Preregistration of Maluch_LearnInst_2015_5Awm

Existing Data Replication

Study Information

1. Title (provided by SCORE)

RR TEAM INSTRUCTIONS: *This has been determined by SCORE.*

Replication of a research claim from Maluch et al. (2015) in *Learning and Instruction*.

2. Authors and affiliations

RR TEAM INSTRUCTIONS: *Fill in the names and affiliations of your team below.*

RR LAB LEAD[1]
Marco Ramljak A[2]
Belén Fernández-Castilla[3]
David Santos[4]

1 Affiliation 1
2 Student, Utrecht University 2
3 UNED, Madrid, Spain 3
4 Instituto Empresa, Madrid, Spain 4

3. Description of study (provided by SCORE)

RR TEAM INSTRUCTIONS: *This description has been provided by SCORE. Please review and make a SCORE project coordinator aware of any edits, additions, and corrections you would suggest to the paragraph. You are free to add additional descriptions of your project in a separate paragraph.*

The claim selected for replication from Maluch et al. (2015) is as follows: The general trend suggests that, given comparable background characteristics, children who speak a minority language at home have, on average, stronger foreign language achievement (Hypothesis 1a). This reflects the following statement from the paper's abstract: "Controlling for cognitive abilities, age, gender, socio-economic status, parental education, and indicators of cultural capital, the analysis revealed a general positive trend between bilingualism and English foreign language achievement." Given the authors' substantive interest in immigrant bilingualism and foreign language achievement (research question 1a), they fit two regression models testing if immigrant bilingualism is positively associated with English foreign language achievement

(Table 3). Model B is selected, and the focal variable is ‘Bilingual (=1)’ (see Model B in Table 3 for details). ...However, this negative relation is reversed once the background characteristics of general cognitive abilities, age, gender, socio-economic status, parental education, and cultural capital have been taken into account (Model B). Given comparable individual and familiar background characteristics bilingual group membership is positively associated with English foreign language achievement (from Model B in Table 3, ‘Bilingual (=1)’ variable: estimate = 2.68; p < .01).

4. Hypotheses (provided by SCORE with possible Data Analyst additions)

RR TEAM INSTRUCTIONS: *The focal test for SCORE is indicated as H*. If you will test additional hypotheses (or use alternate analyses) that help you to evaluate the claim your replication/reproduction is testing, number them H1, H2, H3 etc. (You can place H* in the list wherever makes sense). Please make sure that any additional hypotheses are logical deductions/operationalizations of the selected SCORE claim or are necessary to properly interpret the focal H* hypothesis. Research that is outside this scope should be described in a separate preregistration.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Are the listed hypotheses specific, concise, clearly testable, and specified at the level of operationalized variables?*
- *Are hypotheses identified as directional or non-directional, and, if applicable, have the direction of hypotheses been stated? (Example: “Customers’ mean choice satisfaction will be higher in the CvSS architecture condition than in the standard attribute-by-attribute architecture condition.”)*
- *Does the list of hypotheses/tests indicate whether additional hypotheses are taken from the original study or modified/added by the team?*

H*: Bilingual group membership will be positively associated with foreign language achievement when controlling for background variables.

Design Plan

5. Study type

NOTE: *The study type selected should be based on the data collected for the replication, and not necessarily the data used in the original study.*

- Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.
- **Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.**
- Meta-Analysis - A systematic review of published studies.
- Other

6. Blinding

RR TEAM INSTRUCTIONS: *Select any/all of the below that apply for your study by bolding them. You will give a longer description in the next question.*

- **No blinding is involved in this study.**
- For studies that involve human subjects, they will not know the treatment group to which they have been assigned.
- Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments. (Commonly known as "double blind")
- Personnel who analyze the data collected from the study are not aware of the treatment applied to any given group.

[QUESTION 6 - BOLD YOUR RESPONSE ABOVE]

7. Blinding

RR TEAM INSTRUCTIONS: *Since all existing data replications are based on data that has already been collected, in most cases it will not be necessary to comment on participant blinding. In the rare instance when an existing experiment is being re-analyzed for an existing data replication and blinding is a relevant consideration, please provide below any details regarding blinding that are important for a reviewer to be aware of.*

No blinding is involved in this study to the data finder's knowledge.

8. Study Design

RR TEAM INSTRUCTIONS: *Please describe how data was collected in the original study and how it compares to the data that was selected for the replication attempt. Explain why the data selected for the replication study is suitable for a replication and if any substantial deviations exist between the two.*

If the data used in the replication combines observations from the original study with new observations (e.g. if the data selected for the replication attempt comes from the same longitudinal survey as the original study), describe how 'original' and 'new' observations relate to each other and an estimate for what proportion of the final dataset's observations will be comprised of original vs. new observations.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify the unit of analysis?*
- *Does the preregistration provide sufficient detail about how the data selected for the replication attempt deviates from or is congruent with the data employed in the original study?*
- *Does the preregistration describe whether and how 'original' and 'new observations' are combined together for the replication dataset?*

The analyses of the original paper are secondary analyses of data from the Assessment of Reading and Mathematics Development Study (ELEMENT), which followed a cohort of children from the fourth grade to the sixth grade (ISCED2). The longitudinal study ELEMENT pursues the aim of examining the learning situation and progress of pupils in the transition area between primary and secondary level at primary schools in Berlin and undergraduate high schools. A design with three measuring times from fourth to sixth grade was realized. The pupils' performance in the areas of German and mathematics were the focus of the surveys. At the third time of measurement, they were supplemented by a performance measure in the area "English as a foreign language". The parents also received a questionnaire to determine the family living conditions of the children and their own educational attitudes. The original study focused on data from the sixth grade elementary school sample (third measurement point). The study's sample is representative for the public elementary school students (N = 2946), whose mean age is 14.97 (sd = 0.92), ranging from 11 to 23 years old.

The original study uses this dataset to investigate the effect of immigrant bilingualism on learning English as a foreign language. The SCORE focal claim is concerned with hypothesis 1a: "Given comparable background characteristics, children who speak a minority language at home have, on average, stronger foreign language achievement". The results of this hypothesis can be found in Table 3, Model b. The dependent variable is "English foreign language achievement", the focal independent variable is "Bilingualism" and the control variables are general cognitive abilities, age, gender, SES, parental education, cultural capital.

The proposed replication dataset is the European Survey on Language Competences (ESLC) from 2012. It is designed to collect information about the foreign language proficiency of students in the last year of lower secondary education (ISCED2) or the second year of upper secondary education (ISCED3) in 14 participating countries (Belgium, Bulgaria, Croatia, England, Estonia, France, Greece, Malta, Netherlands, Poland, Portugal, Slovenia, Spain, Sweden). The ESLC is a collaborative effort among the 16 participating educational systems[1] and the company SurveyLang partners to measure the language proficiency of approximately 53,000 students across Europe, to assist the European Commission in establishing a European Indicator of Language Competence to monitor progress against the March 2002 Barcelona European Council conclusions. These conclusions called for 'action to improve the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age' and also for the 'establishment of a linguistic competence indicator'.

Each educational system tested students in two languages; the two most widely taught of the five most widely taught European languages: English, French, German, Italian and Spanish. This effectively meant that there were two separate samples within each educational system, one for the first test language, and one for the second. Each sampled student was tested in one language only. Students' proficiency was assessed in two of the three skills of Listening, Reading and Writing. The ESLC sets out to assess students' ability to use language purposefully, in order to understand spoken or written texts, or to express themselves in writing. Their observed language proficiency is described in terms of the levels of the Common European Framework of Reference (CEFR) (Council of Europe 2001), to enable comparison across participating educational systems. Next to students' test and ability (Plausible Values) scores in Listening, Reading and Writing, the ESLC also contains Student-, Teacher- and Principal Questionnaire on contextual information.

For the potential replication dataset it is proposed to only use students who correspond to ISCED2. The reason for this is that in many educational systems this applies to the average age range of about 12 to 15 years, which is very close to the test assessment year of the original study (6th grade in Germany ~ 12 years old). In the ESLC, students who correspond to ISCED2 can differ in age between educational systems, however they all have in common the number of schooling years learning English as a foreign language. A potential age effect would be however noticed as the analysis model introduces age as a control variable. The subsample of the ESLC ISCED2 students entails 20,026 students.

While the original study used the Cloze test (consisting of four texts with 91 word completion questions measuring reading proficiency, vocabulary, grammar and spelling simultaneously) the proposed replication dataset includes measurements of students' listening, reading and writing abilities. Therefore, the dependent variable "English as a foreign language-ability" divided into reading, writing and listening abilities in the proposed replication dataset would be more refined. The proposed replication dataset contains all necessary variables needed for replication

[1] Belgium entails three different country communities with respective differing educational systems. Therefore, the number of countries in the overall ESLC sample does not equal the number of educational systems.

besides one control variable “general cognitive abilities” which will be further explained in section 12c. Furthermore, the average age of the students might slightly differ, however this is considered marginal as the findings/interpretations of the original study are not limited to this age group.

9. Randomization (free response)

RR TEAM INSTRUCTIONS: *If the variables used for this replication attempt were randomized, state how they were randomized, and at what level.*

No randomization was executed after the stage of the official data collection by ESLC, especially not by the data finder.

