEng_T1_LOC11: “When a child of my age wants to hit me, there’s little I can do about it.”

CQEng_T1_LOC12: “When someone’s mean to me, it’s usually without reason.”

CQEng_T1_LOC13: “I can do something so that no bad things happen to me.”

CQEng_T1_LOC14: “It’s usually no use to try and get what I want at home.”

CQEng_T1_LOC15: “If another child wants to exclude me, there’s little I can do about it.”

CQEng_T1_LOC16: “Normally I don’t have a say in what we eat at home.”

CQEng_T1_LOC17: “When someone doesn’t like me, there’s little I can do about it.”

CQEng_T1_LOC18: “It’s usually of little use to make an effort in school because the other children are smarter than me.”

CQEng_T1_LOC19: “Things work out better if you plan them in advance.”

CQEng_T1_LOC20: “I have no say in what we do at home.”

Chapter 6

Parental questionnaire

The variables in this chapter were reported by means of a questionnaire filled out by the students' parents. The parental questionnaire was only filled out at the first data collection in which the student participated; for the vast majority of participants this means T1. Consequently, these data aren't differentiated by data collection. All questions were asked in German.

6.1 Personal and linguistic background

`PQ_T1_CountryChild`: The student's country of birth. In most cases, a two-letter country code was used, though `NA` means *not available* and not Namibia. See https://www.worldatlas.com/aatlas/ctycodes.htm. `Kosovo`, `Montenegro` and `Tibet` were spelt out.

`PQ_T1_CountryFather`: The student's father's country of birth. (two-letter country code)

`PQ_T1_CountryMother`: The student's mother's country of birth. (two-letter country code)

`PQ_T1_KG1_ZH`: Did the student go to the 1st year of kindergarten in a German-language school in the Canton of Zurich? (`0` = no, `1` = yes; `NA`)

`PQ_T1_KG2_ZH`: Did the student go to the 2nd year of kindergarten in a German-language school in the Canton of Zurich? (`0` = no, `1` = yes; `NA`)

`PQ_T1_K1_ZH`: Did the student go to 1st grade in a German-language school in the Canton of Zurich? (`0` = no, `1` = yes; `NA`)

`PQ_T1_K2_ZH`: Did the student go to 2nd grade in a German-language school in the Canton of Zurich? (`0` = no, `1` = yes; `NA`)

PQ_T1_K3_ZH: Did the student go to 3rd grade in a German-language school in the Canton of Zurich? (0 = no, 1 = yes; NA)

PQ_T1_K4_ZH: Did the student go to 4th grade in a German-language school in the Canton of Zurich? (0 = no, 1 = yes; NA)

PQ_T1_NativeLanguages: A comma-separated list of the student's native language(s).

PQ_T1_FamilyLanguage: The language(s) spoken most often in the family. German = mostly German or Swiss German; GermanMulti = German and another language or other languages; other = mostly another language. (or NA)

PQ_T1_OtherFamilyLanguages: If PQ_T1_FamilyLanguage = GermanMulti or other, this variable specifies the other family language(s). (or NA)

PQ_T1_RWLanguages: A list of the languages in which the child first learnt to read and write.

PQ_T1_EducationFather: The father's highest level of education: Primary school, Secondary school, Apprenticeship, ProfessionalMaturity, AcademicMaturity, HigherVocationalEducation or UniversityDegree (or NA).

PQ_T1_EducationMother: The mother's highest level of education: Primary school, Secondary school, Apprenticeship, ProfessionalMaturity, AcademicMaturity, HigherVocationalEducation or UniversityDegree (or NA).

6.2 Household and budget

PQ_T1_NrBooks: "How many books are there in your household?": 0-10, 11-25, 26-100, 101-200, 201-500, 500+ (or NA)

The following four variables could take the values 1 (not at all sufficient), 2 (rarely sufficient), 3 (sometimes sufficient), 4 (usually sufficient), 5 (always sufficient), and NA (not available).

PQ_T1_MonthlyBills: "Is there enough money to pay the monthly bills?"

PQ_T1_MedicalCare: "Is there enough money for medical and dental care?"

PQ_T1_Saving: "Is there enough money for saving?"

PQ_T1_Holidays: "Is there enough money for travelling and holidays?"

PQ_T1_Earnings: "What's the family's monthly income?": 1 = less than 5000 Swiss francs, 2 = 5000-10000 Swiss francs, 3 = 10000-15000 Swiss francs, 4 = 15000-20000 Swiss francs, 5 = more than 20000 Swiss francs. (or NA)

6.3 School and education

PQ_T1_DaZ: Is the child currently attending or has the child ever attended German as a second language classes? (no, yes; NA)

PQ_T1_KG1_DaZ, PQ_T1_KG2_DaZ, PQ_T1_K1_DaZ, PQ_T1_K2_DaZ, PQ_T1_K3_DaZ, PQ_T1_K4_DaZ, PQ_T1_K5_DaZ: Did the child attend German as a second language classes in 1st kindergarten, 2nd kindergarten, 1st grade, 2nd grade, 3rd grade, 4th grade or 5th grade, respectively? (0 = no, 1 = yes; NA)

PQ_T1_HSK: Is the child currently attending or has the child ever attended heritage language and culture classes? (no, yes; NA)

PQ_T1_KG1_HSK, PQ_T1_KG2_HSK, PQ_T1_K1_HSK, PQ_T1_K2_HSK, PQ_T1_K3_HSK, PQ_T1_K4_HSK, PQ_T1_K5_HSK: Did the child attend heritage language and culture classes in 1st kindergarten, 2nd kindergarten, 1st grade, 2nd grade, 3rd grade, 4th grade or 5th grade, respectively? (0 = no, 1 = yes; NA)

PQ_T1_HomeworkEnglish: How many hours a week does the child spend on homework and learning for English class? 0 = 0 hours, 1 = up till 0.5 hours, 2 = up till 1 hour, 3 = up till 1.5 hours, 4 = up till 2 hours, 5 = up till 3 hours, 6 = up till 4 hours, 7 = more than 4 hours; NA.

6.4 Additional comments

PQ_T1_Country_Comments: Text field with parents' comments about the participants' and their parents' country of birth.

PQ_T1_SchoolZH_Comments: Text field with parents' comments about the participants' years of schooling in the canton of Zurich.

PQ_T1_NativeLanguages_Comments: Text field with parents' comments about the participants' and their parents' first language.

PQ_T1_RWLanguages_Comments: Text field with parents' comments about the participants' reading and writing skills.

PQ_T1_Resources_Comments: Text field with parents' comments about the money-related questions in the questionnaire.

PQ_T1_GeneralComments: General comments about the parental questionnaire by the parents.

Chapter 7

Teacher questionnaire

The class teachers provided information about any special pedagogical measures students received.

PM_T1_IFGeneral, PM_T2_IFGeneral, PM_T3_IFGeneral: “Allgemeine integrierte Förderung?”: no, yes, unspecified, yes, low, yes, middle, yes, high

PM_T1_IFMaths, PM_T2_IFMaths, PM_T3_IFMaths: “Integrierte Förderung Mathe?”: no, yes, unspecified, yes, low, yes, middle, yes, high

PM_T1_IFGerman, PM_T2_IFGerman, PM_T3_IFGerman: “Integrierte Förderung Deutsch?”: no, yes, unspecified, yes, low, yes, middle, yes, high

PM_T1_IFEnglish, PM_T2_IFEnglish, PM_T3_IFEnglish: “Integrierte Förderung Englisch?”: no, yes, unspecified, yes, low, yes, middle, yes, high

PM_T1_ISR, PM_T2_ISR, PM_T3_ISR: “Integrierte Sonderschulung Regelklasse?”: no, yes

PM_T1_DaZ, PM_T2_DaZ, PM_T3_DaZ: “DaZ?”: no, yes

PM_T1_BBF, PM_T2_BBF, PM_T3_BBF: “Begabten- und Begabungsförderung?”: no, yes

PM_T1_Logopedics, PM_T2_Logopedics, PM_T3_Logopedics: “Logopädie?”: no, yes

PM_T1_Ergotherapy, PM_T2_Ergotherapy, PM_T3_Ergotherapy: “Ergotherapie?”: no, yes

PM_T1_Psychomotorics, PM_T2_Psychomotorics, PM_T3_Psychomotorics: “Psychomotorik?”: no, yes

PM_T1_HearingImpaired, PM_T2_HearingImpaired, PM_T3_HearingImpaired: Is the child hearing impaired?: no, yes

PM_T1_VisuallyImpaired, PM_T2_VisuallyImpaired, PM_T3_VisuallyImpaired: Is the child visually impaired?: no, yes

PM_T1_Comment, PM_T2_Comment, PM_T3_Comment: A text field for additional comments.

Chapter 8

Language and cognitive tests

The first part of the variable name refers to the task or test battery the variable refers to; the second part (`T1`, `T2`, `T3`) to whether the variable was collected at the first, second or third data collection; and the third part specifies which variable specifically was extracted from the task.

8.1 Alphabet task

The alphabet task was administered at T1 only. Students were instructed to write down the alphabet from memory as quickly as possible without sacrificing legibility during 60 seconds. After the first 15 seconds, they were instructed to put a bar after the letters they had already written down.

`Alpha_T1_Score15`: Number of correctly provided letters in the first 15 seconds (legible and in correct order).

8.2 CFT

The CFT was administered at T1 only. A short version of two subtests was used: 3. Matrices (3 minutes); 4. Topological deductions (3 minutes).

`CFT_T1_CFT1`: Score on part 1 of the CFT. [0–15; NA]

`CFT_T1_CFT2`: Score on part 2 of the CFT. [0–11; NA]

`CFT_T1_Total`: Sum of `CFT_T1_CFT1` and `CFT_T1_CFT2`. [0–26; NA]

`CFT_T1_GradeNormed`: `CFT_T1_Total` converted to an IQ score depending on the participants' grade (Grade 4 vs. Grade 5) as per Table 8.1. If a participant's `CFT_T1_Total` score was too high or too low to be converted to a grade-normed IQ score (`off-scale`), the nearest grade-normed IQ score was used. For instance, if a 5th-grader obtained a `CFT_T1_Total` score of 2, their grade-normed IQ was set at the nearest value in the conversion table, viz., 53.2. The variable `CFT_T1_GradeNormedCensored` identifies these cases explicitly.

`CFT_T1_GradeNormedCensored`: Whether and how `CFT_T1_GradeNormed` was censored. [`no` = no censoring; `yes, high` = off-scale at the higher end (did not occur); `yes, low` = off-scale at the lower end]

8.3 Computer-administered cognitive tests

These variables were only collected at T1.

Procedure:

- Forward Digit Span (FDS), Backward Digit Span (BDS): Start with 2 digits, 3 trials per level. 1 in 3 trials must be correct to reach next level.
- Corsi blocks: Start with 2 squares, 3 trials per level. 1 in 3 trials must be correct to reach next level.

`Computer_T1_BDSCompleted`: Did the participant fully complete the Back Digit Span task or not (e.g., due to technical glitches)? [0 = no; 1 = yes]

`Computer_T1_BDSSpan`: Backward Digit Span, memory span.

`Computer_T1_BDSTotalCorrect`: Backward Digit Span, total number of correct words.

`Computer_T1_BDSTotalTime`: Backward Digit Span, total time (in minutes).

`Computer_T1_CorsiBlock`: Corsi Blocks, block span. This is computed as the longest length at which at least one pattern was correctly recalled.

`Computer_T1_CorsiCompleted`: Did the participant fully complete the Corsi Block task or not (e.g., due to technical glitches)? [0 = no; 1 = yes]

`Computer_T1_CorsiMemorySpan`: Corsi Blocks, memory span. ‘Memory span takes the minimum list length, adds the total number correct, and divides by the number of lists at each length.’ (http://pebl.sourceforge.net/wiki/index.php/Corsi_Blocks)

`Computer_T1_CorsiTotal`: Corsi Blocks, total score. This is computed by multiplying the block span by the total number of correct trials, but isn’t too meaningful a measure.

