Replication attempt to evaluate a claim from Maluch_LearnInst_2015_5Awm

Replication team: Belén Fernández-Castilla, David Santos.

SCORE RR ID: 2y2g

OSF Project: https://osf.io/k6xm9

Instructions

For each claim that you are evaluating in this replication attempt, please complete all sections below. Please feel free to remove these and the other instructions for the final report that you provide on OSF.

Each claim evaluation **must** include a binary determination of whether or not the claim was replicated, given the inferential criteria. In other words, a sentence along the lines of "This claim [was/was not] replicated." should appear in every claim evaluation. You can (and should!) also articulate in your discussion why this interpretation may be limited, but programmatic requirements of the SCORE project at large include the binary decision so it must be present.

There are also specific requirements for reporting the results of inferential tests:

- Any p-value should be reported as an exact value, not a threshold, e.g."p = .047" rather than "p < .05". Please also indicate whether it is from a one- or two-tailed test.
- Report the full text of your results with as much precision as possible. For example, a satisfactory report for the focal coefficient might be "unstandardized OLS coefficient for *income* term = 0.342367" vs an unsatisfactory report of "beta = 0.34".
- Please verify that for any test that was used as the basis for a power analysis, you are reporting the same effect size type used in that power analysis. It may not be the same type of effect size as was reported in the original paper (e.g., we convert to Cohen's f squared to power many regression analyses, which is not commonly reported).
- For any test that does not have a corresponding power analysis, please provide an effect size that corresponds to one from the original paper to the extent possible.

- In the case of a regression analysis, please report the most unstandardized coefficient possible, e.g., the log-odds instead of the odds ratio. You may also report standardized values.

- If there are multiple cells/conditions or levels to your analysis, please provide the sample sizes for each cell or level, e.g., “234823 participants nested within 32 countries” or “angry condition n = 83, happy condition n = 85”.

- If multiple claims are evaluated using the same model, it is ok to group them together. In that case, please ensure that it is still clear which specific test/coefficient corresponds to which claim ID, and that there is still an evaluation of whether each claim replicated or not.

- Scientific notation is acceptable for especially small or large values.

These requirements apply to all “simple” tests (one inferential test expected to be significant) and any inferential test contained within a “complex” test of a claim.

If applicable, please also distinguish which results come from an analysis of (a) entirely new data, (b) “hybrid” data (composed of both data that was used in the original study and new data), and/or (c) entirely original data.

Description of generalizability

[Please add any other details to describe how this study departs from the original study, and how it speaks to the original claim(s).]

-The nationality of the sample is different. In the original study, they used a sample of 2,835 German 6th-graders (Arabic-German: *n* = 105, Chinese-German: n = 110, Polish-German: n = 57, Turkish-German: n = 383, heterogeneous bilingual: n = 284, and monolingual German group: n = 1896). In the replication study, we used a sample of 20,026 students from 14 differente participating countries (Belgium, Bulgaria, Croatia, England, Estonia, France, Greece, Malta, Netherlands, Poland, Portugal, Slovenia, Spain, Sweden). The claim that the original authors make might only apply to German participants, and not to participants from other countries.

-In the original study, authors assessed English ability with a different instrument. Specifically, original autors used the Cloze test (consisting of four texts with 91 word completion questions measuring reading proficiency, vocabulary, grammar and spelling simultaneously). In contrast, the proposed replication dataset includes measurements of students listening, reading and writing abilities taken from The European Survey on Language Competences (ESLC) from 2012. In the replication, a composite English ability score between these three dimensions was calculated. This composite score might be biased because it is an average of three dimensions (students' listening, reading and writing abilities). In fact, check up analyses carried out by the replication team found that students that responded to the wirting ability test scored, on average, significantly lower than those who responded to the reading and listening skills. That is, it seems that these three scales might measure different dimensions of the construct “English ability”. In this regard, check up analyses were performed, and the main claim was tested for each category of English ability.

-The average age in the replication sample is statistically larger than the mean age of the original sample. The mean age in the original study was 12.71 for the bilingual group (*SD* = 0.70, 95% CI [12.87, 12.75]) and 12.49 for the monolingual group (*SD* = 0.49, 95% CI [12.47, 12.51). In contrast, in the replication sample, the mean age for the bilingual group was 15 (*SD* = 0.89, 95% CI [14.99, 15.01]) and for the monolingual group was also 15 (*SD* = 0.92, 95% CI [14.97, 15.03]). The 95% confidence intervals show that the mean age of the replication sample was significantly larger than the man age of the original sample. This difference in the mean age could lead to substantial differences in the findings of the original and replication study.