However, ESLC is a representative student assessment framework which underlies general sampling theoretical cornerstones in conducting their data collection process, such as which schools and which students within schools are tested. Furthermore, students have been tested in two of the three English ability skills (reading, writing, listening). Individual students have administered tests concerning either Reading and Listening, Reading and Writing, or Listening and Writing. Students have been assigned randomly to one of these three groups. To the data finder’s best knowledge this procedure is not done for experiments but for efficiency reasons and representative results.

For further information: ESLC Technical report, particularly p. 29 (student English ability), p. 106 and p. 126 (Simple random sampling).

Link: https://osf.io/c6y8r/

Sampling Plan

This section describes how the data sources for the replication were selected, how they were prepared into a replication dataset, and the number of observations that will be analyzed from these data. Please keep in mind that the data described in this section are the actual data used for analysis, so if you are using a subset of a larger dataset, please describe the subset that will actually be used in your study.

10. Existing data (multiple choice question, provided by SCORE)

1.1.1. Registration prior to creation of data
1.1.2. Registration prior to any human observation of the data
1.1.3. Registration prior to accessing the data
1.1.4. Registration prior to analysis of the data
1.1.5. Registration following analysis of the data

11. Explanation of existing data

NOTE: *For replications that rely on existing data sources, this question refers to the data that will be used for the replication analysis (i.e. the final replication dataset), and not (a) the data from the original study or (b) the data sources accessed to construct the replication dataset. Since no new data will be created for 'existing data replications,' 1.1.1 should never be selected. Since all analyses will occur after registration, 1.1.5 should also never be selected.*

The final dataset for replication has been accessed, and cleaned to a very low degree prior to registration. This means that the relevant students (ISCED2) were filtered and necessary variables were selected. Observations with NA values were preserved. The European Survey on Language Competences (ESLC) from 2012 has been chosen for replication. The final dataset includes all variables to conduct the main analysis. The main analysis for replication is a multi-level model considering the individual, the school and country level[2]. The selected variables in the final dataset represent the same constructs (besides the possibly differing levels) than the variables used in the original study. One variable from the original study, General cognitive abilities, is not available in the ESLC.

12. Data collection procedures

RR TEAM INSTRUCTIONS: *Please describe the process for constructing the replication dataset in as much detail as you can. The sections below should be used to provide the following information:*

- *Which variables are needed from the original study to perform a good-faith, high-quality replication.*
- *Which data sources were used, why they were selected, any deviations between the original study design and the replication study design that these selections present, and the procedures used to access the data.*
- *Which of the variables from the original study are available in the replication data sources, including relevant details about each measure.*
- *The procedure for creating the replication dataset, in both narrative and script form.*
- *A data dictionary that documents each variable included in the replication dataset.*

In the sections below, please provide links to the original materials whenever possible -- including descriptions of the original datasets and corresponding codebooks. If materials can be shared on the OSF, please do so, and provide view-only links to those materials.

Specific points to keep in mind for reviewers:

[2] If multiple countries are chosen for final replication analysis (as suggested by the data finder to increase power) the country level needs to be implemented into the regression model. As the original study only collected data from Germany, a country level was not necessary.

- *Does the preregistration describe which data sources were selected for the replication study and why each is suitable?*
- *Does the preregistration make clear how the data sources were used to construct the replication dataset?*

(a) Data Needed

RR TEAM INSTRUCTIONS: *List below the datasets and variables the original author used to analyze the focal claim. Include details regarding the sample size, waves or years used, and other details pertinent to finding an existing dataset for replication. Please include page numbers when excerpting from the original article. If possible, categorize the list of variables as one of the following: dependent variable, focal independent variable, control variable, or sample parameters/clustering variable. Finally, include the sample size of the original study's focal analysis, if it is available.*

As explained above, the original study utilized the ELEMENT data and focused on the sixth grade elementary school sample. While the authors clearly state which constructs they include in the analysis, the identification of concrete measurements is often not quite transparent.

Dependent Variable(s)

English ability

- Variable Data Source: ELEMENT, 2003-2005
- Assessed with a Cloze test (a procedure in which a subject is asked to supply words that have been removed from a passage as a test of their ability to comprehend text).
- The test consists of four texts with 91 word completion questions measuring reading proficiency, vocabulary, grammar and spelling simultaneously.
- The items were scaled based on one-parameter item response theory in ConQuest (Wu et al.,1998). Weighted likelihood estimates (WLEs) for individual person parameters were used. The WLEs were scaled with a mean parameter estimate of M = 100 and a standard deviation of SD = 20.

Focal Independent Variable(s)

Bilingualism

- Variable Data Source: ELEMENT, 2003-2005
- Variable in original study (focal claim): Being bilingual or not (dummy)
- Measured how: Number and frequency of languages spoken at home (parents and/or student questionnaire).
- Operationalized: It can be derived if a foreign language was spoken regularly at home and if so then a student is classified as bilingual

Control Variable(s)

General Cognitive Abilities

- Variable Data Source: ELEMENT, 2003-2005

- Variable in original study: numerical score for general cognitive ability
- Measured how: composite score of two subtests of the CFT4-12R: verbal and figural analogies (Heller & Perleth, 2000). This test consists of 25 picture and 20 word tasks subtests and was administered in the fourth grade.
- Operationalized: Numerical score of the composite

Age

- Variable Data Source: ELEMENT, 2003-2005
- Variable in original study: numerical age in years (centered)

Gender

- Variable Data Source: ELEMENT, 2003-2005
- Variable in original study: students’ gender (male, female)

Socio-economic status (HISEI)

- Variable Data Source: ELEMENT, 2003-2005
- Variable in original study: numerical score for SES (z-score)
- Measured how: International Socio-Economic Index (ISEI; Ganzeboom & Treiman, 1996), with the highest score of the two parents' socio-economic status (HISEI)
- Operationalized: Numerical score of the index (z-score)

Parental education/qualification

- Variable Data Source: ELEMENT, 2003-2005
- Variable in original study: numerical highest education of the parents (z-score)
- Measured how: Highest education of the parents measured on a five-point scale with 1 indicating “no qualification” and 5 indicating a “college-bound school diploma”.
- Operationalized: numerical score of the index (z-score)

Cultural capital

- Variable Data Source: ELEMENT, 2003-2005
- Variable in original study: numerical score on Cultural capital (z-score)
- Measured how: Number of books at home with one indicating 0-25 books and 4 indicating over 200 books
- Operationalized: numerical score of the index (z-score)

Cluster variable

School ID

Sample size of analysis has 2,946 observations (original paper, p.79). .

(b) Data Access

RR TEAM INSTRUCTIONS: *Describe below the data sources that will provide the replication variables. Include information such as the name of the data source (e.g., Indonesian Family Life Survey), the description and link of the data source, and the waves needed to create a final replication dataset.*

Also describe the process for accessing the data sources that will be used to create the final replication dataset; specify how long long it took for the registration to be approved and what information was required (e.g., writeup of the purpose of the project, email address from an IPCSR institution, etc.); and verify that the data can be opened as expected. If applicable, provide a link to the page where you registered to access the data.

Describe in detail any restrictions on data access and data-sharing, as well as any additional terms of data use that will be relevant for the replication study and final report (e.g. citations that will need to be made). If you were able to access the data because of special permissions that you have, but that you expect other researchers might not have, please document those as well.

Data access is granted by the Joint Research centre of the EU Commission and is limited to people who have registered this access at the JRC. Momentarily (21.08.2020), only the data finder (Marco Ramljak) and the data analyst have access to the raw data files. If further people need access this needs to be worked out via the data finder who will notify the JRC. The data can be freely used for research purposes related to the SCORE project.

It should be cited in the following way:
Costa, P., & Albergaria-Almeida, P. (2015). The European survey on language competences: Measuring foreign language student proficiency. *Procedia-Social and Behavioral Sciences*, *191*, 2369-2373.

(c) Variable Availability

RR TEAM INSTRUCTIONS: *For each variable required for the replication analysis (listed above), describe the variables from the replication data that can be used to measure it (including which data files or sources each measure is found in),* ***any notes a data analyst should consider when using the measure in a replication analysis****, and any important differences between the original variable and the proposed replication variable.*

If there are multiple variables in the replication data that correspond to a required variable (e.g. two different measures of education in the replication data), include all of those options below. If a variable from the original study ***cannot*** *be measured using the replication data, please make that clear as well.* ***Finally, include a description of the identifiers used to merge multiple datasets, if applicable.***

All variables which are needed for the replication of the focal claim can be found in the ESLC besides one control variable: score for general cognitive ability.[3] This is a limitation that needs to be addressed in the replication. Furthermore, data analysis should include a further cluster variable next to the School ID, for the multilevel analysis, namely the Country ID of the respective student.

