Replication of a Research Claim from Seaton et al. (2010),
from the *American Educational Research Journal*

Replication Team: Carolin Nast and James Field

Research Scientist: Melissa Kline

Action Editor: John Lloyd

Independent Reviewers
(add name below when you initiate review, comment “DONE” on your name when you finish):

Reviewer #1: Zorana Zupan

Reviewer #2: [NAME]

Reviewer #3: [NAME]

Review Period: August 12 - August 17

View-only links to: Original Paper, Replication Data, Replication Analysis

Privacy Statement: Other teams are making predictions about the outcomes of many different studies, not knowing which studies have been selected for replication. As a consequence, the success of this project requires full confidentiality of this peer review process. This includes privacy about which studies have been selected for replication and all aspects of the discussion about these replication designs.

Instructions for Data Analysts

The preregistration for this replication study was started by a separate team of researchers who were responsible for identifying data sources and constructing them into a replication dataset(s) for your use in the analysis. They have completed sections 1-13 of the preregistration below, and included additional materials in the OSF project that document how the dataset was constructed.

In cases where all of the underlying data sources were able to be freely shared and posted, the constructed dataset(s) have been posted to the OSF as well, which you are free to use in designing the analysis plan (see below for details). In cases where some or all of the data sources could *not* be freely shared or posted, the replication dataset(s) are not provided on the OSF. Rather, you will need to follow the instructions and code to first reconstruct the datasets, and then proceed with your work. In such cases, the team responsible for creating the dataset(s) has provided summary statistics in the OSF that correspond to the constructed datasets, so you can verify that the datasets you create match what they intended.

You'll be responsible for filling out sections 16-25 of the preregistration below. Before you do so, **please review the original study, sections 1-15 of the preregistration, and the materials provided on the OSF**, so that you are familiar with all of the decisions that have been made to date. In many cases, the 'data preparer' will have left you instructions and suggestions on how the provided data can be used in the analysis, as well as idiosyncrasies and discrepancies in the data that you should be aware of. The data preparers have tried to be thorough in including all variables that you might need, but please keep in mind the following:

- Some of the variables included in the constructed dataset(s) may not be needed in the final analysis, so please do not feel the need to necessarily use all of the provided variables.
- Some of the variables needed might have mistakenly been excluded from the constructed datasets. If you find that this is the case, please let Andrew or Anna know, and they will work with you to supplement the datasets as needed.

For these secondary data replications, we would like the analysis plan to be completed before the preregistration goes through review, so that after review, the only remaining steps are registration and running the analysis code on the full datasets. To facilitate that, we are asking that you include in section 19 a link to the code you will use that takes the constructed dataset(s) provided to you and produces the focal analysis (including all of the cleaning, merging, and transforming required). When developing your analysis plan and code, please randomly sample 5% of the data for use in your work, and **do not use the rest of the data until it is time to run the final analysis**. In section 19, you will find a statement that we are asking you to bold that confirms you've only used 5% of the data when developing and testing your code. If this approach will not work for any reason, please let Andrew or Anna know and disclose deviations from this plan somewhere in the preregistration.

- In cases where we are providing you a complete dataset, you can just sample out 5% of the observations and hold the rest out until you are ready to perform the final analysis.
- In cases where we are providing you multiple datasets that need to be combined prior to analysis, please sample out 5% of the observations in whatever way is most sensible.
    - For example, in cases where each dataset contains complete observations on its own (a typical 'row bind' situation), it makes the most sense to sample out 5% of each dataset separately and then combine them together to develop and test your code.
    - In cases where datasets need to be merged in order to create complete observations (a typical 'column bind' situation), it makes the most sense to merge the separate datasets

into a full dataset first, and then sample out the 5% before proceeding with the rest of the analysis code.

- We leave the decision on how to sample out the random subset of data to you, so long as (a) you are not performing any analyses on the complete dataset until after your study is registered and (b) whatever decision you make is documented in the preregistration.

Finally, in cases where the replication data combines observations from the original study with observations that were not used in the original study (what we are calling 'hybrid replications'), please perform two analyses (details immediately below). This will likely require you to subset your data into the two groups described immediately below, based on the description of the original analysis provided in the study.

- When the 'new' data alone can clear the minimum power threshold, please perform one analysis that combines all available data, and a second that only uses the 'new' data. Please make sure both analyses are documented (with code) in section 19 below.
- When the 'new' data alone *cannot* clear the minimum power threshold, please perform one analysis that combines all available data, and a second that only uses the old data. Please make sure both analyses are documented (with code) in section 19 below.

Please contact Andrew or Anna if you have any questions. After you've completed the remaining sections of the preregistration and uploaded all the necessary materials to the OSF, please contact the SCORE coordinators regarding next steps.

Preregistration of Seaton_AmEduResJourn_2010_Blxd

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 Seaton et al. (2010) in *American Educational Research Journal*.

2. Authors and affiliations

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

RR LAB LEAD[1]
Carolin Nast[2] (data finder)
James G. Field[3] (data analyst)
TEAM MEMBER B[1]

1 Affiliation 1
2 Student, Utrecht University, Utrecht, The Netherlands
3 Assistant Professor, West Virginia University, USA

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 Seaton et al. (2010) is that larger BFLPE (big-fish-little-pond) effects were associated with students who used memorization to a greater extent. Students in high-ability schools who used the memorization technique to a greater extent suffered a larger decline in mathematics self-concept than those who used the technique to a lesser extent. Although all students who used memorization had lower mathematics self-concepts if they attended high-ability schools than students of similar memorization usage who attended average- or low-ability schools, the drop in mathematics self-concept was more pronounced for students who used memorization to a greater extent. This reflects the following

statement from the paper's abstract: "Statistically significant moderating effects emerged in both areas; however, in relation to the large sample (N = 265,180), many were considered small." The claim was tested by regressing mathematical self-concept on predictor variables were individual mathematics ability (linear and quadratic), school-average mathematics ability, the moderator of interest - use of memorization studying, and the interaction of school-average mathematics ability with memorization. Due to the large number of tests of statistical significance being conducted, the significance level was set at $p < .001$. Effect size for the group-level construct of BFLPE was calculated using the equation shown on p. 406; effect sizes that approached 0.10 or greater were considered of interest. Use of memorization moderated the effect of school-average ability on mathematical self-concept (-.089, $p < 0.001$; see Table 3), with an effect size of -0.157 (see Table 4).