-In the original study, authors used as a control variable a measure of cognitive ability. Specifically, they mentioned “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. TFor cognitive ability, the replication study does not have this measure.” (page 79 in the original paper). There is no such a measure in the replication database to control for. Thus, this is another difference between the original study and the replication study. Not controlling for this variable might lead to differences in the findings from the original and replication study.

Claim evaluation

Single-trace claim

Coded claim 4 text (original paper): “…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).”

Replication outcome: Simple test

Inferential criteria: 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) will be considered as a successful replication in the model where control variables are included (replicating Model B of the original paper).

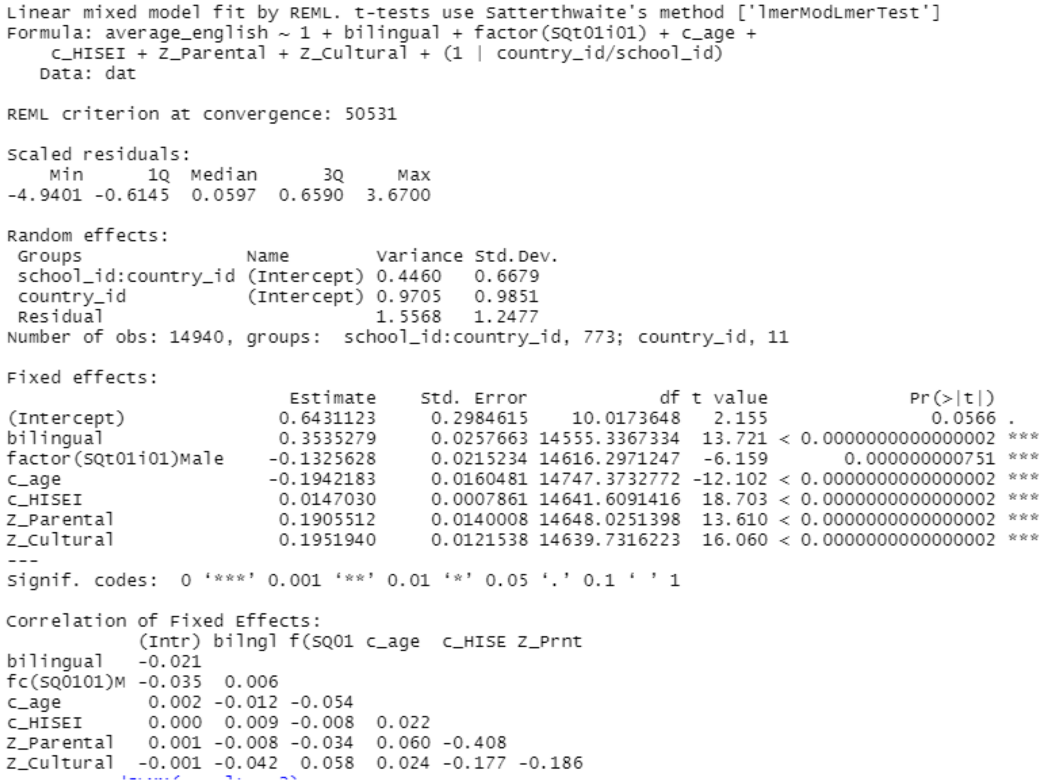
To assess the main claim, a three-level model was fitted with the R package *lmerTest*, with intercept. The dependent variable was a composite English score (average between writing, reading and listening scores). The independent variables were: being bilingual or not (bilingual), gender (SQt01i01), age (c_age), HISEI (c_HISEI), parent qualification (Z_Parental) and cultural capital (Z_Cultural) (please note that cognitive ability was missing in the replication dataset). All these independent variables were introduced as fixed effects. Observations from students were nested within schools and within countries. To have this nestig into account, the factors country_id and school_id were specified as random-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. The regression coefficiens (B) reported below are unstandardized, and they refer to the effect of the foca independent variable “being bilingual or not”. Standardized regression coefficients are also reported for this variable with the symbol ß (the function to calculate them can be found in the R script). The R^2 statistic reported below refers to a *Pseudo-R-squared for Generalized Mixed-Effect* model (*r.squared* function from package *MuMIn*). Specifically, the R^2 statistic reported here refers to the marginal R^2, which is the percentage of variance explained only by the fixed effects in the regression model. The tests were two-sided.

Result:

Main claim: A significant difference between bilinguals and monolinguals in this outcome (a composite English score, alpha = .05, two tailed) will be considered as a successful replication in the model where control variables are included (replicating Model B of the original paper).

