1. ABOUT THE DATASET

------------

Title: A new receptive vocabulary size test for young language learners in England.

Creator(s): Nicola Morea (<https://orcid.org/0000-0003-0623-3078>) [1], Rowena Kasprowicz (<https://orcid.org/0000-0001-9248-6834>) [1], Astrid Morrison (<https://orcid.org/0000-0002-2765-2880>) [2], Carmen Silvestri (<https://orcid.org/0000-0003-1375-8729>) [1].

Organisation(s): 1. University of Reading. 2. Universidad Autónoma de Chile.

Rights-holder(s): University of Reading, Universidad Autónoma de Chile.

Publication Year: 2024

Description: These data were used to evaluate and refine three tests of receptive vocabulary size in French, German and Spanish respectively, which were administered to primary school students aged 7-11 in England in February-March 2023 (initial test) and in January-February 2024 (revised test). There are in total eight files (excluding this one): six data files in Excel format, one test file in PDF format (initial tests) and one test file in Word format (revised tests).

The three data files labelled “Language\_Initial\_Test\_Data” contain participant responses to the **initial** tests. Each initial test included three test forms sharing a nucleus of common items. Each test form was made of 25 multiple-choice items with four options, one of which was correct. For each item, students both read and heard (twice) the target word in French, German or Spanish, then ticked the option they believed represented the meaning of the word in English (with each option accompanied by a picture). The datasets include information on each participant (i.e., ID, school code, grade level, test form), participants’ responses to each item (1-4), participants’ scores for each item (1=correct, 0=incorrect) and participants’ total test scores (0-25).

In light of the poor psychometric properties of the initial tests (see related publication, reference below), the tests were re-designed and administered to the same sample of participants in January-February 2024. The three data files labelled “Language\_Revised\_Test\_Data” contain participant responses to these revised tests. Each revised test included two test forms sharing a nucleus of common items. Each test form was made of 40 multiple-choice items with four options, of which one was correct. For each item, students both read and heard (twice) the target word in French, German or Spanish, then ticked the option that they believed represented the meaning of the target word in English (with each option accompanied by a picture). The datasets include information on each participant (i.e., ID, school code, grade level, test form), participants’ responses to each item (1-4), participants’ scores for each item (1=correct, 0=incorrect) and participants’ total test scores (0-40).

Finally, the Word document labelled “Revised\_Tests” contains the revised tests for the three target languages. The PDF document called “Initial\_Tests” contains the initial tests. These tests have been included for reference only, and they should not be used due to their poor psychometric properties.

Related publication: Morea, N., Kasprowicz, R., Morrison, A., & Silvestri, C. (2024). Diverse population, homogenous ability: The development of a new receptive vocabulary size test for young language learners in England using Rasch analysis. *Research Methods in Applied Linguistics*, 3(3). 100166. https://doi.org/10.1016/j.rmal.2024.100166

2. TERMS OF USE

------------

Copyright 2024 University of Reading, Universidad Autónoma de Chile. This dataset is licensed under a Creative Commons Attribution 4.0 International Licence: https://creativecommons.org/licenses/by/4.0/.

3. PROJECT AND FUNDING INFORMATION

------------

Title: Progression in Primary Languages

Dates: January 2022-December 2025

Funding organisation: UK Research and Innovation (UKRI) Future Leaders Fellowship

Grant no.: MR/V023470/1

4. CONTENTS

------------

File listing

French\_Initial\_Test\_Data.xlsx

German\_Initial\_Test\_Data.xlsx

Spanish\_Initial\_Test\_Data.xlsx

These three files contain student responses to the **initial** tests. Each file has two tabs. The first tab contains the dataset. The second tab (“key”) lists the correct response to each item, by test form. The table below provides an overview of the variables in the datasets:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable name** | **Description** | **Unit** | **Coding** |
| Participant\_ID | Participants’ anonymous ID. | nominal | n/a |
| School\_Code | Anonymous code for participating schools. | numeric | n/a |
| Yeargroup\_2223 | Participants’ grade level/year group in the academic year 2022-2023. | numeric | 1=Year 3  2=Year 4  3=Year 5 |
| Class\_2223 | Participants’ class code, as some participating schools had more than one class per grade level/year group. | numeric | 1=class 1  2=class 2 |
| YearofLanguageStudy\_2223 | Number of years of formal instruction in the target language (as per academic year 2022-2023). | numeric | n/a |
| Vocab\_Test\_Version\_2223 | Test form/version completed by participants. | numeric | 1: version A  2: version B  3: version C |
| Q1\_Answer\_2223 – Q25\_Answer\_2223 | Participants’ response to each test item (Q1=item 1). | numeric | 1: option 1  2: option 2  3: option 3  4: option 4  5: multiple selection  6: missing answer (student present)  : missing answer (student absent) |
| Q1\_Score\_2223 – Q25\_Score\_2223 | Participants’ score on each test item (Q1=item 1).  Approach to converting answers into scores:  multiple selections (‘5’) and missing answers (‘6’) were scored as incorrect (‘0’); uncoded answers (‘’) were scored as missing (‘’). | numeric | 0: incorrect  1: correct  : missing answer |
| Tot\_Score\_2223 | Total test score. | numeric | n/a |