`Computer_T1_CorsiTotalCorrect`: Corsi Blocks, total number of correct trials, i.e., the number of trials that were correctly recalled.

Table 8.1: Conversion table: CFT total score to grade-normed IQ.

CFT_T1_Total	IQ (Grade 4)	IQ (Grade 5)
0	(off-scale)	(off-scale)
1	(off-scale)	(off-scale)
2	(off-scale)	(off-scale)
3	52.9	(off-scale)
4	57.2	53.2
5	61.5	57.5
6	65.8	61.8
7	70.2	66.2
8	74.5	70.5
9	78.8	74.8
10	83.1	79.1
11	86.4	83.4
12	90.7	86.7
13	95	91
14	100.3	95.3
15	104.6	100.6
16	110.4	104.9
17	116.2	110.8
18	120.5	116.5
19	126.8	120.8
20	133.3	127.2
21	141.7	133.9
22	149.5	142.5
23	158.2	150.2
24	(off-scale)	158.8
25	(off-scale)	(off-scale)
26	(off-scale)	(off-scale)

Computer_F1_FDSCompleted: Did the participant fully complete the Forward Digit Span task or not (e.g., due to technical glitches)? [0 = no; 1 = yes]

Computer_T1_FDSSpan: Forward Digit Span, memory span.

Computer_T1_FDSTotalCorrect: Forward Digit Span, total number of correct words.

Computer_T1_FDSTotalTime: Forward Digit Span, total time (in minutes).

Computer_T1_LLama: The percentage of correct answers on the LLAMA-D (sound recognition) task. [0–100, NA]

8.4 ELFE

All three subparts of the ELFE test were used at T1, but only sentence subpart was used at T2 and T3.

To combat ceiling effects, a time limit was imposed on the sentence subpart, see Table 8.2.

Table 8.2: Time limit for the ELFE sentence subpart.

Cohort	Limit at T1	Limit at T2	Limit at T3
4th grade at T1	3’00”	2’30”	1’45”
5th grade at T1	2’00”	1’45”	1’30”

ELFE_T1_ELFESentence: Score on the sentence part in the ELFE test at T1. [0–28, NA]

ELFE_T1_ELFEText: Score on the text part in the ELFE test at T1. [0–20, NA] Time limits: 7’00” in 4th grade, 6’00” in 5th grade.

ELFE_T1_ELFETotal: Sum of ELFE_T1_ELFESentence, ELFE_T1_ELFEText and ELFE_T1_ELFEWord. [0–120, NA]

ELFE_T1_ELFEWord: Score on the word part in the ELFE test at T1. [0–72, NA] Time limits: 3’00” in 4th grade, 2’00” in 5th grade.

ELFE_T2_ELFESentence: Score on the sentence part in the ELFE test at T2. [0–28, NA]

ELFE_T3_ELFESentence: Score on the sentence part in the ELFE test at T3. [0–28, NA]

ELFE_T1_SentencePerMinute: ELFE_T1_ELFESentence divided by the number of minutes the participants had at their disposal for the sentence part (i.e., 3 or 2 minutes).

ELFE_T2_SentencePerMinute: ELFE_T2_ELFESentence divided by the number of minutes the participants had at their disposal for the sentence part (i.e., 2.5 or 1.75 minutes).

ELFE_T3_SentencePerMinute: ELFE_T3_ELFESentence divided by the number of minutes the participants had at their disposal for the sentence part (i.e., 1.75 or 1.5 minutes).

8.5 English proficiency tests

Different tests were used at T1 on the one hand (Oxford Placement Test) and at T2 and T3 on the other hand (five C-tests) for fear of ceiling effects on the Oxford Placement Test. A time limit of four minutes was imposed on each C-test. Each C-test was scored twice: once without taking spelling errors into account, and once penalising spelling errors.

`Eng_T1_LevelListening`: Level associated with the listening subscore on the Oxford Placement Test. [A0, A1, A2, B1, B1+; NA]

`Eng_T1_LevelTotal`: Level associated with total score on the Oxford Placement Test. [A0, A1, A2, B1, B1+; NA]

`Eng_T1_LevelUse`: Level associated with the language use subscore on the Oxford Placement Test. [A0, A1, A2, B1, B1+; NA]

`Eng_T1_ScoreListening`: Score on the listening part of the Oxford Placement Test [0–81; NA].

`Eng_T1_ScoreTotal`: Total score on the Oxford Placement Test. If the pupil participated in one only subtask (in all cases: just use) and not in the other, the total score could not be computed (`NA`). [0–81; NA]

`Eng_T1_ScoreUse`: Score on the language use part of the Oxford Placement Test [0–81; NA].

`Eng_T2_CTest1`: Score on the first C-test at T1; spelling errors not penalised. [0–20; NA]

`Eng_T2_CTest1RS`: Score on the first C-test at T1; spelling errors penalised. [0–20; NA]

`Eng_T2_CTest2`: Score on the second C-test at T1; spelling errors not penalised. [0–20; NA]

`Eng_T2_CTest2RS`: Score on the second C-test at T1; spelling errors penalised. [0–20; NA]

`Eng_T2_CTest3`: Score on the third C-test at T1; spelling errors not penalised. [0–20; NA]

`Eng_T2_CTest3RS`: Score on the third C-test at T1; spelling errors penalised. [0–20; NA]

`Eng_T2_CTest4`: Score on the fourth C-test at T1; spelling errors not penalised. [0–20; NA]

`Eng_T2_CTest4RS`: Score on the fourth C-test at T1; spelling errors penalised. [0–20; NA]

`Eng_T2_CTest5`: Score on the fifth C-test at T1; spelling errors not penalised. [0–20; NA]

Eng_T2_CTest5RS: Score on the fifth C-test at T1; spelling errors penalised. [0–20; NA]

Eng_T2_Total: Sum of Eng_T2_CTest1, Eng_T2_CTest2, Eng_T2_CTest3, Eng_T2_CTest4 and Eng_T2_CTest5. [0–100; NA]

Eng_T2_TotalRS: Sum of Eng_T2_CTest1RS, Eng_T2_CTest2RS, Eng_T2_CTest3RS, Eng_T2_CTest4RS and Eng_T2_CTest5RS. [0–100; NA]

Eng_T3_CTest1, Eng_T3_CTest1RS, Eng_T3_CTest2, Eng_T3_CTest2RS, Eng_T3_CTest3, Eng_T3_CTest3RS, Eng_T3_CTest4, Eng_T3_CTest4RS, Eng_T3_CTest5, Eng_T3_CTest5RS, Eng_T3_Total, Eng_T3_TotalRS: Same as above but for T3.

8.6 GEFT

The GEFT was only administered at T1. The test consists of three subparts (part 1: 7 practice items, not scored; parts 2 and 3: 9 items per subtest). Total time: 18 minutes.

GEFT_T1_GEFT1, GEFT_T1_GEFT10, GEFT_T1_GEFT11, GEFT_T1_GEFT12, GEFT_T1_GEFT13, GEFT_T1_GEFT14, GEFT_T1_GEFT15, GEFT_T1_GEFT16, GEFT_T1_GEFT17, GEFT_T1_GEFT18, GEFT_T1_GEFT2, GEFT_T1_GEFT3, GEFT_T1_GEFT4, GEFT_T1_GEFT5, GEFT_T1_GEFT6, GEFT_T1_GEFT7, GEFT_T1_GEFT8, GEFT_T1_GEFT9: Accuracy on each of the 18 items comprising part 2 and 3 of the GEFT. [0 = incorrect, 1 = correct; NA]

GEFT_T1_GEFTUnderstood: Did the participant understand the task (i.e., at least one correct item in the unscored first part)? [no, yes; NA]

8.7 Aptitude tests

Two aptitude tests were conducted at each data collection: The MLAT and the PLAB. The following subparts were administered:

- MLAT: grammatical sensitivity test (adapted and translated version of MLAT-E, part 2)
- PLAB: inductive ability test (adapted and translated version of PLAB, form 4)

MLAT_T1_MLAT1, MLAT_T1_MLAT10, MLAT_T1_MLAT11, MLAT_T1_MLAT12, MLAT_T1_MLAT13, MLAT_T1_MLAT14, MLAT_T1_MLAT15, MLAT_T1_MLAT16, MLAT_T1_MLAT17, MLAT_T1_MLAT18, MLAT_T1_MLAT19, MLAT_T1_MLAT2, MLAT_T1_MLAT20, MLAT_T1_MLAT21, MLAT_T1_MLAT22, MLAT_T1_MLAT23, MLAT_T1_MLAT24, MLAT_T1_MLAT25, MLAT_T1_MLAT26, MLAT_T1_MLAT27, MLAT_T1_MLAT28, MLAT_T1_MLAT29, MLAT_T1_MLAT3, MLAT_T1_MLAT30, MLAT_T1_MLAT4, MLAT_T1_MLAT5, MLAT_T1_MLAT6, MLAT_T1_MLAT7, MLAT_T1_MLAT8, MLAT_T1_MLAT9:

Accuracy on each of the 30 items comprising the MLAT-E task at T1. [0 = incorrect, 1 = correct; NA]

MLAT_T2_MLAT1, MLAT_T2_MLAT10, MLAT_T2_MLAT11, MLAT_T2_MLAT12, MLAT_T2_MLAT13, MLAT_T2_MLAT14, MLAT_T2_MLAT15, MLAT_T2_MLAT16, MLAT_T2_MLAT17, MLAT_T2_MLAT18, MLAT_T2_MLAT19, MLAT_T2_MLAT2, MLAT_T2_MLAT20, MLAT_T2_MLAT21, MLAT_T2_MLAT22, MLAT_T2_MLAT23, MLAT_T2_MLAT24, MLAT_T2_MLAT25, MLAT_T2_MLAT26, MLAT_T2_MLAT27, MLAT_T2_MLAT28, MLAT_T2_MLAT29, MLAT_T2_MLAT3, MLAT_T2_MLAT30, MLAT_T2_MLAT4, MLAT_T2_MLAT5, MLAT_T2_MLAT6, MLAT_T2_MLAT7, MLAT_T2_MLAT8, MLAT_T2_MLAT9: Accuracy on each of the 30 items comprising the MLAT-E task at T2. [0 = incorrect, 1 = correct; NA]

MLAT_T3_MLAT1, MLAT_T3_MLAT10, MLAT_T3_MLAT11, MLAT_T3_MLAT12, MLAT_T3_MLAT13, MLAT_T3_MLAT14, MLAT_T3_MLAT15, MLAT_T3_MLAT16, MLAT_T3_MLAT17, MLAT_T3_MLAT18, MLAT_T3_MLAT19, MLAT_T3_MLAT2, MLAT_T3_MLAT20, MLAT_T3_MLAT21, MLAT_T3_MLAT22, MLAT_T3_MLAT23, MLAT_T3_MLAT24, MLAT_T3_MLAT25, MLAT_T3_MLAT26, MLAT_T3_MLAT27, MLAT_T3_MLAT28, MLAT_T3_MLAT29, MLAT_T3_MLAT3, MLAT_T3_MLAT30, MLAT_T3_MLAT4, MLAT_T3_MLAT5, MLAT_T3_MLAT6, MLAT_T3_MLAT7, MLAT_T3_MLAT8, MLAT_T3_MLAT9: Accuracy on each of the 30 items comprising the MLAT-E task at T3. [0 = incorrect, 1 = correct; NA]