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*: The interaction of the use of memorization and school-average ability will be negative in its association with mathematical self-concept.

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 was involved 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 original study utilizes the Programme for International Student Assessment (PISA) by the OECD from 2003, which had a focus on students' mathematical abilities. Participants were 15-year-old students from 41 countries (N = 276,165). PISA assesses students' mathematics, literacy, and science, but PISA also contains a vast amount of additional information about students' families, home backgrounds, and SES. Students also report their learning habits, how they feel about their achievements, and how they connect with other students and teachers within their school. From the full PISA data, students were excluded who did not complete the math self-concept items, as well as schools with 10 or fewer students. This led to a sample of 265,180 students who attended 10,221 schools in 41 countries. I propose to replicate this study with a newer PISA wave, namely PISA 2012. This is the most recent wave with the necessary focus on mathematical abilities. The main analysis is a multilevel model considering the individual, the school and the country level. It is assumed that the focal claim is mostly interested in the moderation effect between memorization (independent variable on the individual level) and school-average ability (independent variable on the school level) on the mathematical self-concept (dependent variable). It should be noted that two constructs, (Cooperative Orientation and Competitive Orientation) were officially deleted from the PISA student questionnaire after 2003 and were not yet adequately replaced[1]. Therefore, a replication attempt would need to be followed through without these two constructs. This could have

[1] Source: Page 191, https://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf

implications on the replication attempt, as these independent variables were significant, however, their standardized effect sizes are considered very small (< 0.1). Furthermore, given the fact that these constructs were deleted due to potential invalidity, the results from the original study concerning these independent variables might also be not totally understood/valid. Consequently, it is expected that the non-availability of these independent variables for the replication attempt is minor due to potential invalid previous results.

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 PISA, especially not by the data finder.

For further information: OECD 2014, Technical report, particularly p. 59 (student questionnaires), p. 68 (student selection), p.233ff., p. 260ff. (mathematics test). Link: https://osf.io/2m9ga/

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. The PISA study 2012 has been chosen for replication, because it is the most recent wave with a focus on mathematical abilities. The final dataset includes all variables to conduct the main analysis. The main analysis is a multilevel model considering the individual, the school and the country level. The selected variables in the final dataset are the same ones that were used in the original study. Two variables included in the original study are not included in the final dataset as they are not included in the PISA study anymore.

The replication dataset can be found here: https://osf.io/ym4g5/

Full disclosure statement from Data Analyst: It is important to note that the Data Analyst (i.e., the individual responsible for preparing the analytic script for the replication of the focal claim) made an error when preparing the analytic script for the replication of the focal claim. Specifically, the Data Analyst inadvertently ran a portion of the final analytic script on the *entire* replication data set – not just a 5% random sample of the replication data set, which they were instructed to do. Consequently, the Data Analyst observed the results derived from the set of multilevel modeling regression analyses. Taken together, the Data Analyst inadvertently performed Steps 1-12 (see statistical analysis and procedure information in Section 19). Before realizing their error, the Data Analyst attempted to model the final parameter estimate and corresponding standard error and *p*-value. The Data Analyst realized that their calculations for these results were incorrect *after* they realized that the full data set was being used, not just a random 5% subset. The Data Analyst immediately corrected their mistake upon identifying the error (i.e., deleted all results and began working on the analytic script using just a 5% random sample of the entire data set). Following this, the Data Analyst examined additional sources cited in the original article (Seaton et al., 2010) and identified the correct procedures for estimating the final parameter, standard error, and *p*-value. Although this is non-ideal, it is important to note that the Data Analyst *did not* run the entire final analytic script on the full data set. In other words, the Data Analyst identified their error before the correct procedures for estimating the final parameter estimate and corresponding standard error and *p*-value were identified and added to the analytic script. As such, although the Data Analyst observed the results from the set of separate multilevel modeling regression analyses, they *did not* observe the final estimate parameter, standard error, and *p*-value that are calculated using the correct analytic procedures. This means that the Data Analyst does not know if the focal claim will replicate. A copy of the original analytic script (i.e., the one that failed to create a random 5%

subset and uses the incorrect procedures for estimating the final estimates) can be found at the project website [see filename “Seaton_AmEduResJourn_2010_Blxd_beta.R” at https://osf.io/mu4rs/ or to see the script directly, please visit https://osf.io/7t9g4/]. [End of full disclosure statement]

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:

- *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.*

The analysis of the original study entails about 90 raw items from the PISA dataset 2003 (Student questionnaire) which comprise 18 constructs (1 dependent variable, 16 independent variables on the individual level, 1 independent variable on the school level) which are utilized in

the focal analysis. The original study mainly used pre-defined and pre-calculated indices by PISA to comprise their constructs. These constructs are available in the raw data as well.

For all pre-defined and pre-calculated PISA indices only their construct acronym will be listed here, and not every item used for calculating the index. Three indices are pre-defined by PISA, but not pre-calculated. Furthermore, it should be noted that two constructs, (Cooperative Orientation and Competitive Orientation) were officially deleted from the PISA student questionnaire after 2003 and were not yet adequately replaced. Therefore, a replication attempt would need to be followed through without these two constructs. The original study derived significant effects for both of these independent variables, however their standardized effect size is considered tiny (< 0.1).

