Replication of a Research Claim from Kim et al. (2016),
from *Social Science & Medicine*

Replication Team: Anna Abatayo, Andrew Tyner, and Esteban Méndez-Chacón

Research Scientist: Andrew Tyner

Action Editor: Rich Lucas

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

Reviewer #1: Michael Mullarkey[NAME]

Reviewer #2: [NAME]

Reviewer #3: [NAME]

Review Period: August 12 - August 17

View-only links to: Original Paper, Original Materials, 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 Kim_SocSciMed_2016_AqDO

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 Kim & Radoias (2016), in *Social Science & Medicine*.

2. Authors and affiliations

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

RR LAB LEAD[1]
Anna Abatayo[2]
Andrew Tyner[2]
Esteban Méndez-Chacón[3]

1 Affiliation 1
2 Center for Open Science, Charlottesville, VA
3 Central Bank of Costa Rica

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 Kim & Radoias (2012) is that, for the specific case of asymptomatic disease detection, education should have a clear positive effect for individuals in poor health status (and at the aggregate level), and a smaller (possibly zero or negative) effect for individuals in good health status; the positive effect for individuals in poor health status is the portion of this claim selected for the SCORE program. This reflects the following statement from the paper's abstract: "In terms of disease detection, more educated respondents have a higher probability of being diagnosed, but only conditional on being in poor general health." Evidence in support of the claim is found in Table 2, which contains the probit regression results for the determinants of hypertension under-diagnosis. The dependent variable is a dummy equal to one for those respondents who were found to be hypertensive during the IFLS [Indonesian Family Life Survey] screenings, but were not previously diagnosed by a doctor. The three separate

columns represent, in order, the results for the entire sample, the results for the subsample consisting of respondents in good general health, and the results for the subsample consisting of respondents in poor general health. For the SCORE program, the analysis of the subsample consisting of respondents in poor general health is selected. The predictor of interest is Years of Education (see the right column of Table 2 for details of the model). Education matters for these people, as more educated persons generally have higher opportunity costs of feeling sick and hence value their health higher, which pushes them harder to look for a cure in a doctor's office (marginal effect for Years of Education = -0.00867, SE = 0.00420, significant at 5% level).

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*: Among the sample of respondents in poor general health who were found to be hypertensive during a screening, the probability of being undiagnosed decreases with education.

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 secondary data collectors’ 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 relied on data from the Indonesian Family Life Survey (IFLS) [https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS.html], which is a longitudinal study that involves a combination of new and existing panel members in each wave. As the authors explain, they specifically used the fourth wave in the original study, since that wave included a module on elicited risk and time preference parameters (p. 17). The fifth wave of the IFLS (IFLS 5) [https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/ifls5.html] contains the core variables needed to replicate the claim, and as such was selected for this replication study.

Though IFLS is a panel study, only wave 5 will be used in this replication. This mirrors the analysis of the original study (based on wave 4), where respondents were analyzed without data or reference from earlier waves of the IFLS. In this way, many of the respondents in IFLS 5 are the same as in the original study, though the focal measures defining the dependent variable, independent variable, and sample parameters could have changed since IFLS 4. Also, roughly 19% of the respondents included in the replication dataset are from a 'new household' (as opposed to an existing panel household).

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.*

The focal independent variable (education) was not randomized, but the ordering of one of the modules used to construct two of the control variables ('Time Preference' and 'Risk Preference') was randomized into two levels (A and B). Since this small randomization is unlikely to be relevant to the replication's validity, it will not be further discussed here.

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.*

Data have already been downloaded, merged, and cleaned from IFLS 5. Variables were selected based on their expected relevance to the replication analysis, as determined from the description in the codebook and questionnaire and the values present in the data. None of the variables were selected because of their likelihood (or not) of leading to a confirmatory result.

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.*

IFLS Wave 4 was used for every variable in the original study.

Dependent Variable(s)

Hypertension under-diagnosis

- "Our dependent variable is a dummy equal to one for those respondents who were found to be hypertensive during the IFLS screenings, but were not previously diagnosed by a doctor." (p. 18)

- This variable is defined specifically for those found to be hypertensive via the blood pressure screening *and* self-report poor general health; as such, the sample parameter variables below are needed in order to define this dependent variable for the focal analysis.

Focal Independent Variable(s)

Education

- "Explanatory variables include respondents' education (measured in years of formal education)..." (p. 18)

Control Variable(s)

Age/age-squared

- "Explanatory variables include...respondents' age and age squared (to allow for possible non-linear effects)" (p. 18)

Risk preference

- "Explanatory variables include...respondents' individual risk and time preferences" (p. 18)

Time preference

- "Explanatory variables include...respondents' individual risk and time preferences" (p. 18)

Distance to health center

- "Explanatory variables include...the distance from the closest health center (to proxy for the ease of access to medical care)" (p. 18)

Household per capita expenditures (PCE)

- "Explanatory variables include...household per capita expenditures (PCE)...Per capita expenditures is used here as a proxy for household income." (p. 18)

Sex

- "Explanatory variables include...a sex dummy." (p. 18)

Sample Parameters

Hypertensive

- "Our sample consists of 4209 hypertensive adults...As part of the survey, trained nurses measured respondents' blood pressure three different times. The first measurement was dropped because many people get nervous at first which can cause false high measurements. We then used the average of the other two measurements to construct the hypertension variable. **Following WHO standards, a person is considered**

hypertensive if his systolic is greater than 140 or his diastolic is greater than 90." (p. 18).