Dependent Variable(s)

English ability

- Variable Data Source: ESCL, 2012
- Plausible values for the students' performance scores in Listening, Reading and Writing (Randomly assigned, students administered up to two out of the three skills mentioned). Plausible values for two English ability skills per student will be provided.
  - Acronym (e.g. for writing): PV1_WRIT_C - PV5_WRIT_C
- Alternative: Plausible level of students English ability skills in Listening, Reading and Writing. The Language test measures achievement of levels A1 to B2 of the Common European Framework of Reference (CEFR). The pre-A1 level which is also reported indicates failure to achieve A1 (Technical report, p. 5).
  - Acronym (e.g. for writing): PL1_WRIT_C - PL5_WRIT_C
- Further note: The data analyst has to decide if it is preferred to use plausible values or plausible levels for analysis. It is suggested by the data finder to use plausible values, as these are on a larger interval scale while plausible levels rely on a 5 point ordinal scale. This means that plausible values contain larger variation across the students.
- Difference to original study: While the original study used the Cloze test (consisting of four texts with 91 word completion questions measuring reading proficiency, vocabulary, grammar and spelling simultaneously) the proposed replication dataset includes measurements of students listening, reading and writing abilities.

Focal Independent variable

[3] General cognitive ability (GCA) is very often controlled for in similar studies as these studies assume a relation between having high GCA and high abilities in school subjects such as languages. When replicating this study without GCA I believe that the regression coefficient of the focal independent variable will be higher and the Adj.R^2 will be lower as in the original study. These assumptions stem from comparing Model A and Model B (focal model) in the original study (Table 3) - the effect size of the focal independent variable is larger in Model A then in Model B, the adj.R^2 is lower in Model A than in Model B. Unfortunately, no standard errors nor more granular models are reported, for example a model with only the focal independent variable and GCA implemented, to compare the unstandardized effect sizes with their respective coefficients in other models. This limitation needs to be addressed in a potential replication as the implementation of the background variables as whole into the regression equation are responsible for the positive sign of the focal independent variable regression coefficient. To the data finders' knowledge, no statistically profound assumption can be made - based on the available information reported in the original study - if controlling for GCA is solely responsible for the positive sign of the focal independent variable's coefficient.

Bilingualism

- Variable Data Source: ESCL, 2012
- Predefined compound index: Number of languages used at home
- Acronym: I03_ST_A_S26A
    - Measured how: The index equals the number of selected options in question SQ26 'Which language(s) do you, yourself, speak regularly at home?'.
    - Units: The index has the following categories: 1=”One language” (sum score=1); 2=”Two languages” (sum score=2); 3=”Three or more languages” (sum score≥3).
    - Proposed operationalization: Everyone who has more than one language indicated in the index will be categorized as bilingual (with a label = 1) and the others as monolingual (with a label = 0)
- In addition: Predefined index: Target language use at home
    - Acronym: I03_ST_A_S27B
    - Units: 0 = No; 1= Yes
    - Proposed operationalization: With this variable it can be determined if students speak English regularly at home. It is up to the data analyst if these should be filtered out for analysis.
    - Further note: It is not clear if students speaking English at home have been excluded from the analysis conducted in the original study.

Control variables

Age

- Variable Data Source: ESCL, 2012
- Predefined compound index: Age
- The index equals the difference between the date of the middle of the testing window in each Educational system and the date of birth SQ2 ‘What is your date of birth?’.
- Acronym: I08_ST_A_S02A

Gender

- Variable Data Source: ESCL, 2012
- Variable: Gender
- Acronym: SQt01i01
- Units: Female (0), Male (1)

Socio-economic status (HISEI)

- Variable Data Source: ESCL, 2012
- Predefined index: Parental occupation
- Acronym: HISEI
- The students’ answers to the four questions about parental occupation were coded in each educational system using the International Standard Classification of Occupations (ISCO-88) developed by ILO, including the PISA modifications:

    - SQ7 'What is your mother's main job?'
    - SQ8 'What does your mother do in her main job?'
    - SQ10 'What is your father's main job?'
    - SQ11 'What does your father do in his main job?'
- The codes for parental occupation (ISCO_M "International Standard Classification of Occupation mother" and ISCO_F "International Standard Classification of Occupation father") were transformed into the international socio-economic index of occupational status (ISEI) (Ganzeboom & Treiman 1996). The higher ISEI scores indicated higher levels of occupational status. The component "parental occupation (HISEI)" corresponds to the higher ISEI score of either parent or the only available parent's ISEI. For further information see technical report, p. 246.

Parental education

- Variable Data Source: ESCL, 2012
- Predefined index: Highest parental education expressed in years
- Acronym: PARED
    - The calculation of this component is based on a transformation of the answers to two questions:
        - SQ13 'What is the highest level of schooling completed by your mother?'
        - SQ14 'What is the highest level of schooling completed by your father?'
    - The responses to these questions were converted into estimated years of schooling using the mapping of PISA 2006 (OECD 2007) with a few small changes, because not all educational systems participating in the ESLC were represented in the PISA table. The component "higher parental education expressed as years of schooling" (PARED) corresponds to the higher PARED score of either parent or the only available parent's PARED. See technical report, p. 246.

Cultural capital

- Variable Data Source: ESCL, 2012
- VariableHome possessions
- Acronym: SQt21i01
- Question: How many books are there in your home?
- Units: 0-10 books (0), 11-25 books (1), 26-100 books (2), 101-200 books (3), 201-500 books (4), More than 500 books (5).

Cluster variable

Country ID

- Variable Data Source: ESCL, 2012
- Acronym: country_id

School ID

- Variable Data Source: ESCL, 2012
- Acronym: school_id

Student ID

- Variable Data Source: ESCL, 2012
- Acronym: respondent_id

Final Student Weight Reading trimmed

- Variable Data Source: ESCL, 2012
- Acronym: FSW_READ_TR

Final Student Weight Listening trimmed

- Variable Data Source: ESCL, 2012
- Acronym: FSW_LIST_TR

Final Student Weight Writing trimmed

- Variable Data Source: ESCL, 2012
- Acronym: FSW_WRIT_TR

Final Student Weight Questionnaire trimmed

- Variable Data Source: ESCL, 2012
- Acronym: FSW_QUES_TR

Target language

- Variable Data Source: ESCL, 2012
- Further note: This variable is needed to select students tested in English.
- Acronym: targetLanguage_id

Program level

- Variable Data Source: ESCL, 2012
- Further note: This variable is needed to select students on the ISCED 2 level.
- Acronym: I14_ST_A_S06A

(d) Data Creation

RR TEAM INSTRUCTIONS: *Create a dataset using the data sources and variables listed above. Provide a detailed narrative describing how the various datasets were cleaned and merged into a final replication dataset. Provide a view-only link to a clearly commented script on the OSF that produces the replication data as described in the narrative. Our preference is that this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible.*

- *If the data can be freely shared and posted to OSF, please post it in your OSF project and provide a link to the completed dataset below.*

- *If any part of the dataset cannot be shared between researchers or posted to the OSF, please leave the final dataset off the OSF. Instead, include either below or in your script (commented out at the bottom) two pieces of information that will help an independent team verify they have created the dataset according to your instructions:*
    - *The dimensions of the final dataset(s) you've created (# of rows, # of columns)*
    - *A summary of 8-10 variables in the replication dataset. For numeric variables, the summary should include the mean, standard deviation, and count of NAs. For categorical variables, the summary should include each level present in the data and its count, as well as a count of NAs. If multiple datasets are submitted as part of your work, at least one variable should be included from each dataset.*

The data from the replication sources should be preserved in as 'raw' a form as possible, in order to give the data analyst the most latitude to clean the variables as they see fit. Variables from the original source should be preserved in their original form (e.g. do not recode values of 99 to NA). New variables should only be created when they're needed to complete the merge or combine the datasets; in those cases, please preserve a version of the original, unaltered variable in the new dataset.

Please also use this section to describe:

- *Any deviations between the original study design and the replication design that would result from using this replication dataset.*
- *Any notes about using these variables that you would like to pass along to the data analyst.*

The attached uncompiled R-notebook script is very detailed in its sections and can be found here: https://osf.io/y742m/

The most important issue for the data analyst concerns the dependent variable. In the replication dataset I included the plausible values for student's English ability skills in reading, writing and listening, as well as the plausible level for student's English ability skills in reading, writing and listening. The data analyst has to decide which set of variables to use for the analysis. Another issue related to the dependent variable is that each student was only tested in two of the English ability skills (see explanation in the randomization section). Consequently, each student has only plausible values for two of the three English ability skills (Reading and Listening, Reading and Writing, or Listening and Writing). No overall value for a student's English abilities in reading, writing and listening is available in the dataset. General information on how to handle plausible values can be found in the Technical report in Chapter 12.

Additionally, relevant student weights needed for analysis are included in the replication dataset. General information on how to handle these weights can be found in the Technical report in Chapter 13.7.

Link to the report: https://osf.io/c6y8r/

Lastly, concerning the focal independent variable bilingualism: The replication dataset includes a variable on the “Number of languages used at home”. For the suggested operationalization of this variable see Variable Availability. I, the data finder, suggest controlling for the fact if a student uses the target language (English) at home. This information can be found in the variable “Target language use at home”.