The main source for the acronyms and the predefined indices is the PISA technical report 2012: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf → (OECD Technical report, 2014).

As mentioned above, the necessary variables correspond to the same wording and acronyms between 2003 and 2012.

Dependent Variable

Mathematics self-concept

- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (SCMAT)
- Measurement direction: A high score was associated with a higher mathematics self-concept.
- Source: OECD Technical report (2014), p.323, Table 16.14

Independent variables on the student level

(1) Individual ability

- Variable Data Source: PISA
- Waves / Years: 2003
- Variable: Individual PISA scores for mathematical ability (PV1MATH - PV5MATH)
- Variable explanation: The PISA database does not contain a single mathematics ability measure. Rather, it provides five plausible values to estimate a student's academic ability, which avoids biased population estimates being obtained.
- Units: PISA scores
- Operationalization: The PISA documentation advises researchers not to average these plausible values but to conduct analyses with each plausible value separately and then average all resulting parameters (see OECD, 2005a). This was the course of action followed in the original study.

- Measurement direction: A high score was associated with a higher individual mathematics ability.
- Source: OECD Technical report (2014), p.143-163, Chapter 9 on Scaling PISA Cognitive Data, especially p. 146-148

Socioeconomic Status (SES)
(2) Highest parental occupation
- Variable Data Source: PISA
- Waves / Years: 2003
- Variable: The parental occupation measure was based on the higher occupational status: either the mother's or the father's.
- Number of items: 2

(3) Highest in education
- Variable Data Source: PISA
- Waves / Years: 2003
- Variable: The parental education measure was based on whichever educational level was higher, the mother's or father's.
- Number of items: 2

(4) Home educational resources
- Variable Data Source: PISA
- Waves / Years: 2003
- Number of items: 5
- Sample item: 5 (Sample Item: "In your home do you have books to help with your schoolwork?")

(5) Cultural possessions
- Variable Data Source: PISA
- Waves / Years: 2003
- Number of items: 3
- Sample Item: "In your home do you have books of poetry?"

Academic self-regulation (4 dimensions) - study methods, motive, behavior, social dimension

Study methods: 3 constructs
(6) Control Strategies
- Variable Data Source: PISA
- Waves / Years: 2003

- Pre-defined and pre-calculated index by PISA (CSTRAT)
- Measurement direction: A high score was associated with a higher preference for this learning strategy (control strategies).
- Source: OECD Technical report (2005), p.295. Table 17.18

(7) Memorization
- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (MEMOR)
- Measurement direction: A high score was associated with a higher preference for this learning strategy (memorization).
- Source: OECD Technical report (2005), p.296. Table 17.20

(8) Elaboration
- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (ELAB)
- Measurement direction: A high score was associated with a higher preference for this learning strategy (elaboration).
- Source: OECD Technical report (2005), p.295. Table 17.19

Motive: 3 constructs

(9) Extrinsic
- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (INSTMOT)
- Measurement direction: A high score was associated with higher extrinsic motivation.
- Source: OECD Technical report (2014), p.322, Table 16.10

(10) Intrinsic
- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (INTMAT)
- Measurement direction: A high score was associated with higher intrinsic motivation.
- Source: OECD Technical report (2014), p.321, Table 16.09

(11) Math Self-Efficacy
- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (MATHEFF)
- Measurement direction: A high score was associated with a higher level of confidence.
- Source: OECD Technical report (2014), p.322, Table 16.12

Behavior: 1 construct

(12) Math Anxiety

- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (ANXMAT)
- Measurement direction: A high score was associated with a higher level of anxiety.
- Source: OECD Technical report (2014), p.323, Table 16.13

Social dimension: 4 variables

(13) Cooperative Orientation

- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (COOPLRN)
- Measurement direction: A high score was associated with a higher preference for co-operative learning situations.
- Source: OECD Technical report (2005), p.298, Table 17.24
- Further note: This index and its items are missing from 2003 on.

(14) Competitive Orientation

- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (COMPLRN)
- Measurement direction: A high score was associated with a higher preference for competitive learning situations.
- Source: OECD Technical report (2005), p.298, Table 17.23
- Further note: This index and its items are missing from 2003 on.

(15) Student-Teacher Relations

- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (STUDREL)
- Measurement direction: A high score was associated with a higher student's perception of teachers' interest in student performance.
- Source: OECD Technical report (2014), p.333, Table 16.36

(16) Sense of Belonging

- Variable Data Source: PISA
- Waves / Years: 2003
- Pre-defined and pre-calculated index by PISA (BELONG)
- Measurement direction: A high score was associated with a higher sense of belonging.
- Source: OECD Technical report (2014), p.334, Table 16.37

- Further note: It is indicated that next to the 6 relevant items included in 2003, 3 additional items were asked in 2012 for the construct. If trend analysis is being conducted, PISA suggests to only use the six overlapping items.

Independent variable on the school level

School-average mathematics ability

- Variable Data Source: PISA
- Waves / Years: 2003
- Variable explanation: A school-average mathematics ability variable was calculated for each plausible value by averaging each one separately within each school.
- Operationalization: School-average mathematics ability variable was calculated for each plausible value by averaging each one separately within each school. This is a typical psychometric analysis workflow (described in more detail in section 12d). This school-average mathematics ability variable was not restandardized, thus keeping all variables in the same metric as the individual test scores and in a metric that was consistent across all schools and countries.
- Source: Seaton et al. (2010), p. 404

Cluster variables

Country ID

- Variable Data Source: PISA
- Waves / Years: 2003
- Acronym: CNT

School ID

- Variable Data Source: PISA
- Waves / Years: 2003
- Acronym: SCHOOLID

Student ID

- Variable Data Source: PISA
- Waves / Years: 2003
- Acronym: STIDSTD

Final student weight

- Variable Data Source: PISA
- Waves / Years: 2003
- Acronym: W_FSTUWT

Sample size of analysis has 265,180 observations.