PLAB_T1_PLAB1, PLAB_T1_PLAB10, PLAB_T1_PLAB11, PLAB_T1_PLAB12, PLAB_T1_PLAB13, PLAB_T1_PLAB14, PLAB_T1_PLAB15, PLAB_T1_PLAB2, PLAB_T1_PLAB3, PLAB_T1_PLAB4, PLAB_T1_PLAB5, PLAB_T1_PLAB6, PLAB_T1_PLAB7, PLAB_T1_PLAB8, PLAB_T1_PLAB9: Accuracy on each of the 15 items comprising the PLAB task at T1. [0 = incorrect, 1 = correct; NA]

PLAB_T2_PLAB1, PLAB_T2_PLAB10, PLAB_T2_PLAB11, PLAB_T2_PLAB12, PLAB_T2_PLAB13, PLAB_T2_PLAB14, PLAB_T2_PLAB15, PLAB_T2_PLAB2, PLAB_T2_PLAB3, PLAB_T2_PLAB4, PLAB_T2_PLAB5, PLAB_T2_PLAB6, PLAB_T2_PLAB7, PLAB_T2_PLAB8, PLAB_T2_PLAB9: Accuracy on each of the 15 items comprising the PLAB task at T2. [0 = incorrect, 1 = correct; NA]

PLAB_T3_PLAB1, PLAB_T3_PLAB10, PLAB_T3_PLAB11, PLAB_T3_PLAB12, PLAB_T3_PLAB13, PLAB_T3_PLAB14, PLAB_T3_PLAB15, PLAB_T3_PLAB2, PLAB_T3_PLAB3, PLAB_T3_PLAB4, PLAB_T3_PLAB5, PLAB_T3_PLAB6, PLAB_T3_PLAB7, PLAB_T3_PLAB8, PLAB_T3_PLAB9: Accuracy on each of the 15 items comprising the PLAB task at T3. [0 = incorrect, 1 = correct; NA]

Chapter 9

Language group

On the basis of the child, parent and teacher questionnaire data, three additional variables were created.

Multilingual: Children were considered to be multilingual if at least one of the following conditions was met (yes, no; NA):

- PQ_T1_NativeLanguages is not or not only German.
- PQ_T1_FamilyLanguage is not or not only German.
- PQ_T1_RWLanguages is not or not only German.
- PQ_T1_DaZ is yes.
- PQ_T1_HSK is yes.
- PM_T1_DaZ is yes.
- PM_T2_DaZ is yes.
- PM_T3_DaZ is yes.

L1German: Children were considered to be native speakers of German if German was among the languages listed in PQ_T1_NativeLanguages (yes, no; NA).

L1English: Children were considered to be native speakers of English if English was among the languages listed in PQ_T1_NativeLanguages (yes, no; NA).

The NAs concern participants for whom no pertinent data is available.

Chapter 10

About missing values

For **questionnaire** items, all missing data were coded as `NA`. This applies to the following cases:

- The questionnaire as a whole was not administered or returned.
- No permission was granted to use a filled-out questionnaire.
- A question did not apply to the respondent's case (typically a follow-up question).
- A question in an otherwise filled-out questionnaire was skipped.
- The answer provided was uninterpretable.

For **language and cognitive tests**, missing data were coded as either `NA` or 0 in `all_data.csv` and derived files, depending on the reason why the data were missing. However, the file `NAReasons.csv` specifies the type of missing data for all cases concerned. In `NAReasons.csv`, we distinguished between the following four categories:

- `NA1`: The data are absent or invalid and consequently provide little information about the participant's abilities. This category was coded as `NA` in `all_data.csv` and derived files and covers the following cases:
  - The test as a whole was not administered to the participant, typically due to absence.
  - No permission was granted to use the test data. Other data for the same participant could be used, however.
  - The test item data are missing or otherwise invalid because of a known or plausible technical glitch.

  - The test needed to be aborted early. Depending on the test in question, no valid data could then be recorded for the test as a whole or no data could be recorded from a given test item onwards.

- `NA2`: The item in question was left blank, and the blank response could not be attributed to a technical glitch. However, the participant was administered the test, seemed to have understood the instructions, and did attempt to tackle other items of the same test. We take such blank responses to suggest that the participant did not know the correct answer or ran out of time. Since in either case, the blank response provides information about the participant's abilities, these cases were coded as 0 in `all_data.csv` and derived files.

- `NA3`: The participant did not understand the instructions (most cases) or received additional help (Participant `4.25.07`: CFT at T1). We coded these cases as `NA` in `all_data.csv` and derived files.

- `NA4`: The participant did not attempt to tackle a single item in the test as a whole. Since some of these cases *could* be due to a technical glitch, we coded these cases as `NA` in `all_data.csv` and derived files, but other researchers could reasonably prefer to code these cases as 0 instead. All cases in the `NA4` category are listed in Table 10.1 for good measure.

Table 10.1: Participants who showed no attempt at tackling a particular test (NA4 category). We coded these data as NA, but other researchers may prefer to code some or all of these entries as 0 instead.

StudentID	Item
4.06.04	Alpha_T1_Score15
4.06.11	Alpha_T1_Score15, CFT_T1_CFT1, CFT_T1_CFT2
4.06.12	ELFE_T2_ELFESentence
4.06.12	Eng_T1_ScoreListening, and as a result also Eng_T1_ScoreTotal
4.06.12	MLAT_T2_...
4.08.08	Computer_T1_FDS...
4.11.04	Computer_T1_BDS...; technical glitch possible
4.11.04	Computer_T1_Corsi...; technical glitch possible
4.11.04	Computer_T1_FDS...; technical glitch possible
4.12.18	MLAT_T1_...
4.13.15	Computer_T1_Corsi...; technical glitch possible
4.18.10	MLAT_T1_...
4.18.18	GEFT_T1_...
4.24.05	MLAT_T1_...
4.25.08	Computer_T1_BDS...; technical glitch possible
4.30.07	MLAT_T1_...

Part IV

Predictive modelling

Chapter 11

Modelling strategy

This chapter outlines how we went about building the predictive models. As their name suggests, the goal of these models is primarily prediction: How can the information available at the first data collection best be marshalled in order to make an educated guess about a student's performance on the English test at the third data collection?

It is, of course, also possible to fit models that predict a student's performance on the English test at the second data collection using information available at the first or that predict their performance at the third data collection using information available at the first and second. Indeed, in preliminary analyses we also fitted such models. However, it seems to us that predicting T3 performance using T1 information has the greatest potential value so the following only focuses on this goal.

In terms of building predictive models, there are many ways to skin a cat, and different models (with different predictors or architectures) may have comparable predictive utility; see Breiman (2001).

Our modelling strategy consisted of the following steps:

1. Split up the dataset into a training and a test set.

   The **training set** was used for trying out different models and for gauging the strength of these different models. The predictive strength of the models was one important factor that was considered when selecting the final model; the cost and effort involved in collecting the required predictor variables and the model's ease of use were other considerations (e.g., a linear regression model can easily be written down as an equation or programmed in a spreadsheet; a random forest can't).

   The **test set** was used for validating the selected model. Modelling decisions (including how to select and transform predictors, how to deal with missing data, and how to specify the model) were not affected by the test set data. See Chapter 12.

2. Compute scores for constructs such as intrinsic motivation and locus of control.

   As shown in Chapter 13, so-called ‘optimally-weighted’ construct scores derived from a confirmatory factor analysis based on the training set were all strongly correlated ($r > 0.90$) with scores for which all items corresponding to a construct were weighted equally. Since construct scores derived from a factor analysis depend not only on a participant’s own responses but also on the responses of other participants (this is how the weights are estimated), their computation should be part of the cross-validation (see below). This would have added significantly to the modelling effort, and the strong correlations between the optimally- and equally-weighted construct scores suggested that there was little to be gained from doing so. Therefore only the **equally-weighted construct scores** were used when building models. The files `laps2_full_dataset.csv` and `construct_scores.csv` contain these equally-weighted construct scores for all participants.

3. Exclude students that are not of interest for the present research question. These are students whose native language is English or who were exempted from English classes.

4. Reduce the number of predictors. We removed predictor variables with little variance in the training set. Furthermore, when a construct score was available, item-level responses were not used as possible predictors. Lastly, we removed some predictors showing very strong intercorrelations with others in the training set.

5. Impute missing data. Missing values in the predictor variables that were retained were imputed using the median of the available values of the same variable. More sophisticated imputation strategies exist, e.g., the nearest-neighbour approach (see Kuhn & Johnson, 2013, Section 3.4). In the end, we settled on **median imputation** because it would make the selected model easier to use in class settings: The median values of the predictors can easily be listed, whereas you need specialised software to use the nearest-neighbour algorithm. Moreover, in cross-validation, median imputation performed at least as well as nearest-neighbour imputation.

6. Fit models and cross-validate them in the training set. In order to gauge these models’ predictive strength without turning to the test set, cross-validation was applied. This is a technique that essentially mimicks the partitioning of the overall dataset into a training and test set; see Chapter 12 for details.

   We fitted a whole family of models:

   (a) First, we fitted a **‘no-costs spared’ model**. All available T1 information, from all possible sources, was allowed to enter into this model, without regard to how difficult or costly it was to collect this information. To arrive at the final model in this category, a host of models were fitted on the training data. These included multiple linear regression, robust regression, ridge regression, elastic net, multivariate adaptive regression splines, generalised additive models, partial least squares regression, k-nearest neighbours, regression trees, random forests, support vector machines, stochastic gradient boosting, and Cubist. We do not discuss the architecture of all these models here (see Kuhn & Johnson, 2013, Chapters 5–8); in the end, a multiple linear regression with a limited number of predictors compared favourably to the alternatives. The performance of the more complex models in cross-validation can be consulted in the online materials.

(b) Second, we fitted two simple **baseline models** so that we could get a sense of how much better the **'no-costs spared' model** actually performed in cross-validation. The first baseline model was a **'no predictor' model**, which predicted each unseen data point to be equal to the mean of the seen data points. The second was an **'English-only' model**, which only contained the participants' T1 English test score as the predictor of their T3 English test score.

(c) Third, we fitted a few **'cheap' models**, the input data to which could be collected within a single class hour. These models could potentially be applied in classroom settings. These contained as predictors information that we assumed a class' teacher would already have at their disposal and didn't have to be collected (viz., the variables `AdditionalSupport_T1`, `Grade`, and `L1German`) but that wouldn't lead to discrimination based on sex (hence no `CQEng_T1_Sex`) or possibly socio-economic background (hence no `Gemeindetyp`). Additionally, these models contained either

- the participants' T1 English score;
- the participants' T1 ELFE scores (`ELFE_T1_ELFESentence`, `ELFE_T1_ELFEText`, `ELFE_T1_ELFEWord`, and `ELFE_T1_SentencePerMinute`);
- the participants' T1 questionnaire-based construct scores (equally-weighted); or
- the participants' T1 questionnaire-based construct scores (equally-weighted) and their T1 PLAB score.

7. Select the final models. The final 'no-costs spared' and the final 'cheap' models were decided on by the whole research team based on the candidate models' likely predictive strength (estimated by cross-validation) and the costs involved in obtaining the predictor information required.

8. Assess the predictive strength of the final model using the test set. The final model was refitted on all of the training data and its predictive strength was then tested on the test set. Importantly, the model's parameters and settings were not reestimated or tweaked using the test set.

Chapter 12

Data partition and cross-validation

12.1 Training and test sets

The analyses in this project were to a large part be exploratory. Exploratory analyses entail the substantial risk that the models tightly fit the dataset analysed but does not generalise well beyond it. To offset this risk, we partitioned the dataset into a training set and a test set (see Kuhn & Johnson, 2013, Section 4.3).