(e) Data Dictionary

RR TEAM INSTRUCTIONS: *Create a data dictionary following this template. Provide below a view-only link to the completed data dictionary included in the OSF project. If the Data Analyst will need to create new variables using the variables in the final replication dataset (e.g. recoding the provided education variable to be in a better format for analysis), please document below your recommendation on how the analyst should do so. Please also document any additional notes regarding the variables in the dataset that do not fit within the provided data dictionary template or the other sections above.*

The data dictionary can be found here: https://osf.io/ym5ac/

The codebook can be found here: https://osf.io/j4gu7/

The technical report can be found here:: https://osf.io/c6y8r/

13. Sample size

RR TEAM INSTRUCTIONS: *Please report below the analytic sample size(s) in the replication dataset, with reference to however many units or levels are in the data. Please report as much information here as will be helpful for the review committee to be aware of, including differences in sample size resulting from various analytic decisions (e.g. listwise deletion vs multiple imputation).* ***Finally, when the replication combines observations from the original study with new observations, please estimate what proportion of the analytic sample's observations will be comprised of original vs. new observations.***

Data finders’ response goes here:

For the potential replication dataset it is proposed to only use students who correspond to ISCED2 and who have participated in an English test. For most of the different educational systems this applies to the average age range of about 12 to 15 years. The subsample of the ESLC ISCED2 students entails 20,026 students, nested in 901 schools in 11 countries. When excluding students with NAs for relevant variables as explained above, only the number of students is reduced, not the number of schools nor the countries.

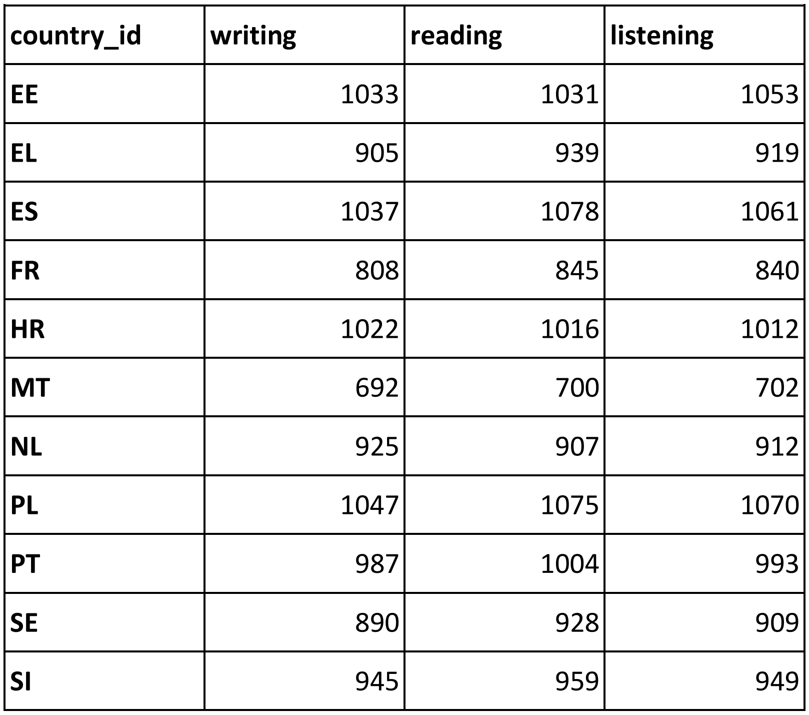
As explained above, students have been tested in two of the three English ability skills (reading, writing, listening). Individual students have received Reading and Listening, Reading and Writing, or Listening and Writing. Consequently, students only have plausible values/levels for the respective two English ability skills.

When only focusing on the English ability skill writing (as a dependent variable), the final replication dataset contains 13,006 observations. After listwise deletion of relevant variables (exclusion of variables related to reading and listening), the final replication dataset includes 10,291 observations.

When only focusing on the English ability skill listening (as a dependent variable), the final replication dataset contains 13,245 observations. After listwise deletion of relevant variables (exclusion of variables related to reading and writing), the final replication dataset includes 10,420 observations.

When only focusing on the English ability skill reading (as a dependent variable), the final replication dataset contains 13,376 observations. After listwise deletion of relevant variables (exclusion of variables related to listening and writing), the final replication dataset includes 10,482 observations.

The following table presents country specific sample sizes for each English ability skill (writing, listening, reading) with listwise deletion applied:



country_id	writing	reading	listening
EE	1033	1031	1053
EL	905	939	919
ES	1037	1078	1061
FR	808	845	840
HR	1022	1016	1012
MT	692	700	702
NL	925	907	912
PL	1047	1075	1070
PT	987	1004	993
SE	890	928	909
SI	945	959	949

Required sample size [to be filled out by the SCORE team]: The primary unit of analysis is the student. An estimate of the minimum viable sample size for the data analytic replication is:

1643. For comparison, the stage1 required sample size would be: 5970 and the stage2 sample size would be: 8950.

14. Sample size rationale

For data analytic replications in SCORE, three sample sizes are calculated:

- *A minimum threshold sample size, defined as the sample size required for 50% power of 100% of the original effect*
- *A stage 1 sample size, defined as the sample size needed to have 90% power to detect 75% of the original effect*
- *A stage 2 sample size, defined as the sample size needed to have 90% power to detect 50% of the original effect*

Details about how those sample sizes were calculated for this project are found here: https://osf.io/yhp8r/?view_only=ce7c2737afbe458aaf8d9a05e78f7d61

15. Stopping rule (provided by SCORE)

RR TEAM INSTRUCTIONS:

Since the replication data does not include any observations from the original analysis, SCORE recommends that only a single analysis using all available observations be performed for the replication.

Variables

RR TEAM INSTRUCTIONS: *The preregistration form divides variables across three questions: manipulated variables, measured variables, and indices (i.e. analytic variables derived from raw variables). For existing data replications, only fill out the "Measured variables' and 'Indices' sections. Please do not fill out anything in the 'Manipulated variables' section.*

The raw data of any transformed variable (e.g. reaction time → log reaction time) or any created index should be defined in the 'Measured variables' section. Details regarding the variable transformation should be specified in the 'Transformations' section. Details regarding the creation of an index should be specified in the 'Indices' section.

Across these questions, you should define all variables that will later be used during your analysis (including data preparation/processing). You can describe all variables in the preregistration and/or summarize and link to a data dictionary (codebook) in your repository to answer these questions.

If you will share data from your replication, this is also the place to state whether any variables will be removed prior to sharing the dataset (e.g. to reduce risk of participant identification or comply with copyright restrictions on scale items.)

16. Manipulated variables

RR TEAM INSTRUCTIONS: *Manipulated variables in this preregistration refer specifically to variables that have been randomly assigned in an experiment. The use of data from an experiment should be rare in existing data replications. If your existing data replication relies on experimental data, please document each manipulated variable as a measured variable, and use the codebook to indicate what each level of the variable corresponds to (e.g. participants assigned to the treatment condition = 1; participants assigned to the control condition = 0). The default language in bold below has been copied into all existing data replication preregistrations.*

N/A -- not documented for existing data replications.

There have been no manipulated variables in this project. All variables were measured.

17. Measured variables

RR TEAM INSTRUCTIONS: *Please use this section to document each variable that was used in the original study's analysis and the role it served (e.g. dependent variable, control variable, sample parameter, etc). For each variable, provide the description of the variable offered in the paper and/or codebook of the original study, the variable in the replication dataset that it corresponds to, and explain any deviations between the two. In cases where an equivalent replication variable was not found, explain how, if at all, you expect it will affect the replication attempt. In cases where you are adding a variable that was not present in the original study, please explicitly state that you are doing so, and explain how, if at all, you expect it will affect the replication attempt.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration surface all of the variables needed to replicate the focal analysis?*
- *Are deviations between the original variables and replication variables documented when needed?*

VARIABLE NAME

- [Use in the analysis]
- [Description from the original study]
- [Variables used in the replication (if it needs to be constructed from multiple measures, include all of them here)]
- [Deviations between the original study and the replication study]

English ability

- Dependent variable
- English language achievement was assessed with a Cloze test (Lehmann & Lenkeit, 2008). The test consists of four texts with 91 word completion questions measuring reading proficiency, vocabulary, grammar and spelling simultaneously. The items were scaled based on one-parameter item response theory in ConQuest (Wu et al., 1998). We used weighted likelihood estimates (WLEs) for individual person parameters. The WLEs were scaled with a mean parameter estimate of M =100 and a standard deviation of SD = 20 (original paper: page 79, section 2.2.1).·
- **ave_writing** (PV1_WRIT_C, PV2_WRIT_C, PV3_WRIT_C, PV4_WRIT_C, PV5_WRIT_C), **ave_reading** (PV1_READ, PV2_READ, PV3_READ, PV4_READ, PV5_READ), and **ave_listening** (PV1_LIST, PV2_LIST, PV3_LIST, PV4_LIST, PV5_LIST)·
- In the original article, the English ability was based on a single measure with different domains (i.e., reading proficiency, vocabulary, grammar and spelling) while in the replication we measured 3 English domains (i.e., writing, reading, and listening) separately. The reason for splitting the dataset was that a check-up analysis showed that, on average, students scored higher in listening and reading compared to writing, and students scored higher in listening compared to reading. Therefore, we decided to analyze these constructs separately, because each student is only measured in two of these three dimensions, which means that students that have scores on *'reading*' and '*listening'* will systematically score higher in English (regardless of the independent variable) than students that have scores on *reading/listening* AND *writing*, biasing the results.