(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.

The data source for replication is the same as in the original study, however a different wave will be utilized. PISA 2012 is freely available for research purposes and offers a very user-friendly platform. The data file and the codebook can be successfully opened. No registration is needed to download the necessary data.

Analyses done on PISA 2012 can be done freely for any kind of institution and research. For citations in academic journals the following should be used:
“OECD (2012), (dataset name),(data source) DOI or URL (accessed on (date))”

(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.***

As the proposed replication dataset - PISA 2012 - is a subsequent wave of the original study’s data - PISA 2003. The replication data set can be found at the following project websiteL

https://osf.io/mu4rs/. All variables listed above are available in the same form and with the same technical acronyms for the proposed new wave for the replication dataset. Two constructs, as mentioned in section 12(a), are not available in any forthcoming PISA wave anymore, as they were discarded due to low quality results. It is assumed that these constructs cannot be replaced/built with different items as PISA discarded the whole concept of the constructs *cooperative and competitive orientation*. However, these independent variables were not mentioned in the focal claim.

Dependent Variable (1)

Mathematics self-concept

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (SCMAT)
- Measurement direction: A high score was associated with a higher mathematics self-concept.
- Source: OECD Technical report (2014), p.323, Table 16.14

Independent variables on the student level (16)

(1) Individual ability

- Variable Data Source: PISA
- Waves / Years: 2012
- Variable: Individual PISA scores for mathematical ability (PV1MATH - PV5MATH)
- Variable explanation: The PISA database does not contain a single mathematics ability measure. Rather, it provides five plausible values to estimate a student's academic ability, which avoids biased population estimates being obtained.
- Units: PISA scores
- Proposed operationalization: The PISA documentation advises researchers not to average these plausible values but to conduct analyses with each plausible value separately and then average all resulting parameters (see section 12(d)). This was the course of action followed in the original study.
- Measurement direction: A high score was associated with a higher individual mathematics ability.
- Source: OECD Technical report (2014), p.143-163, Chapter 9 on Scaling PISA Cognitive Data, especially p. 146-148

Socioeconomic Status (SES)

(2) Highest parental occupation (HISEI)

- Variable Data Source: PISA
- Waves / Years: 2012
- Measurement direction: Higher values indicate higher occupational status.
- Source: OECD Technical report (2014), p.307

(3) Highest in education (HISCED)

- Variable Data Source: PISA
- Waves / Years: 2012
- Measurement direction: Higher values indicate higher levels of education.
- Source: OECD Technical report (2014), p.307

(4) Home educational resources (HEDRES)

- Variable Data Source: PISA
- Waves / Years: 2012
- Measurement direction: Higher values indicate more educational resources.
- Source: OECD Technical report (2014), p.316

(5) Cultural possessions (CULTPOS)

- Variable Data Source: PISA
- Waves / Years: 2012
- Measurement direction: Higher values indicate more cultural possessions.
- Source: OECD Technical report (2014), p.316

(Alternative: Pre-defined and pre-calculated index of Socioeconomic Status (ESCS)

- Variable Data Source: PISA
- Waves / Years: 2012
- Measurement direction: A high score was associated with a higher economic, social and cultural status.
- Source: OECD Technical report (2014), p.352

Academic self-regulation: 4 dimensions, namely study methods, motive, behavior, social dimension

Study methods: 3 constructs

(6) Control Strategies

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined, but not pre-calculated index by PISA
- Index calculated by the data finder (CSTRAT) based on four variables (ST53Q01-ST53Q04), however raw variables are also included in the final dataset. For further explanation see 12(d).
- Measurement direction: A high score was associated with a higher preference for this learning strategy (control strategies).
- Source for index building: OECD Education Working Papers No. 130, p. 77

(7) Memorization

- Variable Data Source: PISA

- Waves / Years: 2012
- Pre-defined, but not pre-calculated index by PISA
- Index calculated by the data finder (MEMOR) based on four variables (ST53Q01-ST53Q04), however raw variables are also included in the final dataset. For further explanation see 12(d).
- Measurement direction: A high score was associated with a higher preference for this learning strategy (memorization).
- Source for index building: OECD Education Working Papers No. 130, p. 78

(8) Elaboration

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined, but not pre-calculated index by PISA
- Index calculated by the data finder (ELAB) based on four variables (ST53Q01-ST53Q04) however raw variables are also included in the final dataset. For further explanation see 12(d).
- Measurement direction: A high score was associated with a higher preference for this learning strategy (elaboration).
- Source for index building: OECD Education Working Papers No. 130, p. 76

Motive: 3 constructs

(9) Extrinsic

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (INSTMOT)
- Measurement direction: A high score was associated with higher extrinsic motivation.
- Source: OECD Technical report (2014), p.322, Table 16.10

(10) Intrinsic

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (INTMAT)
- Measurement direction: A high score was associated with higher intrinsic motivation.
- Source: OECD Technical report (2014), p.321, Table 16.09

(11) Math Self-Efficacy

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (MATHEFF)
- Measurement direction: A high score was associated with a higher level of confidence.
- Source: OECD Technical report (2014), p.322, Table 16.12

Behavior: 1 construct

(12) Math Anxiety

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (ANXMAT)
- Measurement direction: A high score was associated with a higher level of anxiety.
- Source: OECD Technical report (2014), p.323, Table 16.13

Social dimension: 4 variables

(13) Cooperative Orientation

- This index and its items are missing from 2003 on.

(14) Competitive Orientation

- This index and its items are missing from 2003 on.