The training set was used to conduct all exploratory analyses and to decide on such matters as data transformations, the calculation of construct scores, missing data imputation, model specification — in a nutshell, any step in the analysis that requires the analyst to take a decision. Once a suitable predictive model was agreed upon, its predictive power was tested on the test set. Crucially, the chosen predictive model was not reestimated using the test set data.

To respect the hierarchical nature of the data (children in classes), the test and training sets were not random subsets of the children in the study, but rather (largely) random subsets of the classes in the study (see Roberts et al., 2017). Specifically, from the 17 grade-4 classes at T1, 5 were selected to comprise the test set: the smallest class (Class 4, with 5 grade-4 pupils at T1) as well as four randomly picked classes. Similarly, from the 19 grade-5 classes at T1, 6 were selected to comprise the test set: the smallest class (also Class 4, with only 1 grade-5 pupil at T1) as well as five randomly picked classes. The remaining 12 grade-4 and 13 grade-5 classes comprised the training set.

Table 12.1: Training and test sets. The number of classes sums to 36 rather than 32 because four classes had pupils from both 4th and 5th grade at T1. Only pupils for whom T3 English sores were available were included in the predictive models; for the final models, we only included participants who also had T1 English scores.

Set	Cohort	Classes	with English T3	with English T1 and T3
Training set	4th grade at T1	12	169	154
	5th grade at T1	13	187	177
Test set	4th grade at T1	5	70	65
	5th grade at T1	6	85	80

12.2 Cross-validation[1]

When trying out different models on the training data, we used cross-validation to estimate how well the models would work for new data. This was done to ensure that overzealous data exploration and model fine-tuning would not result in a model that fits the training data well but stands little chance of predicting the test data (see Kuhn & Johnson, 2013; Yarkoni & Westfall, 2017). In cross-validation, the training data is split up into a number (k) of folds, and models are fitted on $k - 1$ folds and then used to predict the outcome in the remaining fold. This process is repeated k times, each time leaving out a different fold. The result are k estimates of the models' predictive accuracy on data not used for fitting the model that can then be averaged.

To account for the dependency structure in the data (students in classes), block cross-validation was used (Roberts et al., 2017): rather than constructing the folds randomly, each of the 22 classes in the training data was used 21 times in its entirety for training and once for prediction. This way, the students in each predicted fold were all part of a different class from the students in the other 21 folds.

Figure 12.1 illustrates the principles behind the partitioning of the data and block cross-validation.

[1] This section is adapted from Vanhove et al. (2019).

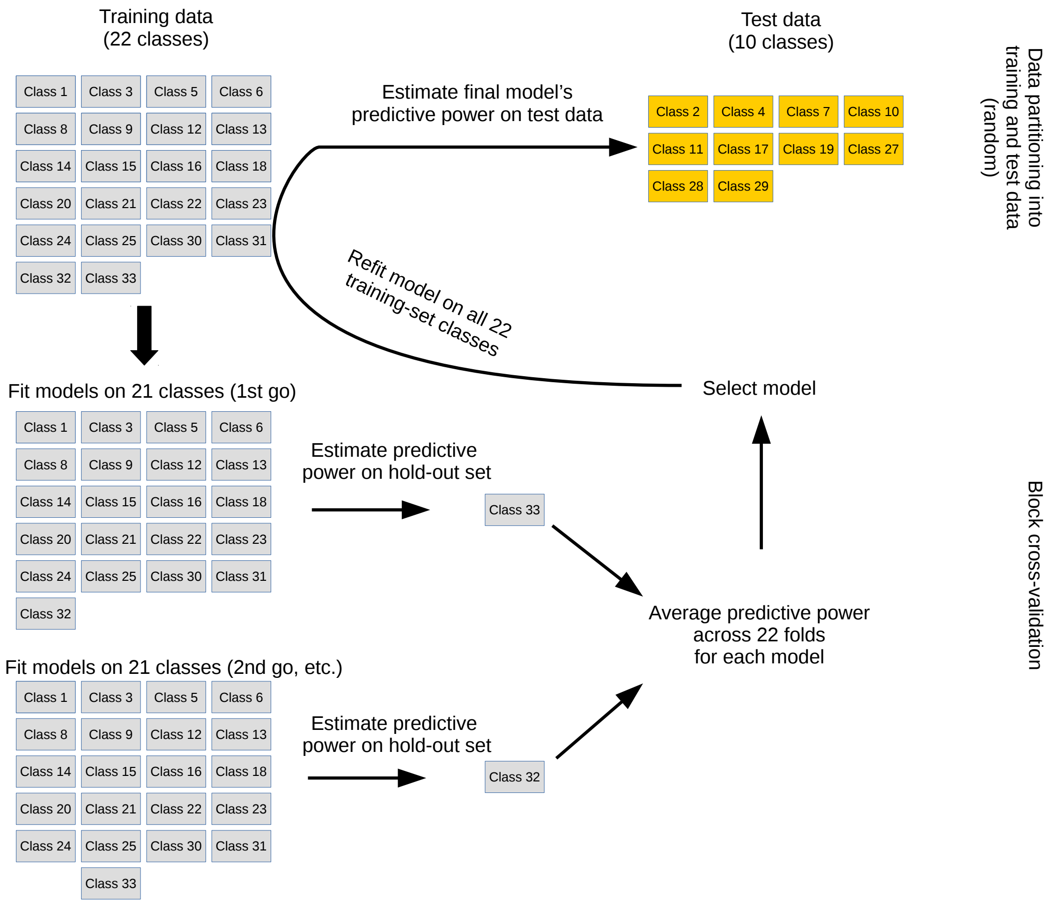
Training data
(22 classes)
Class 1 Class 3 Class 5 Class 6
Class 8 Class 9 Class 12 Class 13
Class 14 Class 15 Class 16 Class 18
Class 20 Class 21 Class 22 Class 23
Class 24 Class 25 Class 30 Class 31
Class 32 Class 33
Test data
(10 classes)
Class 2 Class 4 Class 7 Class 10
Class 11 Class 17 Class 19 Class 27
Class 28 Class 29
Data partitioning into training and test data (random)
Estimate final model's predictive power on test data
Refit model on all 22 training-set classes
Select model
Fit models on 21 classes (1st go)
Class 1 Class 3 Class 5 Class 6
Class 8 Class 9 Class 12 Class 13
Class 14 Class 15 Class 16 Class 18
Class 20 Class 21 Class 22 Class 23
Class 24 Class 25 Class 30 Class 31
Class 32
Estimate predictive power on hold-out set
Class 33
Average predictive power across 22 folds for each model
Block cross-validation
Fit models on 21 classes (2nd go, etc.)
Class 1 Class 3 Class 5 Class 6
Class 8 Class 9 Class 12 Class 13
Class 14 Class 15 Class 16 Class 18
Class 20 Class 21 Class 22 Class 23
Class 24 Class 25 Class 30 Class 31
Class 33
Estimate predictive power on hold-out set
Class 32

Figure 12.1: Illustration of how the data were partitioned into a training and a test set and of how block cross-validation works. Only two iterations of block cross-validation are shown; in reality, 22 took place for each model, each time leaving out a different class. Figure based on Figure 3 in Vanhove et al. (2019).

Chapter 13

Construct scores and their reliability

For many of the constructs that we were interested in, the dataset contains the participants' responses at the item level. For the analyses, however, it was useful to summarise the participants' responses per construct in a single number per participant. In essence, there are two ways for computing such construct scores:

- weighting each response equally. In other words, a participant's performance on a task is the sum or mean of their scores on each individual item.
- weighting responses differentially (or 'optimally'). This can be achieved by fitting a factor analysis on the item-level responses and extracting the participants' factor scores.

The advantages of weighting each response equally are simplicity and the fact that the weights used are independent of the data one has at one's disposal: had the data looked differently, the weights used to compute the construct scores would still have been the same. However, by using equal weights, the reliability of construct score can be negatively affected by poorly functioning items. By contrast, poorly functioning items do not affect the reliability of differentially ('optimally')-weighted scales since their factor loadings will be close to zero. However, the weights are derived from the dataset itself. As a result, had the data looked differently, the weights used to compute the construct scores would have been different, too. This data-dependence should be taken into account if the factor loadings extracted from the factor analysis are used in follow-up analyses: Ideally, you would want to propagate the uncertainty about the factor scores in the follow-up analyses (see Houslay & Wilson, 2017, for a similar point, if not specific to factor analysis). This, however, is not easy to do. Moreover, since the factor loadings and hence the factor score are data-dependent, their computation should be part of the cross-validation scheme, adding another layer of complexity to the analysis.

Because of their conceptual and statistical ease of use, we preferred equally-weighted scales wherever they seemed reasonable. In the following sections, the reliability of these scales will be assessed by the commonly-used Cronbach's α as well as by Revelle's (2019) ω_T (ω_{RT}); see McNeish (2018) for an introduction. Both α and ω_{RT} were computed using the `psych` package for `R` (Revelle, 2018).

Where differentially/optimally-weighted scales also seemed reasonable, their reliability was assessed using Hancock and Mueller's (2001) coefficient H. H was computed as follows (see McNeish, 2018, Equation 6):

$$H = \left(1 + \left(\sum_{i=1}^{k} \frac{l_i^2}{1 - l_i^2}\right)^{-1}\right)^{-1} \tag{13.1}$$

where k is the number of items and l_i is the standardised factor loading for the ith item. We derived these factor loadings (as well as the factor scores as construct scores) by fitting a confirmatory factor analysis using the `lavaan` package (version 0.6-3) for `R` (Rosseel, 2012).

The reliabilities and, if applicable, the factor loadings were computed solely on the basis of the training set. The factor scores for the participants in the test set were then derived from these factor loadings; in other words, the test data were not used in determining the factor solution.

The training set reliabilities of the construct scores are summarised in Table 13.4 at the end of this chapter.

13.1 English proficiency

At T1, the overall construct score (`Eng_T1_ScoreTotal`) was automatically output by the test software as the mean between the two subtask scores (`Eng_T1_ScoreListening` and `Eng_T1_ScoreUse`). These two subtask scores were correlated at $r = 0.64$ ($n = 422$) in the training data.

At T2 and T3, English proficiency was tested using five C-tests. The construct scores were computed as the mean of the five test scores. The C-tests were scored twice: once without penalising spelling errors and once penalising spelling errors.

13.2 Language aptitude

13.2.1 MLAT

The MLAT was scored by tallying the number of correct answers per participant.

13.2.2 PLAB

The PLAB was scored by tallying the number of correct answers per participant.

One participant in the training set (5.33.17) only responded to the first 4 out of 15 items at T1 (`NA` for the remaining 11). This participant's total score was treated as `NA`.

13.3 GEFT

The GEFT was scored by tallying the number of correct answers per participant.

13.4 Child questionnaire data

13.4.1 Motivation

The construct scores for the ten (at T1) motivational constructs (see Table 5.1 on page 24) were computed in two ways.

First, mean scores were computed for which each item that was considered to be tapping into the construct was weighted equally. (The answers to all questions could take values from 1 to 4.) If a child did not respond to all items, only the responses given were used to compute the mean score. If a child did not respond to any of the items subsumed under a construct, no mean score could be computed for it.

Second, 'optimally'-weighted factor scores were computed by fitting all items in a confirmatory factor analysis with ten/eleven latent constructs. This yielded factor scores different from what fitting a separate factor analysis for each latent construct would have done. One advantage of fitting one instead of ten/eleven different factor analyses is that factor scores can still be estimated even for constructs for which a participant did not provide any answers based on the intercorrelations between the latent constructs. In the factor analysis, the responses to each item were treated as ordinal variables.