With 16185 students, nested in 773 schools and 11 countries, we found that, when controlling for sociodemographic variables, the bilingual group scored significantly higher in the composite English score than the non-bilingual group (B = 0.35, SE = 0.03, t(14555) =13.72, *p* = 2.2e-16, R^2 = 9.6%, ß = 0.08). **Therefore, the main claim of the original paper is replicated.**

Results from the model that includes control variables:



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

REML criterion at convergence: 50531

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.9401 -0.6145  0.0597  0.6590  3.6700

Random effects:
 Groups               Name        Variance Std.Dev.
 school_id:country_id (Intercept) 0.4460   0.6679
 country_id           (Intercept) 0.9705   0.9851
 Residual                         1.5568   1.2477
Number of obs: 14940, groups:  school_id:country_id, 773; country_id, 11

Fixed effects:
                       Estimate    Std. Error            df t value             Pr(>|t|)
(Intercept)           0.6431123     0.2984615    10.0173648   2.155               0.0566 .
bilingual             0.3535279     0.0257663 14555.3367334  13.721 < 0.0000000000000002 ***
factor(SQt01i01)Male -0.1325628     0.0215234 14616.2971247  -6.159       0.000000000751 ***
c_age                -0.1942183     0.0160481 14747.3732772 -12.102 < 0.0000000000000002 ***
c_HISEI               0.0147030     0.0007861 14641.6091416  18.703 < 0.0000000000000002 ***
Z_Parental            0.1905512     0.0140008 14648.0251398  13.610 < 0.0000000000000002 ***
Z_Cultural            0.1951940     0.0121538 14639.7316223  16.060 < 0.0000000000000002 ***
---
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   -0.021
fc(SQ0101)M -0.035  0.006
c_age        0.002 -0.012 -0.054
c_HISEI      0.000  0.009 -0.008  0.022
Z_Parental   0.001 -0.008 -0.034  0.060 -0.408
Z_Cultural  -0.001 -0.042  0.058  0.024 -0.177 -0.186
```

Deviations from the preregistration: We have followed the plan stipulated in the pre-registration

Discussion: The main claim of Maluch et al. (2015) study is replicated in this replication study: when controlling for sociodemographic variables (even if cognitive ability is missing), bilingual participantes scored higher in English ability than monolingual individuals. However, the R^2 we have found for this model (R^2 = .096) is less than half of that R^2 reported by the original authors (R^2= .26). A possible explanatiton is that in our replication study, the variable *cognitive ability* was not avaiable, and that variable might explain that piece of variance that we failed to explain with our model.

Additional analyses (optional)

Instructions: If you performed any additional analyses, please report them here. Otherwise you may delete this section.

The same set of three level models were carried out separately for each English outcomes (writing, reading and listening), and the main claim was replicated with all of these three outcomes: when controlling for sociodemographic variables, the English **reading** skills of the bilinguals participants were significantly higher than in the group of monolingual participants (B= 0.23, SE = 0.04, t(4773) = 6.49, *p* = 9.28e-11 , R^2 = 11.74%; ß = 0.07), and the same occured with the English **listening** skills (B= 0.26, SE = 0.02, t(9528) = 12.88, *p* = 2.2e-16, R^2 = 8.8%; ß = 0.08,) and the English **writting** skills (B= 0.57, SE = 0.05, t(9631) = 11.46, *p* = 2.2e-16, R^2= 10.7%; ß = 0.08).

General discussion (optional)

Instructions: If you would like to make any broad comments about this replication attempt, please do so here. Otherwise you may delete this section.

In terms of statistical significance, we consider that our replication study succesfully replicates the original finding of Maluch et al. (2015). We have replicated the finding of Maluch et al. (2015) using different measures of English ability: first a composite score, and then three different dimensions: writing, reading and listening skills. The results were very similar across dependente variables, which could be an indicator of the robustness of the results.

However, both the original and the replication study have a very large sample size, making it very likely that any small effect becomes statistically relevant even if, in reality, it is not. In this sense, the R^2 statistics found in this replication study are smaller than those reported by the original authors. Furthermore, the lack of standardized regression coeffients in the original study did not allow us to compare whether the maginute of the effects are comparable across the original and replication study.

Description of materials provided

Instructions: Please detail the materials that will be available on OSF or another repository from this project. This section should describe both what is available and whether it can be shared publicly or not. If all files can be shared publicly, it is alright to include a general sharing statement to the effect of “All materials on this OSF project may be shared publicly.” Otherwise, please indicate sharing permissions for each file.

For any materials that will not be shared on OSF or another repository by you, and are instead available through other means, please include a description of how someone else might access those materials.

We recommend including an entry for each file that indicates what that file is and its intended use. For example:

- experiment_script.txt - the wording used by the experimenter for consent, instructions, and debriefing