(15) Student-Teacher Relations

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (STUDREL)
- Measurement direction: A high score was associated with a higher student's perception of teachers' interest in student performance.
- Source: OECD Technical report (2014), p.333, Table 16.36

(16) Sense of Belonging

- Variable Data Source: PISA
- Waves / Years: 2012
- Pre-defined and pre-calculated index by PISA (BELONG), including 9 items
- Measurement direction: A high score was associated with a higher sense of belonging.
- Source: OECD Technical report (2014), p.334, Table 16.37
- Further note: It is indicated that next to the 6 relevant items included in 2003, 3 additional items were asked in 2012 for the construct. If trend analysis is being conducted, PISA suggests to only use the six overlapping items.

Independent variables on the school level (1)

School-average mathematics ability

- Variable Data Source: PISA
- Waves / Years: 2012
- Variable explanation: A school-average mathematics ability variable was calculated for each plausible value by averaging each one separately within each school. Given that there are five plausible values, five school-average mathematics ability scores were estimated.
- Proposed operationalization: Calculate school-average mathematics ability variable for each plausible value by averaging each one separately within each school.
- Source: Seaton et al. (2010), p. 404

Cluster variables

Country ID

- Variable Data Source: PISA
- Waves / Years: 2012
- Acronym: CNT

School ID

- Variable Data Source: PISA
- Waves / Years: 2012
- Acronym: SCHOOLID

Student ID

- Variable Data Source: PISA
- Waves / Years: 2012
- Acronym: STIDSTD

Final student weight

- Variable Data Source: PISA
- Waves / Years: 2012
- Acronym: W_FSTUWT

(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.

When combining multiple datasets by binding rows, please be sure that the data type and measurement units are equivalent across each dataset. If there is a discrepancy in how a variable is measured across datasets, rename the variable in each dataset to indicate the original dataset, and then carefully document the resulting measures below and in the data dictionary. See here for an example of how this should work.

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/nydeg/. The final replication data set can be found by following this direct link: https://osf.io/ym4g5/. Alternatively, one can visit the project website (see https://osf.io/mu4rs/) and click on the filename "PISA2012.replication.RDS."

The most important thing the data analyst needs to understand is the use of plausible values to construct the independent variable on the school level, school average mathematics ability. The construction of this index needs to be implemented in the analysis workflow. It might be helpful to read (1) the operationalization by the authors of the original study (see 12a), (2) the provided document on handling plausible values (here: https://osf.io/bzynd/), as well as (3) Chapter 9 of the OECD Technical report (2014) (here: https://osf.io/2m9ga/).

Another important point concerns the definition and construction of the indices CSTRAT, MEMOR and ELAB, which have been defined in the OECD Education Working Papers No. 130 (here: https://osf.io/e3sgh/) and have been constructed by the data finder as explained in the script. The raw variables for indices construction are included in the final replication dataset.

Moreover, the data analyst has to decide if she/he wants to use the four original variables of the socio-economic status construct, namely highest parent education, highest parent occupation, educational resources and cultural possessions, or alternatively use a predefined compound of socio-economic status (ESCS) by PISA (see 12c). This variable (ESCS) was implemented by the data finder into the replication dataset.

Lastly, the data analyst has to pay attention to the fact that the unique school identification has to consist of two variables, which together form a unique identifier for each school, namely CNT (country ID) and SCHOOLID (see OECD Technical report, 2014, p. 400).

(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/w9jh5/

For NAs, missings and invalid values the data dictionary mentions 997, 998 and 999 for all variables whose NAs, missing, and invalid values begin with “99”. This is because these variables always have some variation of these numbers as NAs, missings and invalid values, e.g. 9997 or even 99997 etc.

The complete codebook concerning the student questionnaire from PISA 2012 can be found here: https://osf.io/r4nge/

The student questionnaire form A can be found here: https://osf.io/62acy/

The OECD Technical report (2005) concerning PISA 2003 can be found here: https://osf.io/n3hbr/

The OECD Technical report (2014) concerning PISA 2012 can be found here: https://osf.io/2m9ga/

The OECD Education Working Papers No. 130 concerning the constructing of the indices CSTRAT, MEMOR and ELAB can be found here: https://osf.io/e3sgh/

Chapter 6 on Plausible Values from the PISA Data Analysis Manual: SPSS, Second Edition can be found here: https://osf.io/bzynd/

The raw PISA 2012 student data can be found here:https://osf.io/uk5z7/

The SAS Variable helper file can be found here: https://osf.io/37puy/

The code used for the replication dataset can be found here: https://osf.io/nydeg/

The replication dataset can be found here: https://osf.io/ym4g5/

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:

The proposed replication dataset includes one descending wave of the same data source, PISA. Therefore the sample parameters are the same and can be perfectly replicated.

The initial replication dataset contains 480,174 observations, nested in 18,139 schools and 65 countries. After deleting observations nested in schools with 10 or less (see page 403 of original study; Seaton et al. 2010) observations the dataset contains 468,803 observations, nested in 16,046 schools and 65 countries.

On page 403 of the original study (Seaton et al., 2010), it was noted that students who did not complete any of the math self-concept items (i.e., the dependent variable) were removed from further analyses. In addition, it was noted that only 1% missingness remained following this step. However, the original study did not disclose how the remaining missingness was dealt with. As such, to provide what will most likely be the most conservative estimates, it is proposed that listwise deletion is performed on the entire data set of focal variables.

Required sample size [to be filled out by the SCORE team]: The primary unit of analysis is the school. An estimate of the minimum viable sample size for the data analytic replication is: 972. For comparison, the stage1 required sample size would be: 4,711 and the stage2 sample size would be: 10,619.