The factor analysis for T1 was fitted as follows using the `lavaan` package for `R`:

```
mot_t1.mod <- '
  extrinsic_school =~ CQEng_T1_FB04 + CQEng_T1_FB05 + CQEng_T1_FB08
  extrinsic_leisure =~ CQEng_T1_FB09 + CQEng_T1_FB10 + CQEng_T1_FB11
  intrinsic =~ CQEng_T1_FB01 + CQEng_T1_FB06 + CQEng_T1_FB13 +
              CQEng_T1_FB14
  lingua_franca =~ CQEng_T1_FB02 + CQEng_T1_FB03 + CQEng_T1_FB07 +
              CQEng_T1_FB12
  anxiety =~ CQEng_T1_FB19 + CQEng_T1_FB20 + CQEng_T1_FB21 +
            CQEng_T1_FB22 + CQEng_T1_FB23
  selfconcept_eng =~ CQEng_T1_FB24 + CQEng_T1_FB25 + CQEng_T1_FB26
  selfconcept_ger =~ CQEng_T1_FB27 + CQEng_T1_FB28 + CQEng_T1_FB29 +
                    CQEng_T1_FB30
  parental_encouragement =~ CQEng_T1_FB36 + CQEng_T1_FB37 +
                            CQEng_T1_FB38 + CQEng_T1_FB39 +
                            CQEng_T1_FB40
  teacher_motivation =~ CQEng_T1_FB31 + CQEng_T1_FB32 + CQEng_T1_FB33 +
                        CQEng_T1_FB34 + CQEng_T1_FB35
  dedication =~ CQEng_T1_FB15 + CQEng_T1_FB16 + CQEng_T1_FB17 +
                CQEng_T1_FB18
'
# missing = "pairwise" to deal with missing values in ordinal variables
mot_t1.fit <- cfa(mot_t1.mod, data = mot_t1, missing = "pairwise")
```

Table 13.1 shows the raw and standardised factor loading for the confirmatory factor analysis for the T1 motivational data. Figure 13.1 shows the relationship between the (optimally-weighted) factor scores for the motivational constructs and their corresponding (equally-weighted) mean scores. With correlations ranging between $0.90 < r < 0.99$, the added value of factor scores over mean scores can be expected to be minimal.

The motivational data were collected at T1, T2 and T3. However, the T2 and T3 data will not be used in building predictive models so that no construct scores for T2 or T3 are computed at this point. Since the motivational data with respect to French will not be used in the predictive models either, the same goes for these construct scores.

Table 13.1: Raw and standardised factor loadings for the confirmatory factor analysis fitted on the motivation questionnaire (children's questionnaire, T1, training set only). Item 13 was already recoded such that higher values reflected more intrinsic motivation. All items were coded as ordinal variables.

Construct	Item	Raw loading	Standardised loading
Extrinsic motivation: school	4	1.000	.872
	5	0.851	.742
	8	0.994	.867
Extrinsic motivation: leisure	9	1.000	.712
	10	1.114	.794
	11	0.934	.666
Intrinsic motivation	1	1.000	.755
	6	0.760	.574
	13	0.576	.435
	14	1.142	.862
Lingua franca	2	1.000	.601
	3	1.213	.729
	7	1.196	.719
	12	1.322	.795
Foreign-language anxiety	19	1.000	.643
	20	1.343	.863
	21	1.172	.753
	22	1.213	.780
	23	1.348	.867
Self concept English	24	1.000	.844
	25	1.106	.933
	26	1.008	.851
Self concept German	27	1.000	.872
	28	0.930	.811
	29	0.873	.761
	30	0.995	.868
Parental encouragement	36	1.000	.700
	37	1.171	.824
	38	1.087	.756
	39	0.921	.649
	40	0.953	.667
Teacher motivation	31	1.000	.841
	32	0.952	.800
	33	0.994	.835
	34	0.897	.754
	35	0.691	.581
Dedication	15	1.000	.783
	16	1.000	.784
	17	0.903	.707
	18	1.011	.792

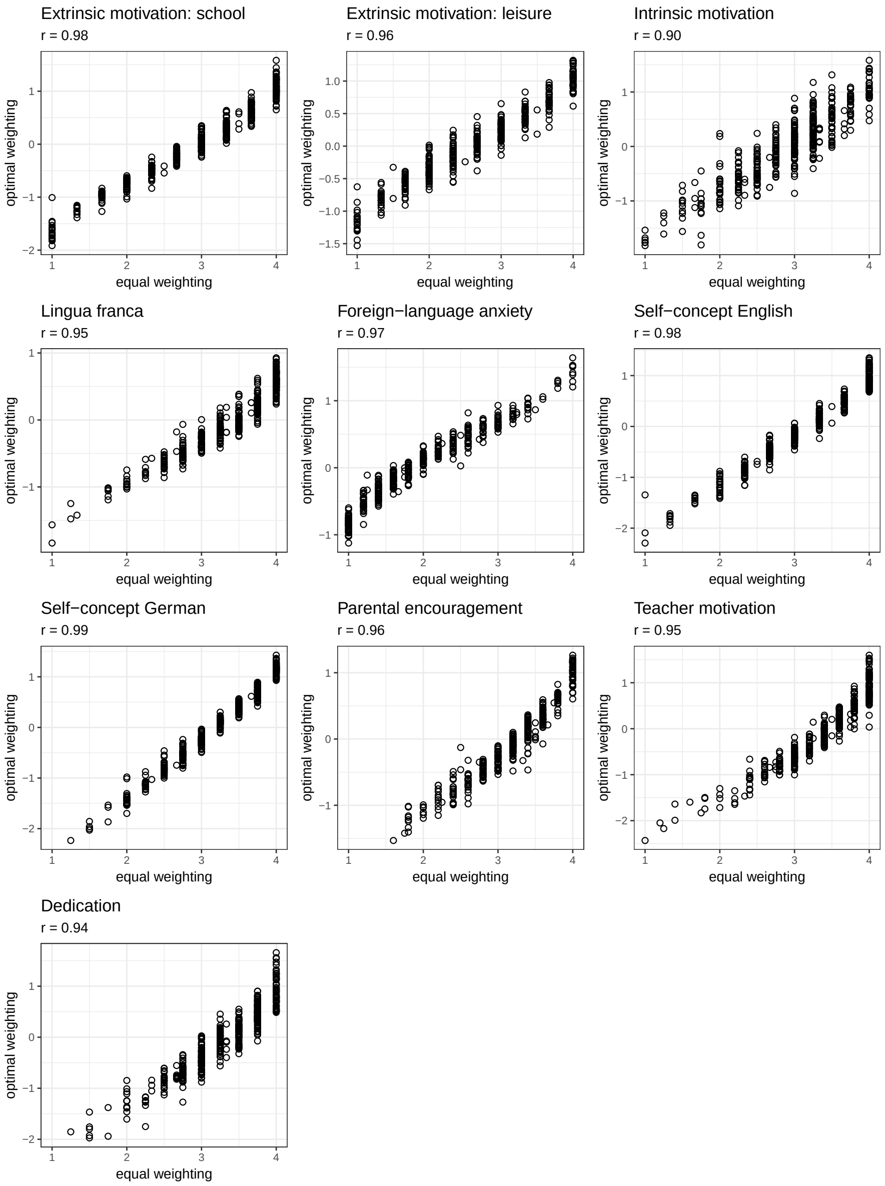

Figure 13.1: Correlations between the (equally-weighted) mean scores and the (optimally-weighted) factor scores for the motivational constructs in the training data at T1.

13.4.2 Locus of control

Since the reliability for a (equally-weighted) mean construct score on this questionnaire was low ($\alpha = .58$, $\omega_{RT} = .63$), the locus of control data were also fitted in a confirmatory factor analysis. The factor scores derived from this analysis are assumed to reflect the participants' external locus of control (i.e., the higher, the more external locus of control). The factor loadings are shown in Table 13.2. As Figure 13.2 shows, there is a strong correlation between the equally-weighted and optimally-weighted LOC construct scores. The locus of control data were only collected at T1.

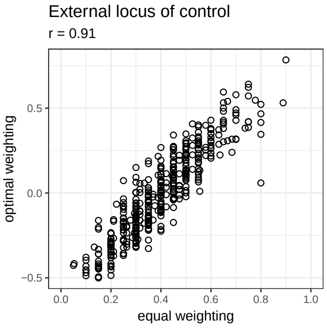

Figure 13.2: Correlation between the (equally-weighted) mean scores and the (optimally-weighted) factor scores for the locus of control construct in the training data at T1.

13.5 Parental questionnaire: Socio-economic status

We attempted to capture the participants' socio-economic status using eight questions on the parental questionnaire. Since these items had different numbers of possible (ordinal) responses, these data were fitted in a confirmatory factor analysis, the loadings of which are shown in Table 13.3.

Additionally, we computed (equally-weighted) mean scores. To this end, the responses to each item were recoded so that they fell in the 0–1 interval, where 0 represents the lowest possible answer and 1 the highest. Thus, for questions with 4 response options, the answers were recoded as 0, 0.33, 0.67 and 1; for questions with 5 response options, they were recoded as 0, 0.25, 0.50, 0.75 and 1, etc. As Figure 13.3 shows, the mean scores thus computed are strongly correlated with the factor scores.

Table 13.2: Raw and standardised factor loadings for the confirmatory factor analysis fitted on the locus of control data (children's questionnaire, T1, training set only). Items 3, 13 and 19 were already recoded so that higher values (i.e., 1 rather than 0) reflected a more external locus of control. All items were coded as binary variables.

Item	Raw loading	Standardised loading
1	1.000	.318
2	1.840	.586
3	-0.144	-.046
4	-0.635	-.202
5	1.341	.427
6	1.660	.529
7	1.132	.360
8	1.250	.398
9	1.087	.346
10	0.330	.105
11	1.543	.491
12	0.919	.293
13	0.378	.120
14	1.639	.522
15	0.779	.248
16	1.514	.482
17	0.637	.203
18	2.186	.696
19	0.272	.086
20	1.761	.561

Table 13.3: Raw and standardised factor loadings for the confirmatory factor analysis fitted on the SES data (parental questionnaire, T1, training set only). All items were coded as ordinal variables.

Item	Raw loading	Standardised loading
`PQ_T1_Earning`	1.000	.804
`PQ_T1_EducationFather`	0.914	.735
`PQ_T1_EducationMother`	0.801	.643
`PQ_T1_Holidays`	1.156	.929
`PQ_T1_MedicalCare`	1.151	.925
`PQ_T1_MonthlyBills`	1.131	.909
`PQ_T1_Saving`	1.126	.904
`PQ_T1_NrBooks`	0.865	.695

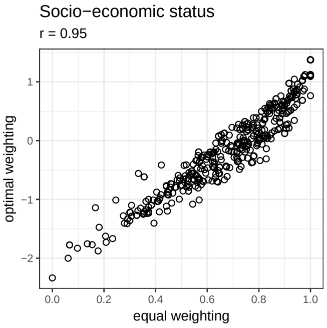

Figure 13.3: Correlation between the (equally-weighted) mean scores and the (optimally-weighted) factor scores for the socio-economic status construct in the training data at T1.

Table 13.4: Reliabilities of the construct scores. These reliabilities were computed on the basis of the training set data only. n: number of data points; if there is a number between brackets, it refers to the number of data points for which equally-weighted mean/sum scores could be computed, whereas the number not between brackets refers to the number of data points for which optimally-weighted factor scores could be computed. α: Cronbach's alpha. ω_{RT}: Revelle's omega total. H: coefficient H; only computed when optimally-weighted factor scores were also computed.