If you have many files which are similar, e.g., stimuli from an experiment, datasets from the same source, etc., it is reasonable to describe them generally instead of individually, including any guidance on naming conventions and a tally of such files so others can verify they have all expected files. For example:

- condition_num.png - There are 60 images (15 per condition) used as stimuli, labeled by condition and ordered by *num* which has a base-0 index, e.g., the 4th image in the “angry” condition is named “angry_03.png”

The minimum requirements for materials include:

1. **Analysis pipeline:** This may take the form of a script (or scripts) that process and then analyze your data, and/or *detailed* instructions for how these steps are accomplished if you do them manually (e.g., in Excel or via the GUI of your chosen statistical software).

2. **Full results/output:** Please provide the full output from your analysis, preferably with comments that identify which claim is being evaluated by each test. Depending on your chosen software, this may be one of a variety of file types. You may provide the default, but please also provide a version that is non-proprietary so that it may be viewed by someone who does not have your given software. Txt, PDF, markdown, and HTML are some common options that are not reliant on proprietary statistical software.

3. **Data:** All versions of your data should be provided. For all projects, the "raw", i.e., earliest tabular version of your data should be available, and if other versions are generated during your analysis pipeline they should also be shared. At minimum, your data should be shared in a non-proprietary file-type such as a csv or tsv; you may also share versions that correspond to your chosen statistical software such as .sav or .dta. It may not always be possible to share data on OSF directly if there are any ethical or legal constraints on sharing, e.g., if it is existing data that are proprietary or include sensitive participant information. In those cases, please provide instructions for how another person might access the data.

4. **Data dictionary:** This is also sometimes called a codebook in some fields. This file (or files) should describe every variable in your dataset. This guide provides more information about how to make one. For projects which rely on existing data, it is possible that a codebook already exists for that data from the source. It is satisfactory to either provide that file again if its license/permissions allow redistribution, or to provide guidance on how to access it if that is not the case.

5. **Study materials:** There is not a one-size-fits-all requirement to describe the other materials that are relevant to every project, but you should also share whatever materials another researcher would need to run your study in their own lab. We may make specific requests during report review if we believe additional materials could be provided that would improve the transparency of your project.

DOCUMENTS:

- **Codebook Students.pdf** - This is a pdf containing the labels of all the variables and a description of each variable.
- **EU Data Dictionary.updated.xlsx** - This is en Excel file containing the variable, the name, the measurement unit, the allowed values, and the description of the variable
- **EU Data Dictionary.xlsx** - This is en Excel file containing the variable, the name, the measurement unit, the allowed values, and the description of the variable
- **language-survey-technical-report_en.pdf** - This is a pdf containing the technical report by the EU regarding the European Survey on Language Competences (ESLC).
- **Preregistration from.pdf** - Pre-registration of the replication study.

- **SCORE Report - Maluch_LearnInst_2015_5Awm - Fernández - 2y2g.docx.** Analysis to test the replication claim and conclusions and discussion.

DATABASE:

- **Final replication dataset.rds.** Database used for the replication.

R CODE:

- **MALUCH-code (1).R** - This is the code to create the final dataset for analysis.
- **MALUCH.updated.code.R** - This is the code that was used to clean and debug the database
- **Preregis_Analysis_5percent.R** - This is the code to test whether the models properly run for the pre-registration.
- **Preregis_Check up analysis.R** - This is the code where we tested whether there were statistical differences between the mean scores of the three dimensions (namely writing, reading, listening).
- **Replication attempt code (FINAL).R** - R code to reproduce the final results of the replication.

IMAGES:

- **Results1-Writing_5percent.png** - This is a picture with the outcome of the linear mixed model for writing using the 5% of the sample in the analysis
- **Results2-Reading_5percent.png** - This is a picture with the outcome of the linear mixed model for reading using the 5% of the sample in the analysis
- **Results3-Listening_5percent.png -** This is a picture with the outcome of the linear mixed model for listening using the 5% of the sample in the analysis

References

Instructions: At minimum, include a full citation of the original study. If you are using existing data for your replication, also cite the source(s) of that data. Literature reviews are not required, but may be reflected here as well.

European Commission. (2012). *First European survey on language competences*. Publications Office of the European Union.

Maluch, J. T., Kempert, S., Neumann, M., & Stanat, P. (2015). The effect of speaking a minority language at home on foreign language learning. *Learning and instruction*, *36*, 76-85.