Notes: With the exception of sample size, these analyses assume that the replication is identical to the original data in every factor influencing power. In this case, it is recommended that the replication team consider the following [this list of factors to consider may not be exhaustive.]:

- Whether all 16 effects can be included

- Whether the required number of school units are feasible

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/u8bxz/?view_only=67c747fbf46f405da2101aa3be20029d

15. Stopping rule (provided by SCORE)

RR TEAM INSTRUCTIONS: *Because all existing data replications that clear SCORE's minimum power threshold will proceed to analysis, the stopping rule is not relevant for these kinds of projects.*

N/A -- all observations will be used in a single analysis.

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.

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?*

As previously mentioned in Section 12, the necessary variables correspond to the same wording and acronyms between 2003 and 2012 (i.e., the original [Seaton et al., 2010] and replication [PISA, 2012] data sets, respectively). As such, given that there are no deviations between the measures across both data sets, in this section we only describe the variables used in the original analysis of the focal claim as well as the role they played.

VARIABLE #1: Country of Origin (Coded as "CNT" in the original and replication data sets)

- Description of variable: This variable denotes the country in which the respondent was located at the time of data collection.
- Purpose of variable in focal claim analysis: This variable did not play a formal role (e.g., dependent variable, moderator) in the test of the focal claim. However, this variable is used to account for the nested nature of the data (i.e., students nested within schools nested within countries) and, thus, used to model random-effects in a multi-level model.

VARIABLE #2: School ID (Coded as "SCHOOLID" in the original and replication data sets)

- Description of variable: This variable, when combined with the respondent's country of origin (more information provided in Section 18), denotes the school in which the responded was located at the time of data collection.
- Purpose of variable in focal claim analysis: Information provided by this variable was used to calculate the multiplicative term that tested the proposed moderating effect in the focal claim. Specifically, information provided by this variable was used to estimate each school's average plausible value scores (i.e., each school had an average score for each of the five measures of math ability), which were combined with the respondent's memorization score to create the five cross-products that were used in separate multilevel models. In addition, this variable is used to account for the nested nature of the data (i.e., students nested within schools nested within countries) and, thus, used to model random-effects in a multi-level model.

VARIABLE #3: Final Student Weight (Coded as "W_STUWT" in the original and replication data sets)

- Description of variable: According to the OCED Technical Report (2005; see page 114), this variable represents "the final weight variable on the data file … which is the final

student weight that incorporates any student-level trimming." In addition, the OCED Technical Report (2005; see page 324) states that "the sum of the weights constitutes an estimate of the size of the target population, i.e. the number of 15-year-old students in grade 7 or above attending school in that country."
- Purpose of variable in focal claim analysis: This variable did not play a formal role (e.g., dependent variable, moderator) in the test of the focal claim. However, this variable should be used in multilevel analysis (see OCED, 2005, p. 325).

VARIABLE #4: Mathematics Self-Concept (Coded as "SCMAT" in the original and replication data sets)

- Description of variable: According to Seaton et al. (2010, p. 392), academic self-concept can be defined as "defined as one's knowledge and perceptions about one's academic ability." As such, it can be assumed that mathematics self-concept can be defined as one's knowledge and perceptions about one's mathematics ability. Importantly, an examination of the sample items listed in Seaton et al. (2010; see page 404) indicate that the same items were used to measure mathematics self-concept in the original and replication data sets (OCED, 2012, see Table 16.14 on page 323).
- Purpose of variable in focal claim analysis: Dependent variable.

VARIABLE #6: Final Student Weight (Coded as "W_FSTUWT" in the original and replication data sets)

- Description of variable: A normalized weight used to ensure that larger schools and countries did not bias the observed results.
- Purpose of variable in focal claim analysis: Weight applied in each multilevel modeling regression analysis.

VARIABLES #6-10: Plausible Values 1-5 (Coded as "PV1MATH" – "PV5MATH" in the original and replication data sets)

- Description of variable: These variables were indicators of an individual's mathematical ability.
- Purpose of variable in focal claim analysis: These variables played two roles in the focal claim analysis. First, these variables were independent variables. Second, these variables were used to estimate each schools' math ability averages (i.e., each school had an average PV1MATH score, PV2MATH score, and so on), which, in turn, were combined with the respective respondent's memorization score to create the required cross-product term (e.g., PV1MATH*MEMOR) for each multilevel analysis.

VARIABLE #11: Memorization (Coded as "MEMOR" in the original and replication data sets)

- Description of variable: According to the original article (Seaton et al., 2010), memorization is a cognitive learning strategy by which students tend to rehearse material and learn by rote (see p. 397).
- Purpose of variable in focal claim analysis: Memorization was the moderating variable in the original analysis of the focal claim.

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?*

COMPOSITE #1 – UNIQUE SCHOOL ID: Country (coded as "CNT") and school ID (coded as "SCHOOL ID") information were combined to create unique school IDs. This important step was performed so that the replication analyses properly accounted for the nested nature of the replication data set.

COMPOSITES #2-6 – SCHOOL AVERAGES: A school-average mathematics ability variable was calculated for each plausible value by averaging each one separately within each school.

COMPOSITES #7-11 – CROSS PRODUCTS WITH SCHOOL AVERAGE ABILITY: Given that separate multilevel model analyses had to be performed for each plausible value (i.e., indicator of math ability), five cross-product terms were created. Each cross-product term was comprised of the individual's memorization score and the respective school's plausible value average score.

COMPOSITES #12-16 – QUADRATIC TERMS: On page 405 of the original study (Seaton et al., 2010), it was written that "for each moderator analysis the fixed components were individual ability (both linear and quadratic), the specific moderator, school-average ability, and the cross-product of the moderator and school-average ability." As such, quadratic terms were created for all five plausible values provided in the replication data set.

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 (a screenshot of the results is preferable). 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?*

The following disclosure statement, which is pasted from a previous section, explains an error made by the Data Analyst when preparing the analytic script for the replication analyses.