French\_Revised\_Test\_Data.xlsx

German\_Revised\_Test\_Data.xlsx

Spanish\_Revised\_Test\_Data.xlsx

These three files contain student responses to the **revised** tests in each language. Each file has two tabs. The first tab contains the dataset. The second tab (“key”) reports the correct response to each item, by test form. It also indicates which items were common between test forms/versions. The table below provides an overview of the variables in the datasets:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable name** | **Description** | **Unit** | **Coding** |
| Participant\_ID | Participants’ anonymous ID. | nominal | n/a |
| School\_Code | Anonymous code for participating schools. | numeric | n/a |
| Yeargroup\_2324 | Participants’ grade level/year group in the academic year 2023-2024. | numeric | 1=Year 3  2=Year 4  3=Year 5  4=Year 6 |
| Class\_2324 | Participants’ class code, as some participating schools had more than one class per grade level/year group. | numeric | 1=class 1  2=class 2 |
| YearofLanguageStudy\_2324 | Number of years of formal instruction in the target language (as per academic year 2023-2024). | numeric | n/a |
| Vocab\_Test\_Version\_2324 | Test form/version completed by participants. | numeric | 1: version A  2: version B |
| Q1\_Answer\_2324 – Q40\_Answer\_2324 | Participants’ response to each test item (Q1=item 1). | numeric | 1: option 1  2: option 2  3: option 3  4: option 4  5: multiple selection  6: missing answer (student present)  : missing answer (student absent) |
| Q1\_Score\_2324 – Q40\_Score\_2324 | Participants’ score on each test item (Q1=item 1).  Approach to converting answers into scores:  multiple selections (‘5’) and missing answers (‘6’) were scored as incorrect (‘0’); uncoded answers (‘’) were scored as missing (‘’). | numeric | 0: incorrect  1: correct  : missing answer |
| Tot\_Score\_2324 | Total test score. | numeric | n/a |

Revised\_Tests.docx

This file contains the **revised** tests for each language, in the following order:

French, version A

French, version B

German, version A

German, version B

Spanish, version A

Spanish, version B

Items crossed out are misfitting items that should not be included when administering the tests. Please refer to related publication (reference above) for more information on misfitting items and psychometric properties of the initial and the revised tests.

The pictures used for the tests are royalty-free images downloaded from the websites Pixabay (<https://pixabay.com/>), Unsplash (<https://unsplash.com/>) and Pexels (<https://www.pexels.com/>).

Initial\_Tests.pdf

This file contains the **initial** tests for each language, in the following order:

French, version A

French, version B

French, version C

German, version A

German, version B

German, version C

Spanish, version A

Spanish, version B

Spanish, version C

These tests are provided for reference only and should not be used.

The pictures used for the initial and revised tests are royalty-free images downloaded from the websites Pixabay (<https://pixabay.com/>), Unsplash (<https://unsplash.com/>) and Pexels (<https://www.pexels.com/>).

5. METHODS

-----------

Detailed information on how the datasets were generated may be found in the related publication (reference above). A summary is provided below.

**Initial tests:**

The initial tests were administered in February and March 2023 in 18 English primary schools. The research team visited each participating school to administer the language tests in the classroom during regular teaching hours using a pen-and-paper format. At the beginning of the session, the researcher reminded participants of the project aims, with the tests presented as language activities that the researchers would use to understand what the children had been learning in the target language. It was emphasised that students should not worry if they did not know some of the language featured in the activities, and that neither their teacher nor their parents would be able to access their individual responses. The researchers used a Power Point presentation to guide students through the tests. After test administration, each participant was assigned a unique code, and all identifying information was removed in the data cleaning process. Student responses were initially entered into Excel files. Occasional instances of ambiguous responses were flagged and discussed within the team until a final decision was reached. The Excel files were imported into the software SPSS (version 29). Responses were systematically scored into binary data (correct = 1, incorrect = 0) using the “convert into a different variable” function. Multiple responses and missing answers from participants present during the session were coded as incorrect. The SPSS datasets were then imported into the programme Winsteps (version 5.7.2) to assess the psychometric properties of the tests using Rasch analysis.

**Revised tests:**

The revised tests were administered in January and February 2024 in 17 primary schools in England to children from Year 3 to Year 6 (7-11 years old). Students were the same as in the previous data collection (but one year older), except for a new cohort of Year 3 students. The approach to test administration, data entry, coding and analysis was analogous to one used for the initial tests.