Bilingualism

- Focal Independent Variable·
- To identify bilingual groups, we analyzed the language spoken at home as well as the frequency of the language spoken at home as reported by the parents and students. Inclusion into one of the language groups was determined when the parents reported that a language other than German (the specific language was also identified) was regularly spoken in the home. If the parents' information was missing (24%), the information provided by the student was used. From the aforementioned criterion, we identified the monolingual group (n = 1896) and the bilingual group (n = 939). (original paper: page 79, section 2.1).
- I03_ST_A_S26A
- In the replication, we did not have information about the language spoken at home. We only had information about how many languages were spoken at home (1, 2 or 3). This variable was re-coded so that if **I03_ST_A_S26A** = 1 (only one language spoken at home) means that the student is monolingual (Bilingual = 0) and if **I03_ST_A_S26A** = 2 or 3 (two or more languages spoken at home), then the student is bilingual (Bilingual = 1). We should warn about the risk of proceeding this way: Maybe students that speak only one language at home (and therefore are considered monolinguals), speak a language different from the instructional (e.g., a minority language), being actually bilingual. There is no way to know this with the replication dataset, because the idiom spoken at home is not coded.

Age

- Control variable
- Gender and age were reported in the fifth and sixth grade student questionnaires (original paper: page 79, section 2.2.2).
- c_age (centered variable based on I08_ST_A_S02A)
- There are no deviations from the procedure used in the original paper.

Gender

- Control variable
- Gender and age were reported in the fifth and sixth grade student questionnaires (original paper: page 79, section 2.2.2).
- SQt01i01
- There are no deviations from the procedure used in the original paper.

Socio-economic status (HISEI)

- Control variable
- To measure the family's socio-economics status, we used the International Socio-Economic Index (ISEI; Ganzeboom & Treiman, 1996), with the highest score of the two parents' socio-economic status (HISEI) serving as an indicator of family socio-economic status (original paper: page 80, section 2.2.2).
- c_HISEI (centered variable based on HISEI)
- There are no deviations from the procedure used in the original paper.

Parental education

- Control variable
- The highest education of the parents was measured on a five-point scale with one indicating no qualification and five indicating a college-bound school diploma (original
- paper: page 80, section 2.2.2).
- Z_Parental (Z-scores the variable PARED)
- In the original paper, this variable was coded on a 5-point scale, while in the replication, the variable is a numeric variable representing “Highest parental education expressed in years”

Cultural capital

- Control variable
- Finally, to operationalize cultural capital, we used the number of books at home as a proxy variable, which was assessed on a four-point scale with one indicating 0-25
- books and four indicating over 200 books. (original paper: page 80, section 2.2.2)
- Z_Cultural (Z-scores transformed from Cultural_capital based on SQt21i01)
- The variable was calculated differently than in the original paper. In the original paper, they reported the following values: 1 = 0-25 books and 4 = over 200 books. In the replication, we use the following values: 0 = 0-10 books, 1 = 11-25 books, 3 = 101-200 books, 4 = 201-500 books, and 5 = More than 500 books.

Cognitive abilities

- Control variable
- As general cognitive abilities might systematically differ across groups, we used a composite score of two subtests of the CFT4-12R: verbal and figural analogies (Heller & Perleth, 2000). This test consists of 25 picture and 20 word tasks subtests and was administered in the fourth grade. The test-retest reliability for this age group is $r_{\text{analogies}}$ = 0.83 and r_{figural} = 0.93 (Heller & Perleth, 2000).
- There is no equivalent variable in the replication dataset.
- This was an important variable to control for in the original paper, and we expect that this will affect the replication attempt. In the original paper, after controlling for cognitive abilities, "the Turkish-German bilingual group and the Other Bilingual group have, on average, no longer significant disadvantages and similar levels of English achievement as their monolingual peers." (original paper, page 81). Therefore, if we find a difference we cannot be sure if that difference would not have resulted if this variable would have been controlled for.

Country ID

- Cluster variable
- This variable was not used in the original study
- country_id
- We do not expect this variable to affect the replication attempt as it gives more control.

School ID

- Cluster variable
- This variable was not used in the original study
- School_id
- We do not expect this variable to affect the replication attempt as it gives more control.

18. Indices

RR TEAM INSTRUCTIONS: *If any of the measured variables described in Section 17 will be combined into a composite measure (including simply a mean), describe in detail what measures you will use and how they will be combined. Please be sure this preregistration includes a link to a clearly commented script that constructs the index according to the narrative.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify each of the composite measures (e.g. mean scores, factor scores) that are needed for the focal analysis, and which of the measured variables in Section 17 are used in each one (e.g. the happiness, joy, and satisfaction items will be used to create the 'positive feelings' measure)?*
- *Does the preregistration link to a clearly commented script that constructs the indices according to the narrative description?*

The variables described in point 16 are not part of other indices and will be used in the Analysis separately. The data transformations have been commented on point 17 above and point 19 below.

Analysis Plan

19. Statistical models

RR TEAM INSTRUCTIONS: *This section should describe in detail the analysis that will be performed to replicate the focal result. This analysis must align as closely as possible with the original study's analysis, even if you have identified limitations in the original study. The level of detail should allow anyone to reproduce your analyses from your description below. Examples of what should be specified: the model; each variable; adjustments made to the standard errors and to case weighting; additional analyses that are required to set up the focal analysis; and the software used.*

Beyond the replication of the focal analysis from the original study, it is at your discretion to test the claim using other analytic approaches as a check of the robustness of the claim. The original test should be listed first and be clearly distinguished from any other tests. If you are testing additional confirmatory hypotheses, describe them in the same order as you numbered them in the "Hypotheses" section above and make clear reference to the specific hypothesis being tested for each.

Please provide a link to a clearly commented script that performs the analysis described in the narrative provided below. Our preference is that this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible. Please also test that the code runs without error on a random subset of 5% of the replication dataset, and provide verification that the code has produced a sensible result below by providing a screenshot of the output (please upload the screenshot to the OSF as well). Finally, please confirm that you have only developed and tested your analysis plan and code using 5% of the data.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify which statistical model will be used to provide the 'focal evidence' for the SCORE test (e.g. a regression coefficient in a larger multiple regression model), and does it correspond closely to the model and evidence from the original study?*
- *Does the preregistration describe each variable that will be included in the focal analysis, and what role each variable has (e.g. dependent variable, independent variable)?*
- *Does the preregistration include a detailed specification of the focal analysis, including interactions, lagged terms, controls, etc., in both narrative form and in a clearly commented script?*
- *Does the preregistration verify that the code runs without error on a random subset of the replication dataset?*

Four steps for data preparation:

1. **Data exclusion.** The following observations were excluded from the analysis:

The following missing data were observed and excluded. From the 20026 observations, 87 did not report information on the variable I03_ST_A_S26A, which is the variable used to classify participants in monolinguals or bilinguals. Therefore, after removing these 87 observations, 19939 participants were kept.

From these 19939 observations, 569 participants were removed because no information was reported on the language they talked at home. Additionally, other 368 participants were excluded because they speak English (the target language) at home. After removing 937 participants, 19002 observations were kept for the next analyses.

At this point, the dataset was split in three datasets: one for writing, another from reading, and finally for listening. No missing values were observed in any of the items that measured each dimension as long as the student was measured on that dimension. That is, if a student was measured in WRITING and READING, no missing data was observed in any of the items that measure these dimensions, although of course the items measuring LISTENING were empty because the student wasn't assessed in this dimension.

According to section 13.7 of the technical report (that can be found here: https://osf.io/c6y8r/), *"only students and schools that meet the formal criteria for participation have a weight in the datasets".* Therefore, for each separate analysis (i.e., reading, writing, and listening), the participants that did not have weight (i.e., missing value) or that had a weight of 0 were removed from the analysis. The names of the variables that have this information are: **FSW_READ_TR** (for reading), **FSW_WRIT_TR** (for writing), **FSW_LIST_TR** (for listening). In the table below it is indicated how many observations remained available after the exclusion of those observation where the weight was missing or equal to zero:

	Number of observations available for each dimension	Number of observations available after removing no weights or weight = 0
Writing	12329	10512
Reading	12723	5406
Listening	12558	10714

The sample size in the last column was the final sample size available for each analysis.

2. Data transformation.

- The independent variable of this study is whether students speak more than 1 language at home. In the replication dataset, there is a variable (**I03_ST_A_S26A**) that measures the number of languages spoken at home: 1, 2 or 3. This variable was re-coded so that if **I03_ST_A_S26A** = 1 (only one language spoken at home) means that the student is monolingual (Bilingual = 0) and if **I03_ST_A_S26A** = 2 or 3 (two or more languages spoken at home), then the student is bilingual (Bilingual = 1)

-The student's performance on the target language (i.e., English) is measured in three dimensions: Reading, Writing and Listening. The dataset includes plausible values and plausible levels, but following the recommendations of the data finder, we used plausible values because they are numeric instead of character variables. Each of the three dimensions are measured through 5 items. Therefore, a global score of each dimension was calculated by obtaining the average across the 5 items (per dimension).