Task or construct	Data collection	n	α	ω_{RT}	H
MLAT	T1	419	.88	.90	
PLAB	T1	424	.69	.75	
GEFT	T1	422	.83	.86	
English C-tests, without spelling	T3	397	.94	.95	
English C-tests, with spelling	T3	397	.94	.95	
Locus of control	T1	418	.58	.63	.81
Extrinsic motivation: school	T1	427 (426)	.81	.82	.88
Extrinsic motivation: leisure	T1	427 (426)	.71	.71	.78
Intrinsic motivation	T1	427 (427)	.67	.73	.83
Lingua franca	T1	427 (427)	.73	.75	.82
Foreign-language anxiety	T1	427 (422)	.85	.86	.90
Self-concept English	T1	427 (422)	.85	.85	.92
Self-concept German	T1	427 (413)	.84	.86	.91
Parental encouragement	T1	427 (411)	.71	.80	.85
Teacher motivation	T1	427 (411)	.82	.85	.89
Dedication	T1	427 (423)	.77	.81	.85
Socio-economic status	T1	443	.89	.93	.96

Chapter 14

Selection of participants

We removed from the dataset children who were exempted from English classes (variable `AddInfo_DispEN`), native speakers of English (variable `L1English`) and participants 4.14.06 (has lived in Canada), 4.18.09 (born in US; went to a bilingual school), 5.21.19 (English may be spoken in the household), and 5.22.17 (English may be one of the native languages, though the parents did not declare it as such).

Pupils who did not take the T3 English test were excluded from the analyses.

Chapter 15

Metrics of model performance

The root mean squared error (RMSE) was used to adjudicate between different models. It is defined as follows

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15.1}$$

where n is the number of out-of-fold cases (in cross-validation) or the number of test set cases (in the final validation), y_i the ith observed outcome value and, $\hat{y}_i$ the corresponding ith predicted outcome value.

The RMSE can be interpreted as being roughly – but not quite – the average difference between a model's predictions and the observed values. (In the same way that a standard deviation can be interpreted as being roughly – but not quite – the average difference between the observations and their mean.) The interpretation of the mean absolute error (MAE) is simpler: it *is* the average (mean) difference between a model's prediction and the observed values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \tag{15.2}$$

Many readers will be more familiar with the R^2 metric of (so-called) 'explained' variance. Some problems that beset R^2 are discussed at https://janhove.github.io/analysis/2016/04/22/r-squared, but perhaps most important of all is that R^2, as it is traditionally computed, does *not* estimate how well the model itself would capture the variance in a new sample. Instead, it estimates (at best) how well a *newly estimated* model would capture the variance in a new sample.

However, there exist different ways of computing R^2 (Kvålseth, 1985).[1] For ordinary regression models, these all yield the same result. However, when the model is used to predict observations that were not used when fitting the model, they do not. One popular method for computing R^2 (and in fact the default in the `caret` package) is to square the correlation between the predicted and observed values. This is problematic since the correlation between predicted and observed values can be excellent even if the former correspond poorly to the latter (e.g., the values 1, 2, 3 correlate perfectly with the values 2000, 4000, 6000 but correspond poorly to them). We therefore computed R^2 as the proportional decrease in the residual sum of squares relative to a baseline model without any predictors (hence R^2_{RSS}):

$$R^2_{RSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \hat{y_0})^2} \tag{15.3}$$

where $\hat{y_0}$ is the predicted outcome value by a baseline model with only an intercept. Such a model predicts each new observation to be equal to the mean of the training data.

[1] This paragraph is adapted from Vanhove et al. (2019, Note 6).

Chapter 16

Predictive modelling of T3 data

16.1 Dataset

The training set for T3 comprised 169 4th-graders and 187 5th-graders. The test set comprised 70 4th-graders and 85 5th-graders.

16.2 Outcome variable

The T3 English test scores for which spelling errors were or were not penalised were highly correlated in the training set, see Figure 16.1; we only analysed the test scores for which spelling errors were not penalised.

16.3 Selection of predictors

Only variables that were available at T1 served as predictor variables.

Item-level responses that formed part of a construct score were not retained as predictors. For instance, `CQEng_T1_FB01` and `GEFT_T1_GEFT18` were not used as predictor variables because they formed part of the construct scores for intrinsic motivation and the GEFT, respectively.

Several variables were discarded because they hardly contained any variance. This pertains to all `PM_T1_` variables. A related reason for excluding predictor variables was the sheer number of different categorical values, with few observations for the majority of values (e.g., `PQ_T1_CountryFather`).

The following categorical variables were retained and recoded as numeric (dummy) variables:

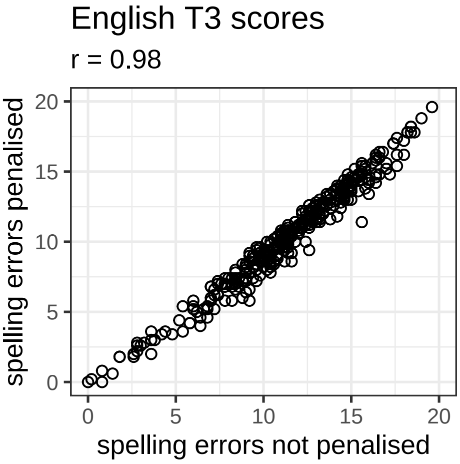

Figure 16.1: T3 English test scores that were and were not penalised for spelling errors; training set data only.

- Grade;
- Gemeindetyp, recoded as four binary dummy variables (Class_RegionalCentre, Class_Suburban, Class_HighIncome, Class_Periurban);
- CQEng_T1_Sex, recoded as a binary dummy variable (Girl);
- L1German;
- Multilingual.

The T1 English variable selected was the total score; the subscores for the listening and use subtests were not used as additional predictors.

Figures 16.2 and 16.3 show the (equally-weighted) construct scores for the questionnaire-based constructs, their intercorrelations as well as their correlations with the overall English test scores at T1 and T3 in the training data. There are no strong intercorrelations between the predictors nor any large outlying values.

As for the cognitive predictors (for want of a better term), CFT_T1_Total was chosen as the CFT predictor, Computer_T1_BDSTotalCorrect as the predictor derived from the backward digit task, Computer_T1_CorsiMemorySpan as the predictor derived from the Corsi block task, and Computer_T1_FDSTotalCorrect as the predictor derived from the forward digit task. As Figure 16.4 shows, these predictors are not strongly collinear with one another nor do they show large outliers.

Figure 16.5 shows the ELFE measures. The total ELFE score was not included as a predictor because of its strong intercorrelations with the other ELFE measures.

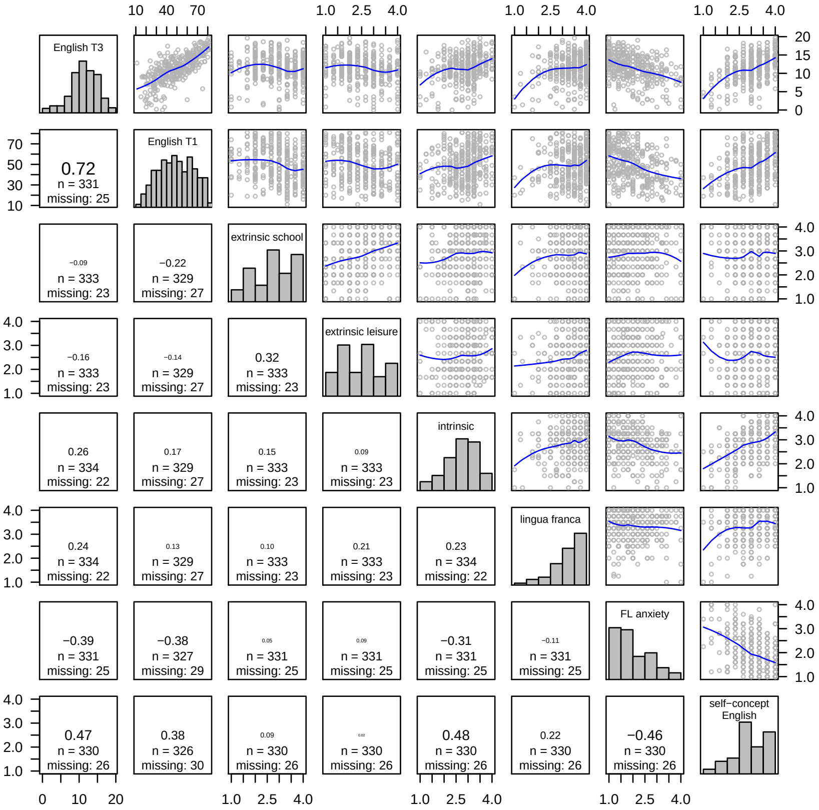

Figure 16.2: Equally-weighted construct scores for questionnaire-based constructs at T1, their intercorrelations and their correlations with the overall English scores at T1 and T3 in the training data. The lower triangle shows Pearson correlation coefficients. (Part 1)

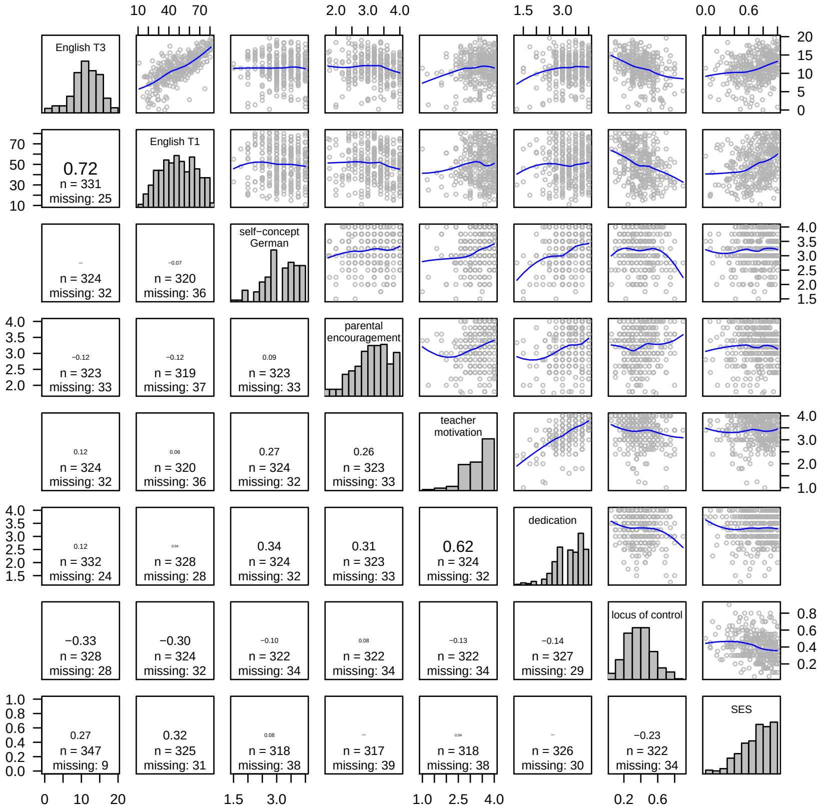

Figure 16.3: Equally-weighted construct scores for questionnaire-based constructs at T1, their intercorrelations and their correlations with the overall English scores at T1 and T3 in the training data. The lower triangle shows Pearson correlation coefficients. (Part 2)

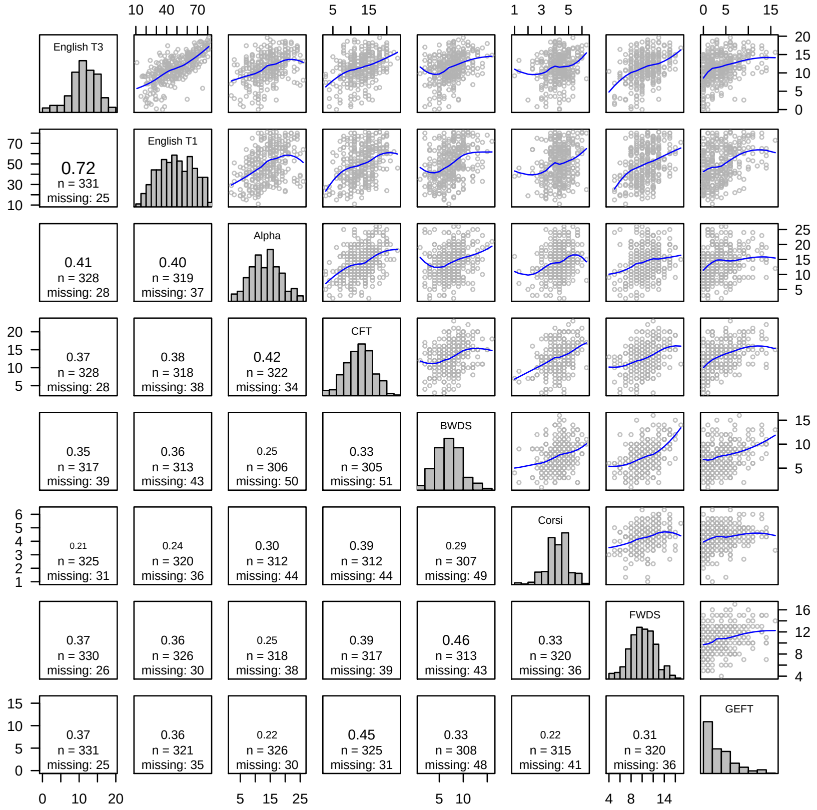

Figure 16.4: The 'cognitive' predictors at T1, their intercorrelations and their correlations with the overall English scores at T1 and T3 in the training data. The lower triangle shows Pearson correlation coefficients.

