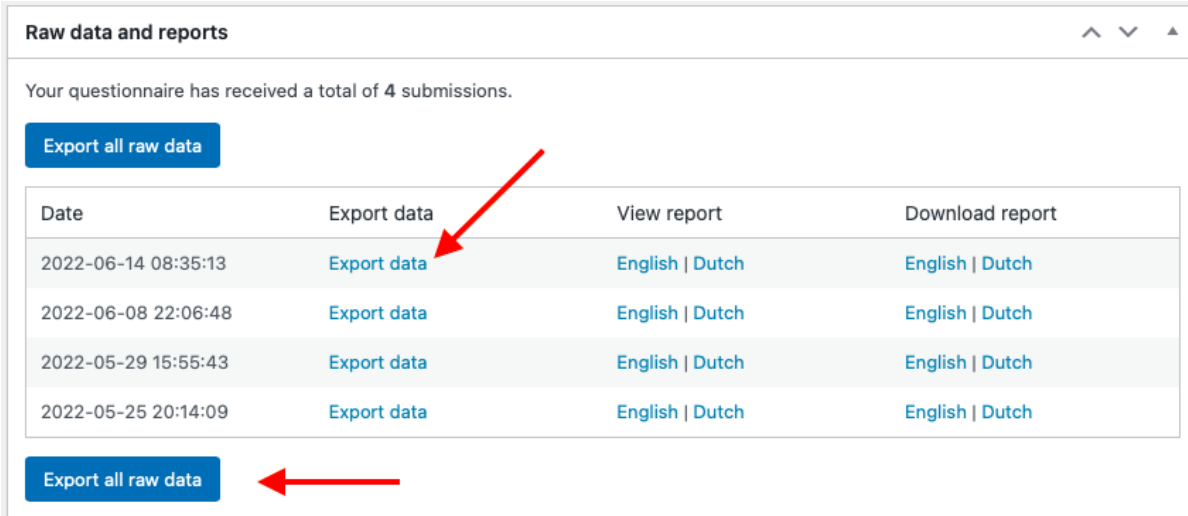# Accessing and interpreting data files

Accessing data files

On the page of each project that you create, you will be able to access four types of output files:
- Raw data
- Reports
- List of respondents' IDs and names
- Ordered response data and calculations

The output file containing the **raw data** can be found under the section *Raw data and reports*. You can either download the data (as a single .csv file) from all respondents by clicking *Export all raw data* or you can download the data for each participant individually by clicking *Export data* next to each submission (see Figure 1). Note that these files contain only raw data (i.e., no derived scores). All the data accessible in raw data files are anonymised as the names of children are automatically replaced by *respondent_id* (a variable present in the raw data file).

Figure 1. Exporting raw data

In the same section, you can view or download **reports** aimed at practitioners, which contain some of the data about participants' language history. Note that these files are not anonymised (i.e., they contain participants' names). You can access the reports in either English or Dutch (by clicking on the name of the language version which you prefer). See Figure 2 on how to access the report files.

Figure 2. Exporting report files

If you want to link individual data to each child, download the **list of respondents' IDs and names** from the section *Respondents* by clicking *Export respondent names* (see Figure 3).

Figure 3. Exporting respondents' names

In the section *Ordered response data and calculations*, you can download **ordered raw and calculated data** for all participants by clicking *Export ordered response data and calculations* (see Figure 4).

Figure 4. Exporting ordered response data and calculations

Interpreting data files

In order to interpret the data, you will require the *Output interpretation guide* excel file which can be downloaded from the resources on our website. The *Output interpretation guide* file contains four sheets: raw data, ordered data and calculations, report, and respondents.

The sheet *raw data* contains variables necessary for the raw data interpretation (see Figure 5). For ease of interpretation, columns E and F contain the questions (the caregiver and the child version respectively), while column A contains the internal reference for each

question. Columns B and C indicate which module and sub-module each question belongs to. In column D, you can find the variable names assigned to each question in the output file. Column G contains possible responses as they will appear in the output file, while column H clarifies what each of those values stands for. Finally, column I contains some additional notes to help the interpretation. Note that throughout the *Output interpretation guide* file, we use square brackets around numbers when these stand for variables. For instance, the answer(s) to question Q.72 ("Please list all the languages that the child speaks and/or understands - no matter how well or how often.") will be stored in the output file as *language 1* (if only one language is specified), and *language 2* (and possibly *language 3*) if more languages are specified by the respondent. Column D of the *Output interpretation guide* file only mentions these variables as *language [1]*, or *adult [1]* or *child [1]*, but they can be instantiated by several columns in the actual output file (depending on the number of languages, adults or children mentioned in the responses).