-The variable "Cultural capital" is recorded to convert it to a continuous variable. This variable has six categories:

- "0-10 books" = 0
- "11-25 books" = 1
- "26-100 books" = 2
- "101-200 books" = 3
- "201-500 books" = 4
- "More than 500 books" = 5

-After this point, three datasets are created, one for each dimension of the variable "English Proficiency": Writing, Reading, and Listening. This procedure deviates from the procedure followed by the original paper, where a global average was calculated between measures of reading, vocabulary, grammar and spelling in English. The reason for splitting the dataset was that a check-up analysis showed that, on average, students scored higher in listening and reading compared to writing, and students scored higher in listening compared to reading. Therefore, we decided to analyze these constructs separately, because each student is only measured in two of these three dimensions, which means that students that have scores on *'reading*' and '*listening'* will systematically score higher in English (regardless of the independent variable) than students that have scores on *reading/listening* AND *writing*, biasing the results. The model used to test the difference between these three dimensions was a repeated-measures four-level model (measurements within cases, nested in schools, nested in countries), using as an independent variable the type of dimension used (i.e., writing, reading, listening). The dataset had to be transformed into a long format to carry out this check-up analysis.

-For each of the three datasets (listening, writing and reading), the variables Age and HISEI (SES) were centered.

-For each of the three datasets (listening, writing and reading), Parental education and cultural capital are transformed to Z-scores.

3. Data selection

-5% of the students that belong to the bilingual group and 5% of the students that belong to the monolingual group were selected for each of the three datasets (which constitute 5% of the total dataset)

4. Model

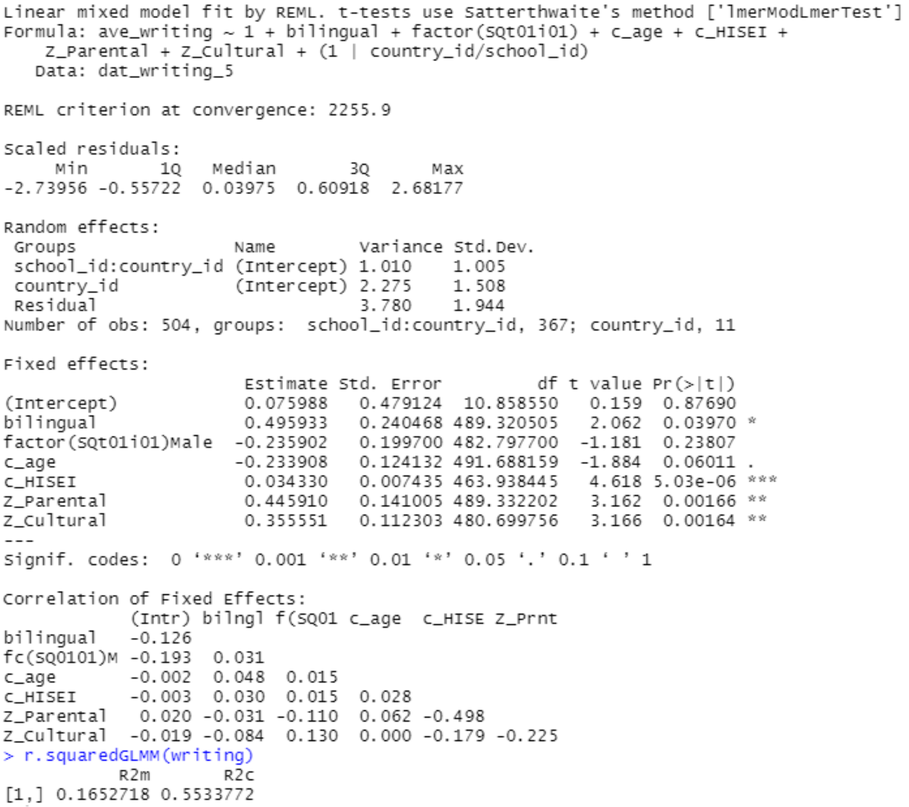
-A multilevel three level model (cases within schools within countries) was fitted with the R package lmerTest, with intercept. The variables country_id and school_id were specified as random-effects, whereas the variables being bilingual or not (bilingual), gender (SQt01i01), age (c_age), HISEI (c_HISEI), parent qualification (Z_Parental) and cultural capital (Z_Cultural) where introduced as fixed effects. The estimation method selected was Restricted Maximum Likelihood (REML). The function used was lmer, from lmerTest Package, which gives by default t-tests and degrees of freedom corrected by Satterthwaite's method.

-R square was calculated with MuMIn package, using the function r.squaredGLMM, which gives the marginal R2 (variance explained by the fixed effects) and the conditional R2 (variance explained by the entire model, including both fixed and random effects)

This statement confirms that only 5% of the data have been randomly sampled in developing the analysis plan and code contained in this preregistration.

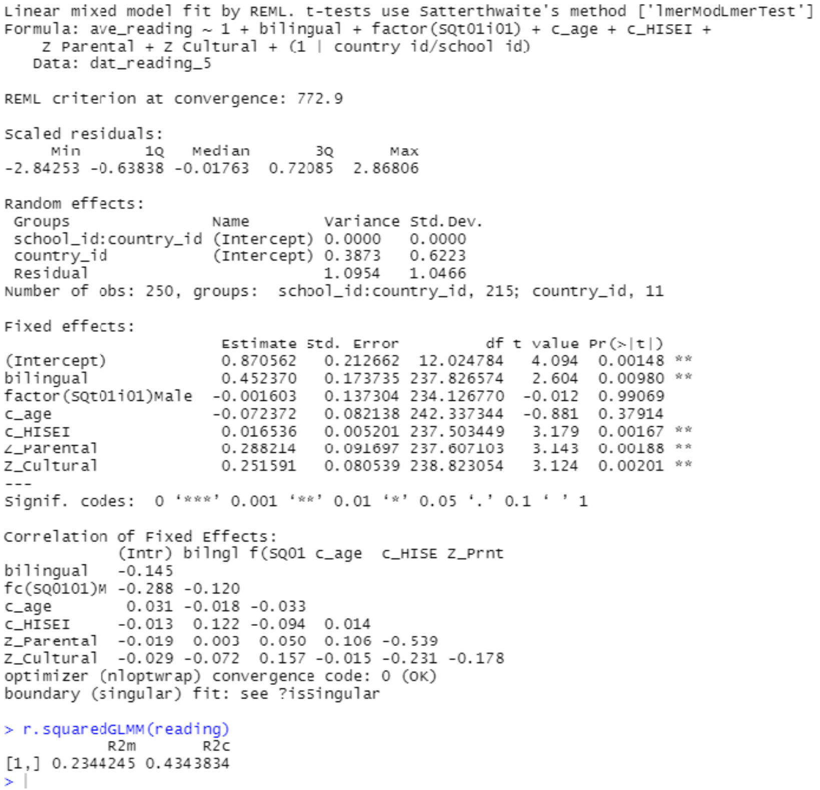
Screenshot of the output

Writing



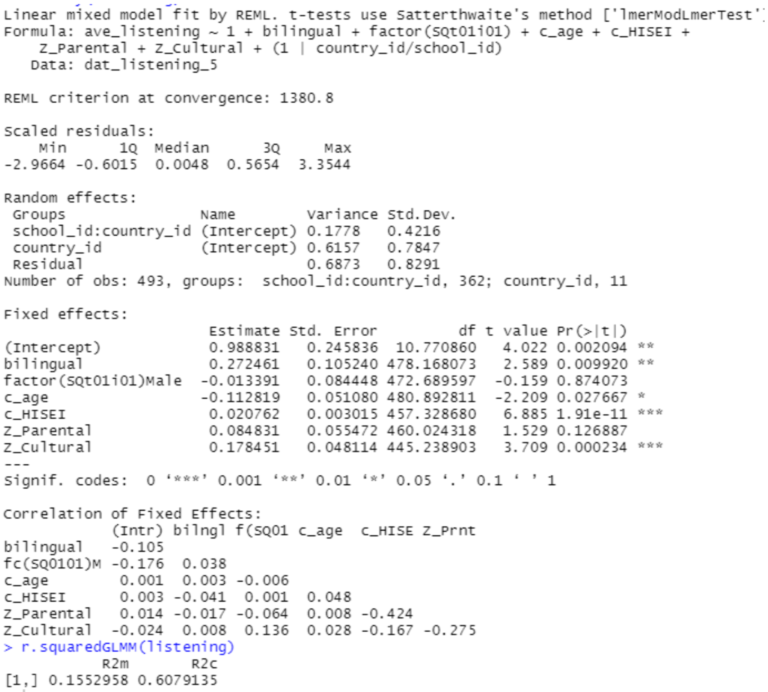
```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: ave_writing ~ 1 + bilingual + factor(SQt01i01) + c_age + c_HISEI +
    Z_Parental + Z_Cultural + (1 | country_id/school_id)
   Data: dat_writing_5

REML criterion at convergence: 2255.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.73956 -0.55722  0.03975  0.60918  2.68177

Random effects:
 Groups               Name        Variance Std.Dev.
 school_id:country_id (Intercept) 1.010    1.005
 country_id           (Intercept) 2.275    1.508
 Residual                         3.780    1.944
Number of obs: 504, groups:  school_id:country_id, 367; country_id, 11

Fixed effects:
                       Estimate Std. Error         df t value Pr(>|t|)
(Intercept)            0.075988   0.479124  10.858550   0.159  0.87690
bilingual              0.495933   0.240468 489.320505   2.062  0.03970 *
factor(SQt01i01)Male  -0.235902   0.199700 482.797700  -1.181  0.23807
c_age                 -0.233908   0.124132 491.688159  -1.884  0.06011 .
c_HISEI                0.034330   0.007435 463.938445   4.618 5.03e-06 ***
Z_Parental             0.445910   0.141005 489.332202   3.162  0.00166 **
Z_Cultural             0.355551   0.112303 480.699756   3.166  0.00164 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) bilngl f(SQ01 c_age  c_HISE Z_Prnt
bilingual   -0.126
fc(SQ0101)M -0.193  0.031
c_age       -0.002  0.048  0.015
c_HISEI     -0.003  0.030  0.015  0.028
Z_Parental   0.020 -0.031 -0.110  0.062 -0.498
Z_Cultural  -0.019 -0.084  0.130  0.000 -0.179 -0.225
> r.squaredGLMM(writing)
           R2m       R2c
[1,] 0.1652718 0.5533772
```