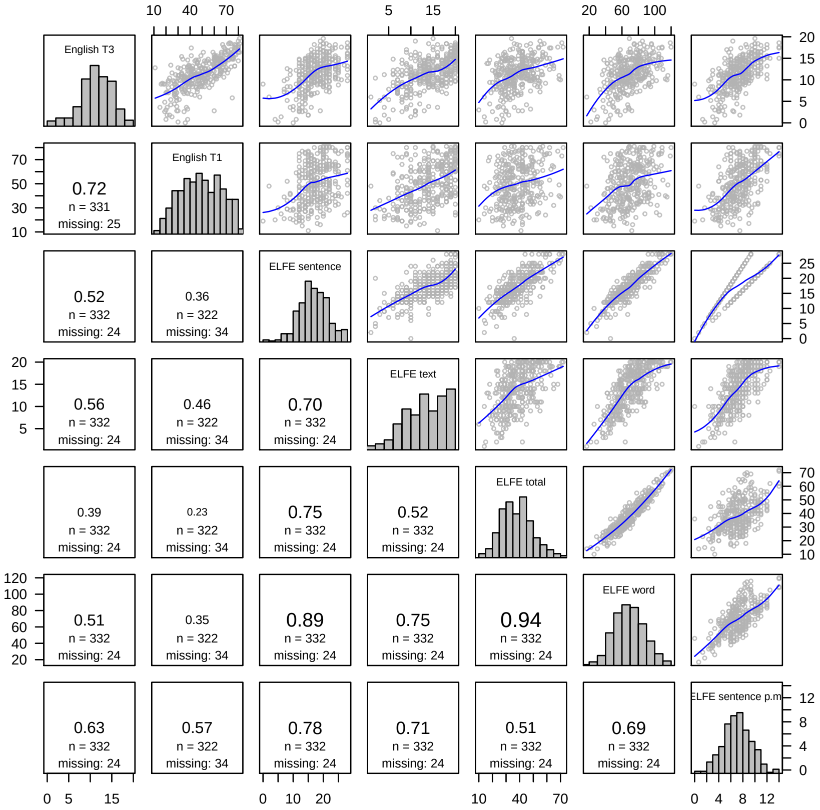

Figure 16.5: The ELFE measures at T1, their intercorrelations and their correlations with the overall English scores at T1 and T3 in the training data. The lower triangle shows Pearson correlation coefficients. The ELFE total score was discarded because of its strong intercorrelations with the other ELFE measures.

Figure 16.6 shows the three aptitude measures. They are not strongly collinear with one another nor do they show large outliers.

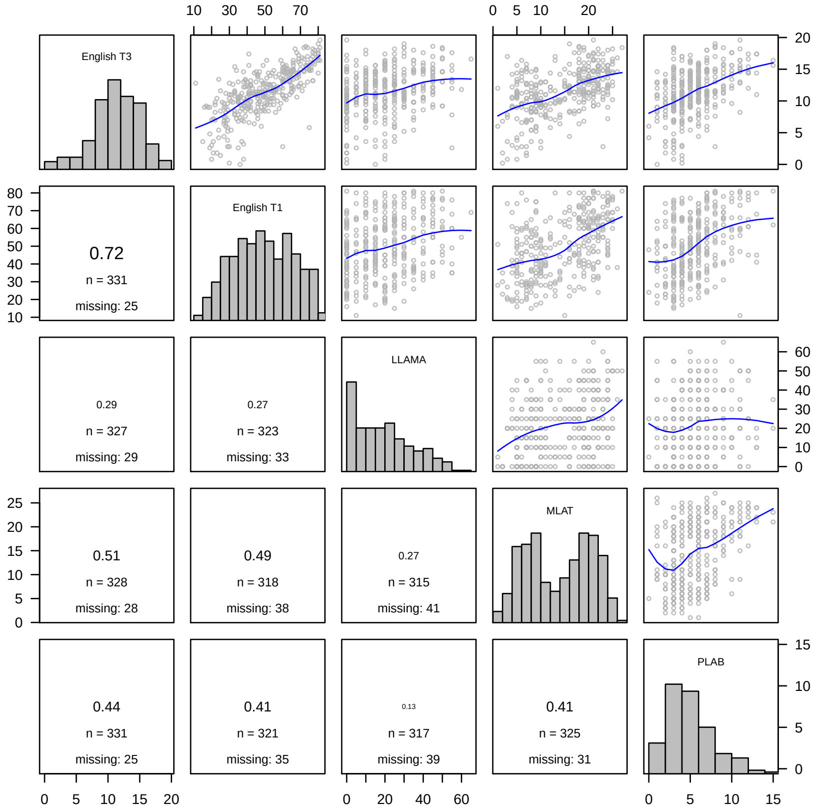

Figure 16.6: The aptitude measures at T1, their intercorrelations and their correlations with the overall English scores at T1 and T3 in the training data. The lower triangle shows Pearson correlation coefficients.

In all, 35 predictors were retained.

16.4 Selection of the 'no costs spared' model

As outlined in Chapter 11, a host of models were fitted and tuned on the training data. However, a multiple linear model with only a handful of predictors and no interactions performed roughly on par with the more complex approaches in cross-validation. **When fitting the final model, we only took into account participants who had T1**

Table 16.1: Multiple linear regression model for predicting T3 English scores. Missing predictor data were imputed using median imputation using the full training set data. Median = the predictor's median in the training set (used in imputation). Estimate = the estimated regression coefficient for the predictor. SE = the naïve standard deviation for the estimated regression coefficient; naïve meaning that its computation did not take into account the fact that this model was selected for its performance in cross-validation.

Term	Median	Estimate	SE
Intercept		0.045	1.4
English T1	49	0.093	0.010
Grade at T1	5	−0.55	0.31
Intrinsic motivation	3	0.48	0.22
Self-concept English	3	0.89	0.22
ELFE sentences/minute	7.33	0.49	0.08
MLAT	15	0.047	0.023
PLAB	5	0.16	0.05

English test scores. The model's estimated coeffients are shown in Table 16.1 and partial effect plots are shown in Figure 16.8.[1]

In cross-validation, the linear model with seven predictors reduced the residual sum of squares by about 58% relative to an intercept-only model (i.e., $R^2_{RSS} = .58$, 95% CI: [.49, .66]). Its root mean square error (RMSE) in cross-validation was 2.24 (95% CI: [2.02, 2.47]), and its mean absolute error (MAE) in cross-validation was 1.77 (95% CI: [1.60, 1.95]). For reference, an intercept-only model yielded a RMSE of 3.69 and a MAE of 2.93. For further reference, we also fitted and cross-validated a linear model with a single predictor, viz., the participants' English score at T1. This model yielded $R^2_{RSS} = .42$, RMSE = 2.61 and MAE = 2.03.

When applied to the test set, the linear model with seven predictors reduced the residual sum of squares by about 62% relative to the intercept-only model (i.e., $R^2 = .62$, 95% CI: [.52, .70]). Its root mean square error (RMSE) was 2.32 (95% CI: [2.02, 2.60]) and its mean absolute error (MAE) 1.85 (95% CI: [1.63, 2.08]).

These results are summarised in Figure 16.7.

[1] We want to draw the attention of any reader who wishes to use this model for *understanding* (as opposed to merely *predicting*) foreign-language learning to what Breiman (2001) calls the 'Rashomon effect': While the presented model worked best in cross-validation, a number of models with different predictors fared only slightly worse. Consequently, one would be jumping to conclusions if one said that these seven predictors are important in foreign-language learning and the others are not. See the online materials for cross-validation results of alternative models.

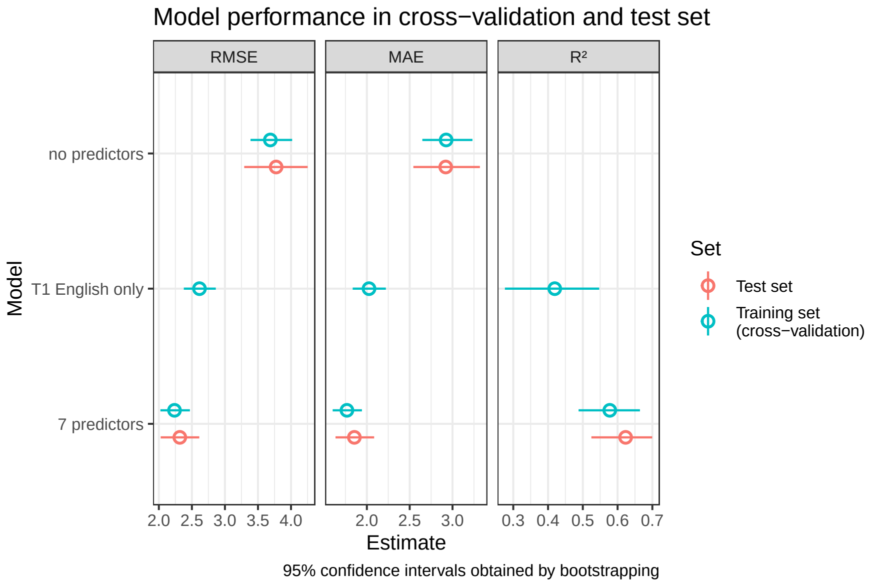

Figure 16.7: Performance of the chosen model relative to two baseline models. The R^2 value for the intercept-only model is not shown as it is 0 by definition. The 95% confidence intervals were obtained by bootstrapping the 22 cross-validation estimates or by bootstrapping the observed and predicted test set values and recomputing the estimates (percentile approach). The English-only model wasn't applied to the test set.