	A	B	C	D	E	F	G	H	I
1	question reference	module	sub-module	column name(s) in data output	question (caregiver)	question (child)	scale wording (data output)	scale wording (as seen by respondents)	notes
2	NA	NA	NA	country of residence	NA	NA	abbreviation of a country or name of a country	NA	This country is normally predetermined by the superuser as a country of residence for all participants. However, if the superuser includes CQ.18 in the questionnaire (in which participants can specify themselves where the child lives), the data in this column will be overwritten by whatever the participant responded to CQ.18.
3	CQ.15	Background information	NA	respondent_id	Please enter the child's name and surname.	Please enter your name and surname.	name replaced by an ID (a number)	insert	
4	NA	NA	NA	start_time	NA	NA	yyyy-mm-dd hh:mm:ss	NA	These variables are meta data collected for every respondent without asking them any questions
5	NA	NA	NA	end_time	NA	NA	yyyy-mm-dd hh:mm:ss	NA	
6	NA	NA	NA	version	NA	NA	version of the questionnaire [currently 8.1.0]	NA	
7	Q.72	Background information	NA	language [1]	Please list all the languages that the child speaks and/or understands - no matter how well or how often.	Please list all the languages that you speak and/or understand - no matter how well or how often.	names of a languages	select from a list of languages	
8	Q.73	Background information	NA	date of birth	What is the child's date of birth?	What is your date of birth?	yyyy-mm-dd	insert date: dd/mm/yyyy	
9 10 11 12 13	Q.75	Background information	NA	gender	What is the child's gender?	What is your gender?	gender-male	male	
							gender-female	female	
							gender-non-binary	non-binary	
							gender-other [it will show what the respondent typed in]	other: [please specify]	
							gender-prefer-not-to-specify	I prefer not to specify	

raw data | ordered data & calculations | reports | respondents | +

Figure 5. Output interpretation, raw data

The sheet *ordered data & calculations* lists the variables that appear in the output file *Ordered response data and calculations* (see Figure 6). While column A lists the names of these variables as in the output file, column B specifies whether the variable is a raw or a calculated measure. Columns C specifies the unit measure of calculated variables as well as an explanation or a reference to how the calculation has been derived. Column D clarifies when you can rely on a certain calculated variable, which primarily depends on which modules and sub-modules were included in the survey. Finally, in columns F-L, you can find additional notes to help you interpret the *Ordered response data and calculations* output.

	A	B	C	D	E	F	G	H	I
1	Variable name	Raw or calculated measure	Calculation explanation or reference	When can you rely on this calculation?					
2	Country of residence	raw	na	na					
3	Respondent ID	raw	na	na					NOTES
						This sheet contains a list of variables which can be f...			
						calculations' output downloadable from each project...			
						(raw and calculated) for all participants in an ordere...			
						number of people, events, languages that a child mi...			
						irrelevant for a child (e.g., 'language 3' because the...			
						output will be empty. The only raw data not include...			
						of hours that the child spends with each person, gro...			
						a typical week day, during odd week days, during a...			
						typical day on holidays. This data can be found in ra...			
						In the column 'Raw or calculated measure', we indic...			
						In the column 'Calculation explanation or reference'...			
						what the calculated measure represents or provide...			
						information can be found. For clarifications about ra...			
						In the column 'When can you rely on this calculation...			
						and sub-modules need to be distributed to the parti...			
4	Start time	raw	na	na					
5	End time	raw	na	na					
6	Version	na	na	na					
7	Language 1	raw	na	na					
8	Language 2	raw	na	na					
9	Language 3	raw	na	na					
10	Date of birth	raw	na	na					
11	Current age (months)	calculated	number of months [from the child's date of birth until the questionnaire completion date]	always					
12	Gender	raw	na	na					
13	Number of siblings	raw	na	na					
14	Child birth order	raw	na	na					
15	Country of birth	raw	na	na					
16	Child attended daycare	raw	na	na					

raw data | **ordered data & calculations** | reports | respondents | +

Figure 6. Output interpretation, ordered data and calculations

The sheet *reports* outlines the data summarised in reports aimed at teachers and speech and language therapists (see Figure 7), while the sheet *respondents* clarifies where to find participants' names linked to their IDs (see Figure 8).

	A	B	C	D
		NOTES		

In the section 'Raw data and reports' on each project's page, we provide downloadable summary reports of each child's language profile. These reports are aimed primarily at practitioners (e.g., teachers and speech and language therapists), yet they can be useful to researchers as well. They do not include a summary of all the data. Specifically the reports include the following:

- Child's name
- Basic demographic data about the child's age, languages, country of birth, number of siblings, date the child started school, languages spoken by their caregivers, age of first exposure to each language
- If relevant (sub-)modules are distributed: cumulative estimates for exposure and use in each language
- concerns about the child's language development and periods when the child missed school (if relevant)
- If relevant (sub-)modules are distributed: current weighted estimates for exposure and use in each language
- If relevant (sub-)modules are distributed: current unweighted estimates for exposure and use in 4 contexts (home, local community, school, during holidays)
- If relevant (sub-)modules are distributed: richness of linguistic experience in each language
- If relevant (sub-)modules are distributed: proficiency estimates for each language (listening, speaking, reading, writing)
- further information on each of the above measures

raw data | ordered data & calculations | **reports** | respondents | +

Figure 7. Output interpretation, reports sheet

Figure 8. Output interpretation, respondents sheet

The formulae used by the calculation functions are provided in the *Back-end calculator example* file, which can be downloaded from the resources on our website. Based on an invented example of a bilingual child, this spreadsheet explains how we obtain calculations for: weighted current exposure and use, unweighted current exposure and use across four contexts, cumulative exposure and use, as well as the richness estimates for each language.