Reading



```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: ave_reading ~ 1 + bilingual + factor(SQt01i01) + c_age + c_HISEI +
    Z Parental + Z Cultural + (1 | country id/school id)
   Data: dat_reading_5

REML criterion at convergence: 772.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.84253 -0.63838 -0.01763  0.72085  2.86806

Random effects:
 Groups               Name        Variance Std.Dev.
 school_id:country_id (Intercept) 0.0000   0.0000
 country_id           (Intercept) 0.3873   0.6223
 Residual                         1.0954   1.0466
Number of obs: 250, groups:  school_id:country_id, 215; country_id, 11

Fixed effects:
                       Estimate Std. Error         df t value Pr(>|t|)
(Intercept)            0.870562   0.212662  12.024784   4.094  0.00148 **
bilingual              0.452370   0.173735 237.826574   2.604  0.00980 **
factor(SQt01i01)Male  -0.001603   0.137304 234.126770  -0.012  0.99069
c_age                 -0.072372   0.082138 242.337344  -0.881  0.37914
c_HISEI                0.016536   0.005201 237.503449   3.179  0.00167 **
Z_Parental             0.288214   0.091697 237.607103   3.143  0.00188 **
Z_Cultural             0.251591   0.080539 238.823054   3.124  0.00201 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) bilngl f(SQ01 c_age  c_HISE Z_Prnt
bilingual   -0.145
fc(SQ0101)M -0.288 -0.120
c_age        0.031 -0.018 -0.033
c_HISEI     -0.013  0.122 -0.094  0.014
Z_Parental  -0.019  0.003  0.050  0.106 -0.539
Z_Cultural  -0.029 -0.072  0.157 -0.015 -0.231 -0.178
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see ?isSingular

> r.squaredGLMM(reading)
           R2m       R2c
[1,] 0.2344245 0.4343834
> 
```

Listening



```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest'
Formula: ave_listening ~ 1 + bilingual + factor(SQt01i01) + c_age + c_HISEI +
    Z_Parental + Z_Cultural + (1 | country_id/school_id)
   Data: dat_listening_5

REML criterion at convergence: 1380.8

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.9664 -0.6015  0.0048  0.5654  3.3544 

Random effects:
 Groups               Name        Variance Std.Dev.
 school_id:country_id (Intercept) 0.1778   0.4216  
 country_id           (Intercept) 0.6157   0.7847  
 Residual                         0.6873   0.8291  
Number of obs: 493, groups:  school_id:country_id, 362; country_id, 11

Fixed effects:
                       Estimate Std. Error         df t value Pr(>|t|)    
(Intercept)            0.988831   0.245836  10.770860   4.022 0.002094 ** 
bilingual              0.272461   0.105240 478.168073   2.589 0.009920 ** 
factor(SQt01i01)Male  -0.013391   0.084448 472.689597  -0.159 0.874073    
c_age                 -0.112819   0.051080 480.892811  -2.209 0.027667 *  
c_HISEI                0.020762   0.003015 457.328680   6.885 1.91e-11 ***
Z_Parental             0.084831   0.055472 460.024318   1.529 0.126887    
Z_Cultural             0.178451   0.048114 445.238903   3.709 0.000234 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) bilngl f(SQ01 c_age  c_HISE Z_Prnt
bilingual   -0.105                                   
fc(SQ0101)M -0.176  0.038                            
c_age        0.001  0.003 -0.006                     
c_HISEI      0.003 -0.041  0.001  0.048              
Z_Parental   0.014 -0.017 -0.064  0.008 -0.424       
Z_Cultural  -0.024  0.008  0.136  0.028 -0.167 -0.275
> r.squaredGLMM(listening)
           R2m       R2c
[1,] 0.1552958 0.6079135
```

20. Transformations

RR TEAM INSTRUCTIONS: *This section should describe how any of the measured variables or composite measures mentioned above will be transformed prior to the analyses listed in Section 19. These are adjustments made to variables* ***after*** *measurement or measure creation, and might include centering, logging, lagging, rescaling etc. Please provide enough detail such that anyone else could reproduce the transformations based on the description below. Please be sure this preregistration includes a link to a clearly commented script that performs the transformations described in the narrative provided below.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify which of the measured variables or composite measures will need to be transformed prior to the focal analysis?*
- *For each variable needing transformation, does the preregistration adequately describe the transformations, including any centering, logging, lagging, recoding, or implementation of a coding scheme for categorical variables?*
- *Does the preregistration link to a clearly commented script that performs each transformation?*

The following transformations were carried out:

- The independent variable of this study is whether students speak more than 1 language at home. In the replication dataset, there is a variable (**I03_ST_A_S26A**) that measures the number of languages spoken at home: 1, 2 or 3. This variable was re-coded so that if **I03_ST_A_S26A** = 1 (only one language spoken at home) means that the student is monolingual (Bilingual = 0) and if **I03_ST_A_S26A** = 2 or 3 (two or more languages spoken at home), then the student is bilingual (Bilingual = 1)

-The student's performance on the target language (i.e., English) is measured in three dimensions: Reading, Writing and Listening. The dataset includes plausible values and plausible levels, but following the recommendations of the data finder, we used plausible values because they are numeric instead of character variables. Each of the three dimensions are measured through 5 items. Therefore, a global score of each dimension was calculated by obtaining the average across the 5 items (per dimension).

-The variable "Cultural capital" is recorded to convert it to a continuous variable. This variable has six categories:

- "0-10 books" = 0
- "11-25 books" = 1
- "26-100 books" = 2
- "101-200 books" = 3
- "201-500 books" = 4
- "More than 500 books" = 5

-After this point, three datasets are created, one for each dimension of the variable "English Proficiency": Writing, Reading, and Listening. This procedure deviates from the procedure followed by the original paper, where a global average was calculated between measures of reading, vocabulary, grammar and spelling in English. The reason for splitting the dataset was that a check-up analysis showed that, on average, students scored higher in listening and reading compared to writing, and students scored higher in listening compared to reading. Therefore, we decided to analyze these constructs separately, because each student is only measured in two of these three dimensions, which means that students that have scores on *'reading*' and '*listening'* will systematically score higher in english (regardless of the independent variable) than students that have scores on *reading/listening* AND *writing*, biasing the results. The model used to test the difference between these three dimensions was a repeated-measures four-level model (measurements within cases, nested in schools, nested in countries), using as an independent variable the type of dimension used (i.e., writing, reading, listening). The dataset had to be transformed into a long format to carry out this check-up analysis.

-For each of the three datasets (listening, writing and reading), the variables Age and HISEI (SES) were centered.

-For each of the three datasets (listening, writing and reading), Parental education and cultural capital are transformed to Z-scores.

21. Inference criteria

RR TEAM INSTRUCTIONS: *This section describes the precise criteria that will be used to assess whether the hypotheses listed above were confirmed by the analyses in Section 19. The default language below only applies to the test of the SCORE claim,* ***H****. *It is at your discretion to describe the inferential criteria you will use for any additional analyses. They need not rely on p-values and/or the same alpha level we have specified for* ***H****.

If the additional analyses will use multiple comparisons, the inference criteria is a question with few "wrong" answers. In other words, transparency is more important than any specific method of controlling the false discovery rate or false error rate. One may state an intention to report all tests conducted or one may conduct a specific correction procedure; either strategy is acceptable.

Criteria for a successful replication attempt for the SCORE project is a statistically significant effect (alpha = .05, two tailed) in the same pattern as the original study on the focal hypothesis test (**H***). The main outcome that will be directly compared to the main dependent variable of the original paper will be the **composite score of English ability** (average of writing, reading, and listeting measures), so a significant difference between bilinguals and monolinguals in this outcome (alpha = .05, two tailed) in two different regression models will be considered as a successful replication. In the first regression model, only the variable bilinguals-monolingual will be included (as in Model A of Table 3 of the original paper) and in a second model, the control variables (Except for cognitive ability) will be entered in the regression (as in Model B of Table 3). **Therefore, a significant effect of the variable "being bilingual or monolingual" on a composite English score in both an empty model and a model with control variables will be considered as a successful replication.**

However, check-up analyses showed that using a composite score in the replication dataset could induce bias. Therefore, the analyses described above will be also done for each English measure separately. For these analyses, a successful replication will be considered if significant differences are observed between the bilingual and monolingual group in at least one of these three dimensions. Correction for multiple comparisons has to be applied, so it is recommended that alpha is lowered from 0.05 to 0.017.