General health assessment

- "Respondents were asked to evaluate their general health status (GHS) on a scale from 1 to 4. Depending on the answers provided, we split the sample in two groups: a healthy group containing respondents who characterized their general health status as being either 'very healthy' or 'somewhat healthy', and an unhealthy group containing respondents who claimed they were either 'unhealthy' or 'somewhat unhealthy'." (p. 18)

Sample size of analysis has 1064 observations (respondents).

(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.

As explained in the Study Design section above, the Indonesian Family Life Survey Wave 5 (IFLS Wave 5) is being used to replicate the focal claim. Accessing IFLS 5 data required registration on the RAND website [https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/access.html]. Two of the relevant terms included with this registration are as follows: "You will not distribute the IFLS Public Use Data files to others. If you plan to work with other people using these data, you will ask them to register or register them yourself. If you are a data librarian, you will ask users to register if they obtain a copy of the data from you...You will acknowledge the IFLS as the source of the data for analysis in all reports and publications based on these data. Desired citations for each wave are found on the IFLS data download page."

Approval of the registration was almost instantaneous. A link to access the data was sent soon after registration.

The IFLS data are split across many different files, corresponding to different sections of the larger survey (as detailed in the codebook [https://www.rand.org/pubs/working_papers/WR1143z3.html]). Each file that was downloaded (in dta format) was included in a distinct folder. To construct the replication dataset documented below, all files were downloaded on 7/16/20 *except* the files in the hh14_trk_dta folder, which were downloaded on 7/23/20. The dta files needed, and the folders they are found in, are as follows:

- hh14_b1_dta folder
    - b1_ks1.dta
    - b1_ks2.dta
    - b1_ks3.dta
    - b1_ks0.dta
- hh14_b3a_dta folder
    - b3a_dl1.dta
    - b3a_dl2.dta
    - b3a_cov.dta
    - b3a_si.dta
- hh14_b3b_dta folder
    - b3b_cd3.dta
    - b3b_rj2.dta
    - b3b_kk1.dta
- hh14_bk_dta folder
    - bk_ar1.dta
    - bk_ar0.dta
- hh14_bus_dta folder
    - bus_us.dta
- hh14_trk_dta folder
    - htrack.dta
    - ptrack.dta

(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.***

Hypertension [under-diagnosis]

Variable name: cd05

- File name: b3b_cd3.dta
- Folder: hh14_b3b_dta
- Description: Diagnosed with chronic condition (Y/N) [“Have a doctor/paramedic ever told you that you had [...] ?”]

Variable name: cdtype

- File name: b3b_cd3.dta
- Folder: hh14_b3b_dta
- Description: Chronic condition type (all equal to ‘A’)
- Additional notes: This variable identifies the chronic condition that the value of cd05 corresponds to. There are many different chronic conditions in the original data; for the replication data, only hypertension (cdtype = A) was retained.

Hypertensive

From the original study: "Our sample consists of 4209 hypertensive adults...As part of the survey, trained nurses measured respondents' blood pressure three different times. The first measurement was dropped because many people get nervous at first which can cause false high measurements. We then used the average of the other two measurements to construct the hypertension variable. **Following WHO standards, a person is considered hypertensive if his systolic is greater than 140 or his diastolic is greater than 90.**” (p. 18).

Variable name: us07b1

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Blood pressure - systolic (2nd measurement)

Variable name: us07b2

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Blood pressure - diastolic (2nd measurement)

Variable name: us07bx

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Blood pressure (2nd measurement) (able to measure)

Variable name: us07c1

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Blood pressure - systolic (3rd measurement)

Variable name: us07c2

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Blood pressure - diastolic (3rd measurement)

Variable name: us07cx

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Blood pressure (3rd measurement) (able to measure)

Additional blood pressure measurements

The following variables were also included in the replication dataset in case the analyst wants to use them, though there is not a clear purpose for them as far as the data finders can tell.

Variable name: us07bp

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Pulse (2nd measurement)

Variable name: us07b_1

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Left or right arm
- Notes: Though not explicit in the questionnaire or codebook, this is presumably a reference to the second blood pressure measurement.

Variable name: us07cp

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Pulse (3rd measurement)

Variable name: us07c_1

- File name: bus_us.dta
- Folder: hh14_bus_dta
- Description: Left or right arm
- Notes: Though not explicit in the questionnaire or codebook, this is presumably a reference to the third blood pressure measurement.