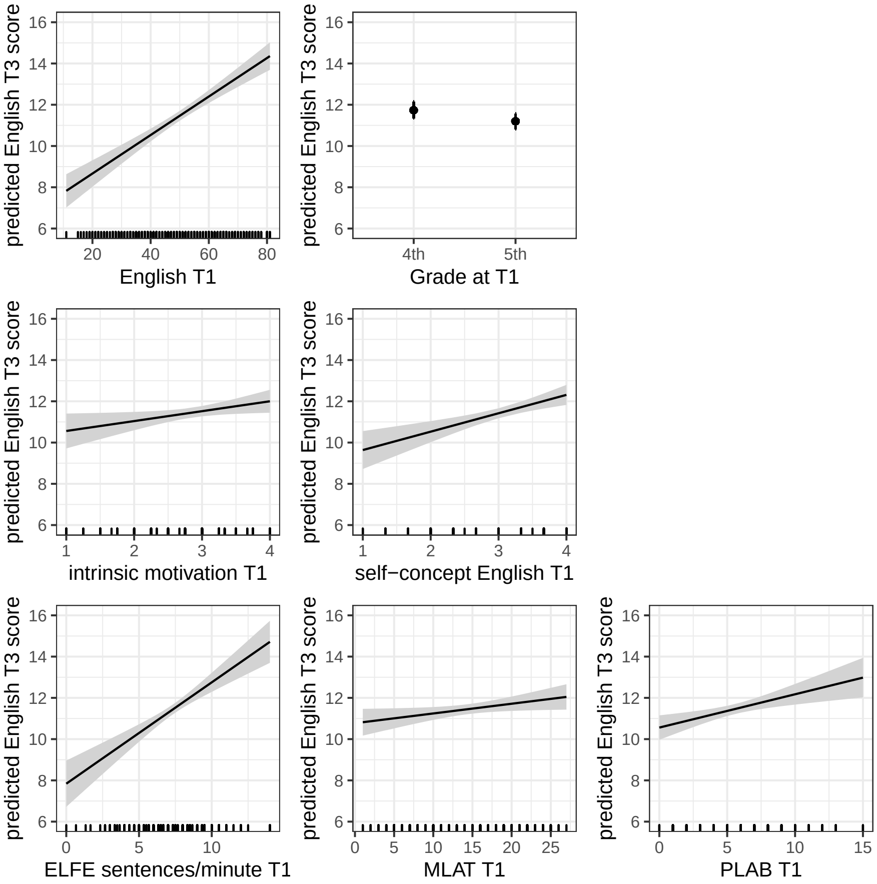

Figure 16.8: Selected model for predicting T3 performance using T1 info. For each effect plot, the six other predictors were centred at their training set mean. Naïve 95% confidence bands are also plotted; naïve meaning that they do not account for the fact that this model was selected for its performance in cross-validation.

Figure 16.9 shows how the model’s predictions compare to the actually observed values in both the training set (out-of-fold predictions) and in the test set.

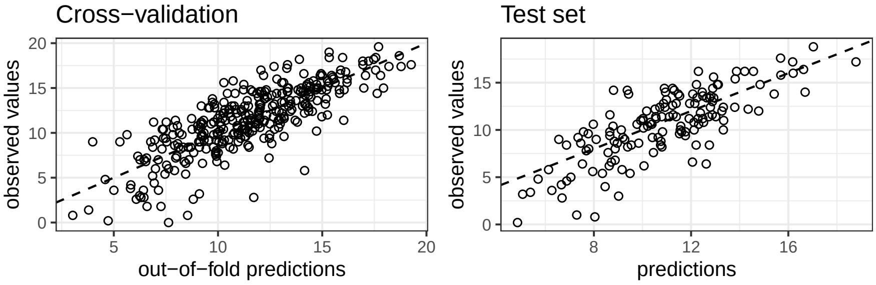

Figure 16.9: Model fit of the ‘no costs spared’ with 7 predictors. Left: Out-of-fold predictions versus actual observations in the training set. Right: Predictions versus actual observations in the test set.

16.5 Selection of the ‘cheap’ models

Four ‘cheap’ models were fitted, see Chapter 11. These were the results in cross-validation.

- English T1 + free variables. RMSE = 2.48, MAE = 1.99, $R^2_{RSS} = .47$.
- ELFE + free variables. RMSE = 2.69, MAE = 2.19, $R^2_{RSS} = .39$.
- Motivation + free variables. RMSE = 2.83, MAE = 2.27, $R^2_{RSS} = .34$.
- Motivation + PLAB + free variables. RMSE = 2.69, MAE = 2.16, $R^2_{RSS} = .40$.

The first and fourth model were then applied to the test set. When applied to the test set, the ‘English’ model had a RMSE of 2.52 (95% CI: [2.15, 2.90]), a MAE of 1.97 (95% CI: [1.72, 2.23]), and a R^2_{RSS} of .55 (95% CI: [.43, .65]). For the ‘Motivation + PLAB’ model: RMSE = 2.82 (95% CI: [2.48, 3.17], MAE = 2.24 (95% CI: [1.97, 2.52], $R^2_{RSS} = .46$ (95% CI: [.30, .59]). See Figure 16.10.

Tables 16.2 and 16.3 list their coefficients.

Figures 16.11 and 16.12 show the fit of these models.

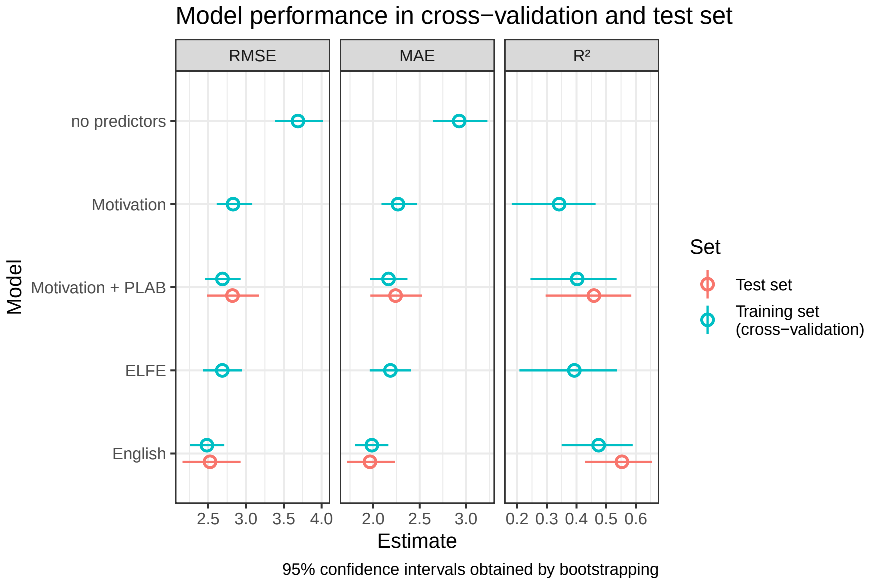

Figure 16.10: Performance of the cheap models. The R^2 value for the intercept-only model is not shown as it is 0 by definition. The 95% confidence intervals were obtained by bootstrapping the 22 cross-validation estimates or by bootstrapping the observed and predicted test set values and recomputing the estimates (percentile approach). Only the 'English' and 'Motivation + PLAB' models were applied to the test set.

Table 16.2: Model for predicting the T3 English results using the T1 English results and three free variables. Estimate = the estimated regression coefficient for the predictor. SE = the naïve standard deviation for the estimated regression coefficient; naïve meaning that its computation did not take into account the fact that this model was selected for its performance in cross-validation.

Term	Estimate	SE
Intercept	2.0	1.3
English T1	0.14	0.010
Grade at T1	0.61	0.33
Additional support?	−2.6	0.4
L1 German?	0.06	0.4

Table 16.3: Model for predicting the T3 English results using the questionnaire-based constructs, PLAB and three free variables. The questionnaire-based constructs are equally-weighted mean scores; see Chapter 13. Estimate = the estimated regression coefficient for the predictor. SE = the naïve standard deviation for the estimated regression coefficient; naïve meaning that its computation did not take into account the fact that this model was selected for its performance in cross-validation.

Term	Estimate	SE
Intercept	−0.44	2.2
Extrinsic motivation, school	0.06	0.20
Extrinsic motivation, leisure	−0.51	0.19
Intrinsic motivation	0.42	0.30
Lingua franca	0.65	0.27
Foreign language anxiety	−0.33	0.22
Self-concept English	1.47	0.28
Self-concept German	−0.53	0.30
Parental encouragement	−0.72	0.30
Teacher motivation	−0.51	0.38
PLAB	0.36	0.06
Grade at T1	1.9	0.3
Additional support?	−2.7	0.4
L1 German?	0.15	0.39

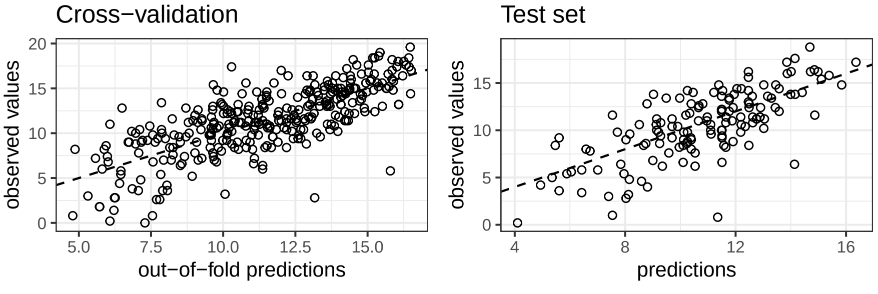

Figure 16.11: Model fit of the ‘cheap English’ model. Left: Out-of-fold predictions versus actual observations in the training set. Right: Predictions versus actual observations in the test set.

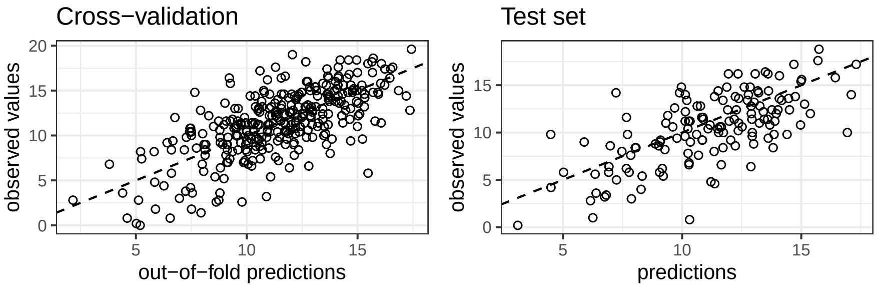

Figure 16.12: Model fit of the 'cheap Motivation + PLAB' model. Left: Out-of-fold predictions versus actual observations in the training set. Right: Predictions versus actual observations in the test set.

Bibliography

Breiman, Leo. 2001. Statistical modeling: The two cultures. *Statistical Science* 16(3). 199–231. doi:10.1214/ss/1009213726.

Houslay, Thomas M. & Alastair J. Wilson. 2017. Avoiding the misuse of BLUP in behavioural ecology. *Behavioral Ecology* 28(4). 948–952. doi:10.1093/beheco/arx023.

Kuhn, Max & Kjell Johnson. 2013. *Applied predictive modeling.* New York: Springer. doi:10.1007/978-1-4614-6849-3.

Kvålseth, Tarald O. 1985. Cautionary note about R^2. *The American Statistician* 4(1). doi:10.2307/2683704.

McNeish, Daniel. 2018. Thanks coefficient alpha, we'll take it from here. *Psychological Methods* 23(3). 412–433. doi:10.1037/met0000144.

Revelle, William. 2018. *psych: Procedures for psychological, psychometric, and personality research.* Northwestern University Evanston, Illinois. https://CRAN.R-project.org/package=psych. R package version 1.8.12.

Revelle, William. 2019. *Using R and the psych package to find ω.* http://personality-project.org/r/psych/HowTo/omega.pdf.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig & Carsten F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8). 913–929. doi:10.1111/ecog.02881.

Rosseel, Yves. 2012. lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48(2). 1–36. http://www.jstatsoft.org/v48/i02/.

Vanhove, Jan, Audrey Bonvin, Amelia Lambelet & Raphael Berthele. 2019. Predicting perceptions of the lexical richness of short French, German, and Portuguese texts using text-based indices. *Journal of Writing Research* 10(3). 499–525. doi:10.17239/jowr-2019.10.03.04.

Yarkoni, Tal & Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives in Psychological Science* 12(6). 1100–1122. doi:10.1177/1745691617693393.