22. Data exclusion

RR TEAM INSTRUCTIONS: *The section below should describe the rules you will follow to exclude collected cases from the analyses described in Section 19. Note that this refers to exclusions* ***after*** *the creation of the replication dataset; exclusion criteria that prevent a case from entering the replication dataset in the first place should be detailed in the 'Data Collection Procedure' section above. Please be as detailed as possible in describing the rules you will follow (e.g. What is the specific definition of outliers you will use? Exactly how many attention checks does a participant need to fail before their removal from the analytic sample?).*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration comment on whether any cases included in the replication dataset will be excluded prior to data analysis?*
- *If yes, does the preregistration provided detailed instructions on how the exclusions will be performed (e.g. Is the definition of outlier provided? Is the number of attention checks failed before a participant is excluded specified?)*

The following observations were excluded from the analysis:

-Participants who spoke the target language (English) at home. This is measured with the variable **I03_ST_A_S27B = 1**. For this reason, a total of 369 participants were removed from the replication dataset.

-According to section 13.7 of the technical report (that can be found here: https://osf.io/c6y8r/), *"only students and schools that meet the formal criteria for participation have a weight in the datasets".* Therefore, for each separate analysis (i.e., reading, writing, and listening), the participants that did not have weight (i.e., missing value) or that had a weight of 0 were removed from the analysis. The names of the variables that have this information are: **FSW_READ_TR** (for reading), **FSW_WRIT_TR** (for writing), **FSW_LIST_TR** (for listening). This exclusion led to the removal of 8314 participants (for reading), 8298 (for writing), and 2827 participants (for listening)

23. Missing data

RR TEAM INSTRUCTIONS: *The section below should describe how missing or incomplete data will be handled. Please be as detailed as possible in describing the exact procedures you will follow (e.g. last value carried forward; mean imputation) and any software required (e.g. We will use Amelia II in R to perform the imputation).*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration comment on how missing or incomplete data will be addressed (e.g. casewise removal, missing data imputation)?*
- *If applicable, does the preregistration specify how many missing variables will lead to a case's removal (e.g. If a subject does not complete any of the three indices of tastiness, that subject will not be included in the analysis.)?*
- *If applicable, does the preregistration describe how missing data imputation will be performed, including relevant software?*

No imputation of missing data was performed. In the original paper, authors did multiple imputation with MPlus following a theoretical model proposed by Lehmann and Lenkeit, 2008. However, in the replication dataset, no missing data was observed in the dependent variable. All the participants that were supposed to fill in, for instance, the reading and writing dimension, filled in all the items that referred to this dimension. There is not one single piece of data missing in this sense. If we assume that each participant was randomly assigned to fill in two out of the three dimensions of the English ability construct, then the fact that some scores are missing for one of these dimensions should not constitute a problem.

24. Exploratory analysis (Optional)

RR TEAM INSTRUCTIONS: *If you plan to explore your data set to look for unexpected differences or relationships, you may describe those tests here. An exploratory test is any test where a prediction is not made up front, or there are multiple possible tests that you are going to use. A statistically significant finding in an exploratory test is a great way to form a new confirmatory hypothesis, which could be registered at a later time. If any exploratory analyses involve additions to the data collection procedure beyond what was performed in the original study (e.g. additional items on the survey; running another condition in the experiment), please describe them below.*

As an exploratory analysis, we calculate an overall average across the three dimensions of the English proficiency construct (calculating the average of the three average scores of writing, listening and reading). A three-level model was run on this new dependent variable on 5% of the dataset. This analysis resembles the analyses done by the original paper, but due to the way data was collected in the replicated dataset (i.e., only two measures by student), this analysis might lead to biased results.

An additional exploratory analysis consisted in analyzing the items of each dimension directly without calculating an overall effect for each dimension. That is, the score of the 5 items of each dimension were kept, and an additional level of variation was added in the multilevel model, leading to a four-level model: items (Level 1) were nested within cases (Level 2), which were nested within schools (Level 3), which were nested within countries (Level 4). These analyses were done for each dimension separately (one analysis for the items of Writing, another analysis for the items of Reading, and a last analysis for the items of Listening), but also for all the items together, without differentiating between dimensions. This last analysis resembles the first exploratory analysis that was done, in which all items were pooled together, regardless of the type of dimensions they measure.

From these exploratory analyses, it could be concluded that the same results are obtained regardless of whether items are pooled together in an overall score and a three-level model is performed, or whether separate item scores are preserved and analyzed within a four-level model.

25. Other

RR TEAM INSTRUCTIONS: *This section serves two purposes. First, please use this section to discuss any features of your replication plan that are not discussed elsewhere. Literature cited, disclosures of any related work such as replications or work that uses the same data, plans to make your data and materials public, or other context that will be helpful for future readers would be appropriate here. Second, please also re-surface any major deviations from earlier in the preregistration that you expect a reasonable reviewer could flag for concern. Give a* *summary of these deviations, focusing on larger changes and any possible challenges for comparing the results of the original and replication study.*

Specific points to keep in mind (please also consult the [Reviewer Criteria](#)):

- *Does the preregistration reference other sections of the preregistration where substantial deviations from the original study have been described (including deviations due to differences in location or time compared to the original study)?*
- *Does the preregistration comment on plans to make the data and materials from the replication study public?*

We have two major concerns about the potential use of these dataset to replicate the results of Maluch et al. (2015). The first one is that the mean age (and it's standard deviation) of the current dataset deviates from the one of the original study:

Age in the original sample:
Bilingual group
mean = 12.71 (SD=0.70)

Monolingual group:
mean = 12.49 (SD=0.49)

Age in the current sample:
Bilingual group
mean = 15.0 (SD=0.89)

Monolingual group:
mean = 15.0 (SD=0.92)

The current sample is on average older than the original sample, and the age varies more in the current sample (i.e., the standard deviation is lager) than in the original sample. Therefore, we are not sure of whether the sample of this current dataset can be considered similar to the original one.

Secondly, we think that there is a major concern with the data available for this replication that does not allow us to confirm that "*we believe it represents a good-faith replication attempt of the original focal claim*" (the last point of the "Final review checklist").

In the original study, the authors insist on controlling for certain socio-demographic variables, that are very relevant for obtaining the final results. One of these variables is '*cognitive ability*'. Find here some textual sentences of the original study:

-Indeed, once controlling for general cognitive abilities, gender and age, the Turkish-German bilingual group and the Other Bilingual group have, on average, no longer significant disadvantages and similar levels of English achievement as their monolingual peers (page 81).
-All comparisons of the different groups of elementary school students were conducted controlling for sociocultural background factors and general cognitive abilities. (page 82).

Unfortunately, this variable is not included in the replication dataset, and we have serious doubt about the validity of the results if analyses are not controlled by this variable.

Another concern relates to the variable '*proficiency of the instructional language*'. The major finding of the original study is that the relationship between bilingualism and English proficiency is highly moderated by the *proficiency of the instructional language at home.*

- Original hypothesis was: "In the present study, we seek to determine whether there is a relation between immigrant bilingualism and foreign language learning outcomes and to what extent the predicted pattern holds across bilingual groups with different instructional language proficiency and diverse home languages."
-For these reasons, it is useful to pay particular attention to the specific language groups as well as language proficiency to better understand the underlying mechanisms.
- Failing to take these factors into account may be leading to biased conclusions in empirical studies and potentially masking possible advantages for immigrant bilingual groups.
-In addition to the large contrast in sociocultural and linguistic background characteristics, which might account for the varying results in the aforementioned studies, another important factor that can affect foreign language learning is the proficiency level in the language of instruction.

However, this other analysis cannot be replicated because that variable is not included in the replicated dataset. We understand that this is not the focal analysis and that is why we see this one less of a concern. Nonetheless, we think it was worth noticing these issues here.

Final review checklist

REVIEWER INSTRUCTIONS: *For the following questions, reviewers please indicate whether you can 'sign off' on the following items by adding a comment. You can update this response as the lab moves through revisions during the review period!*

- Included in this pre-registration are specific materials needed to create a replication dataset:
    - Is the final replication dataset that the research team constructed suitable for performing a high-quality, good-faith replication of the focal claim selected from the original study?
    - Is the procedure for constructing the final replication dataset sufficiently documented that an independent researcher could construct the same dataset following the procedures and code they lay out?
- Included with this pre-registration is a narrative description of how the replication dataset will be used to perform the focal replication analysis, as well as the specific analytic scripts/code/syntax that will be used:
    - Is the analysis plan (including code) that's documented in the preregistration consistent with a high-quality, good-faith replication of the focal claim selected from the original study?
    - Has the data analyst demonstrated that the analysis code works as expected on a random 5% of the final replication dataset?
- I have reviewed all sections of this pre-registration, and I believe it represents a good-faith replication attempt of the original focal claim.