Full disclosure statement (repeated from Section 11): It is important to note that the Data Analyst (i.e., the individual responsible for preparing the analytic script for the replication of the focal claim) made an error when preparing the analytic script for the replication of the focal claim. Specifically, the Data Analyst inadvertently ran a portion of the final analytic script on the *entire* replication data set – not just a 5% random sample of the replication data set, which they were instructed to do. Consequently, the Data Analyst observed the results derived from the set of multilevel modeling regression analyses. Taken together, the Data Analyst inadvertently performed Steps 1-12 (see statistical analysis and procedure information in [the following sections]). Before realizing their error, the Data Analyst attempted to model the final parameter estimate and corresponding standard error and *p*-value. The Data Analyst realized that their calculations for these results were incorrect *after* they realized that the full data set was being used, not just a random 5% subset. The Data Analyst immediately corrected their mistake upon identifying the error (i.e., deleted all results and began working on the analytic script using just a 5% random sample of the entire data set). Following this, the Data Analyst examined additional sources cited in the original article (Seaton et al., 2010) and identified the correct procedures for estimating the final parameter, standard error, and *p*-value. Although this is non-ideal, it is important to note that the Data Analyst *did not* run the entire final analytic script on the full data set. In other words, the Data Analyst identified their error before the correct procedures for estimating the final parameter estimate and corresponding standard error and *p*-value were identified and added to the analytic script. As such, although the Data Analysts observed the results from the set of separate multilevel modeling regression analyses, they *did not* observe the final estimate parameter, standard error, and *p*-value that are calculated using the correct analytic procedures. This means that the Data Analysts does not know if the focal claim will replicate. A copy of the original analytic script (i.e., the one that failed to create a random 5% subset and uses the incorrect procedures for estimating the final estimates) can be found at the project website [see filename “Seaton_AmEduResJourn_2010_Blxd_beta.R” at https://osf.io/mu4rs/ or to see the script directly, please visit https://osf.io/7t9g4/]. [End of full disclosure statement]

Details of the statistical analysis and procedure employed in the original study are provided on pages 404-406 of Seaton et al. (2010). The replication analysis follows the exact same procedure. To better understand how the final parameter, standard error, and p-value estimates were calculated, the replication team gathered information from the PISA Data Analysis Manual, which can be found at the project webpage (https://osf.io/mu4rs/). In the statistical analysis and procedure description that follows, we try to be clear when describing when the PISA Data Analysis Manual had to be referenced. The script that performs the following steps, which are intended to replicate the focal analysis claim and follows the correct procedures, can be found here: https://osf.io/7t9g4/. Importantly, the script has been tested by the Data Analyst and does not present any errors.

Step 1: Delete observations nested in schools with less than 10 individual students

- On page 403 of the original study (Seaton et al., 2010), it was noted that, to maintain comparability with previous studies using the PISA database, schools with 10 or fewer students were deleted from further analyses. As such, the first step is to remove from the data set schools with 10 or fewer students.

Step 2: Subset data set to include focal variables only

- Given that the focal analysis pertains to a subset of variables and the number of total observations is very large (n = ~480,000), it is suggested that the second step should be to remove nonessential variables. Put differently, following this step, only the variables needed to perform the replication of the focal claim will remain in the data set.

Step 3: Remove missing data

- On page 403 of the original study (Seaton et al., 2010), it was noted that students who did not complete any of the math self-concept items (i.e., the dependent variable) were removed from further analyses. In addition, it was noted that only 1% missingness remained following this step. However, the original study did not disclose how the remaining missingness was dealt with. As such, to provide what will most likely be the most conservative estimates, it is proposed that the second step in the analysis procedure should be to perform listwise deletion on the entire data set of focal variables.

Step 4: Standardize focal variables

- On page 404 of the original article (Seaton et al., 2010), the original authors stated that "the five plausible values for mathematics ability, mathematics self-concept, and all the potential moderators were standardized (M = 1, SD = 0)." As such, the third step is to standardize the following variables so that their corresponding mean and standard deviation were 1 and 0, respectively: (1) plausible value #1, (2) plausible value #2, (3) plausible value #3, (4) plausible value #4, (5) plausible value #5, (6) mathematics self-concept, and (7) memorization

Step 5: Calculate school average for each plausible value

- On page 404 of the original article (Seaton et al., 2010), the original authors stated that "a school-average mathematics ability variable was calculated for each plausible value by averaging each one separately within." As such, the fourth step is to estimate each school's average plausible value scores.

Step 6: Calculate cross-products

- On pages 404-405 of the original article (Seaton et al., 2010), the original authors stated that “Cross-products with school-average ability were created for each potential moderator.” As such, given that the moderator in the focal claim is memorization, the fifth step is to create the following five product terms for each school: (1) (average plausibility value #1 score * individual memorization score), (2) (average plausibility value # score * individual memorization score), (3) (average plausibility value #3 score * individual memorization score), (4) (average plausibility value #4 score * individual memorization score), (5) (average plausibility value #5 score * individual memorization score)
- Note that per what was written on page 404 of the original article these cross-products were not restandardized.

Step 7: Calculate quadratic terms

- On page 405 of the original article (Seaton et al., 2010), the original authors stated that each predictor variable’s quadratic term is included in the analyses. As such, the sixth step is to create the following five quadratic terms: (1) PV1MATH^2, (2) PV2MATH^2, (3) PV3MATH^2, (4) PV4MATH^2, and (5) PV5MATH^2

Steps 8-12: Perform a set of multilevel modeling analyses

- On page 405 of the original article (Seaton et al., 2010), the original authors stated that “a separate set of five multilevel regression analyses (one for each plausible value for achievement) was conducted for each moderator of the BFLPE.” This practice is supported by information provided in the PISA Data Analysis Manual and previous OECD Technical Reports. As such, five multilevel models are estimated, one for each plausible value and the corresponding school averages and cross-product.
- It is important to note that each multilevel model regression analysis is weighted by the final student weight variable labeled “W_FSTUWT.” This is done because the 2005 OCED Technical Report, which refers to the PISA 2003 (i.e., the data set used in the original study [Seaton et al., 2010]), states the following:

> “The variable W_FSTUWT is the final student weight. The sum of these weights constitutes an estimate of the size of the target population. When analysing weighted data at the international level, large countries have a greater contribution to the results than small countries. This weighting is used for the OECD total in the tables of the international report for the first results from PISA 2012 (OECD, 2014). To weight all countries equally for a summary statistic, the OECD average is computed and reported. The OECD average is computed as follows. First, the statistic of interest is computed for each OECD country using the final student weights.”

Step 13: Estimate focal analysis final parameter

- To estimate the final parameter – or in other words, the coefficient for the focal claim (i.e., the cross-product term involving memorization and plausible value] – we intend to take the average of the five cross-product terms derived from the set of individual multilevel analyses. This practice follows exactly what was stated on page 405 of the original paper ("parameter estimates were averaged") and the second step of the six-step process to calculating final estimates presented on pages 120-121 of the PISA Data Analysis Manual. We note that this six-step process is described in other places throughout the manual as well.

Step 14: Estimate focal analysis final standard error

- To estimate the final standard error estimate we follow exactly the process outlined on pages 120-121 of the PISA Data Analysis Manual. We note that this six-step process is described in other places throughout the manual as well.
- Specifically, we followed this process:
    - (1) We estimate the required parameters and respective standard error for each plausible value. Put differently, we extract the cross-product parameters and corresponding standard error from the set of five multilevel modeling regression analyses.
    - (2) We estimate the final parameter estimate (i.e., the average of the five cross-product parameters)
    - (3) We estimate the final sampling variance (i.e., sum of the squared standard errors divided by number of plausible values)
    - (4) We estimate the imputation variance (i.e., sum of the squared parameter differences divided by the number of plausible values minus one)
    - (5) We estimate the final error variance using the formula provided on page 121 of the PISA Data Analysis Manual ([result provided by 3] * (1.2[result provided by 4]
    - (6) We estimate the final standard error by taking the square root of the result provided by the previous step.

Step 15: Estimate focal analysis final *p*-value

- To the best of our knowledge, the original article (as well as the corresponding journal webpage, corresponding first author's websites, and PISA Data Analysis Manual) did not describe how final *p*-values should be estimated. Put differently, it was not made clear how to derive a single *p*-value for the focal claim from the set of five separate multilevel analysis.
- We intend to follow Fischer's combined probability test method (see https://en.wikipedia.org/wiki/Fisher%27s_method) to combine the *p*-values produced by the five multilevel modeling regression analyses. This method will produce a single *p*-value that can be used to examine the replicability of the original focal analysis result/claim.

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?*

Aligned with what was stated on page 404 of the original article (Seaton et al., 2010), the five plausible values, mathematics self-concept, and the moderator of interest (i.e., memorization) were standardized so that the respective M = 1 and SD = 0.

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 hypothesis to be tested is: The interaction of the use of memorization and school-average ability will be negative in its association with mathematical self-concept. As such, the inference criteria that will be used to assess whether or not the focal claim replicates will be (1) the direction of the final parameter estimate (needs to be negative) and (2) the magnitude of the final *p*-value (needs to be less than less than .001 [see page 405 of Seaton et al., 2010]. Both conditions need to be met in order for the focal claim to replicate.

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?)*

Variables not needed for reassessment of the focal claim are removed from the replication analysis. The following variables are included in the replication data set

1. Country
2. School ID
3. Student ID
4. Final student weight
5. Mathematics self-concept
6. Highest parental occupation
7. Highest in education
8. Home educational resources

9. Cultural possessions
10. Socioeconomic Status (SES)
11. Learning Strategies - Important Parts vs. Existing Knowledge vs. Learn by Heart
12. Learning Strategies - Improve Understanding vs. New Ways vs. Memory
13. Learning Strategies - Other Subjects vs. Learning Goals vs. Rehearse Problems
14. Learning Strategies - Repeat Examples vs. Everyday Applications vs. More Information
15. Motive, Extrinsic
16. Motive, Intrinsic
17. Motive, Math Self-Efficacy
18. Behavior, Math Anxiety
19. Social dimension, Student-Teacher Relations
20. Social dimension, Sense of Belonging
21. Individual ability, Plausible value 1
22. Individual ability, Plausible value 2
23. Individual ability, Plausible value 3
24. Individual ability, Plausible value 4
25. Individual ability, Plausible value 5
26. Study methods, Control Strategies
27. Study methods, Memorization
28. Study method, Elaboration

However, irrelevant variables are removed, meaning only the following variables are used in the replication attempt:

1. Country ID
2. School ID

3. Student ID
4. Final student weight
5. Mathematics self-concept
6. Individual ability, plausible value 1
7. Individual ability, plausible value 2
8. Individual ability, plausible value 3
9. Individual ability, plausible value 4
10. Individual ability, plausible value 5
11. Study methods, Memorization

The linked script (see https://osf.io/mu4rs/), which performs all replication analyses, removes the irrelevant data.

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?*

The original article stated that observations with missing data in the dependent variable were removed from the data set (Seaton et al., 2010, p. 403). Following this, 1% missingness was observed in the remaining variables (Seaton et al., 2010, p. 403). However, the authors did not make clear how they handled the remaining missingness. As such, it is proposed that listwise deletion be performed on the data set of focal variables so that a database without any missing data is used for the replication analyses. This recommended practice should not present sample size/statistical power concerns given the large size of the replication data set (n = ~480,000).

The linked script (see https://osf.io/mu4rs/), which performs all replication analyses, removes the irrelevant data.

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.*

Exploratory analyses will not be conducted

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?*

NA

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.