General health assessment

Variable name: kk01

- File name: b3b_kk1.dta
- Folder: hh14_b3b_dta
- Description: Generally how is your health?
- Notes: Analytic sample is limited to respondents who report poor general health (3:Somewhat unhealthy; 4:Very unhealthy).

Education

Preliminary Notes

- In conjunction, dl06 and ar16 are the most promising education variables identified in the IFLS 5 data, though they should be cleaned in concert with data from dl04, dl05b, and dl07. As documented below, dl06 and ar16 are very similar, but they each have advantages and disadvantages. It's probably best to create a new variable based on data from each, rather than relying on either in isolation. Details are provided below.
- The original study suggests education should be measured in years; ar16 and dl06, by contrast, are categorical variables, with some portions of the response categories being ordinal. As such, the final education variable used for the replication analysis will need to be carefully cleaned to convert the raw measures into an ordered variable.
- Below the main discussion, additional education variables have been provided, as well, in case those are useful to the data analyst, but they are not recommended for use in the replication analysis.
- Finally, it is recommended that the data analyst consult the categorization of school types (dl2type) listed under dl10 in the questionnaire (page 82 of the pdf) when cleaning the final education variable for the replication analysis. This categorization has been recreated here for reference:
    - Elementary:
        - Elementary 02
        - Adult Education A 11
        - School for Disabled 17
        - Madrasah Elementary 72
        - Other 95
    - Junior High
        - Junior high general 03
        - Junior high vocational 04
        - Adult Education B 12
        - School for Disabled 17
        - Madrasah Junior High School 73
        - Other 95

- Senior High
    - Senior high general 05
    - Senior high vocational 06
    - Adult Education C 15
    - School for Disabled 17
    - Madrasah Senior High School 74
    - Other 95
- D1, D2, D3//University
    - College (D1, D2, D3) 60
    - University (BA) 61
    - University (MA) 62
    - University (PhD) 63
    - Open University 13
    - Other 95

- **Note:** The following values present in dl06 are not categorized into the four levels above, but should be incorporated in some way by the data analyst if dl06 is used:
    - 14:Islamic School (pesantren) [78 observations]
    - 90:Kindergarten [1 observation]
    - 98:Don't Know [11 observations]
    - 99:MISSING [2 observations]

Variable name: dl06

- File name: b3a_dl1.dta
- Folder: hh14_b3a_dta
- Description: Highest level of education attended
- Notes: This is one of the closest measures found in IFLS 5 to a ‘years of education’ variable. As noted immediately below it largely overlaps with ar16. It’s main advantage relative to ar16 is that it contains fewer responses of 98 [Don’t Know]. It’s main disadvantage relative to ar16 is that it does not contain a valid response for people without any schooling.

Variable name: ar16

- File name: bk_ar1.dta
- Folder: hh14_bk_dta
- Description: HHM highest level of education [provided during listing of household members]
- Notes: ar16 is quite similar to the dl06 variable documented above.
    - When ar16 is merged into the dataset containing dl06 and directly compared, 29,946 of the 34,464 observations have the same value.
    - Of the cases that are different, the majority seems to be due to different ways of recording schooling types that correspond to the same general level (e.g.

schooling could be recorded as 73 [73:Islamic Junior/High School (Madrasah Tsanawiyah)] in one variable and as 3 [3:General jr. high] in the other, but both correspond to the same overall level of education ('Junior High'), per the documentation above.

- ar16 has the inverse of the advantages and disadvantages documented above for the dl06 variable. Specifically, it has many more responses of 98 [Don't Know], but it also contains valid values for the respondents without schooling. Of the 1815 respondents with a value of 1:Unschooled on ar16, 1695 have a value of NA for dl06 [the next largest category is 107 respondents listed as 2:Elementary school for dl06].

Recommended way to combine: The data finders recommend using ar16 as the foundation of a cleaned education variable for the replication analysis. Values from dl06 should be used when the value of ar16 is 98 [Don't Know].

- This is still an imperfect solution, since there will be cases where the differences can't easily be resolved (e.g. a respondent is listed as 6 [Senior high vocational] on one variable and 61 [University S1] on the other, though those are different levels of educational attainment according to the coding scheme above). But absent any indication that one variable is more accurate than the other, the data finders recommend using the variable that records respondents without schooling as a valid response (i.e. ar16) as the foundation.

Variable name: dl04

- File name: b3a_dl1.dta
- Folder: hh14_b3a_dta
- Description: Have you ever attended/are you attending school?
- Notes: Of the 1741 respondents with a response of 'No' to this question, all have a value of NA for dl06. By contrast, 1682 of those respondents have a value of 1:Unschooled on ar16. The largest other category in ar16 is 42 respondents with a value of 2:Grade school.

Variable name: dl05b

- File name: b3a_dl1.dta
- Folder: hh14_b3a_dta
- Description: Did you attend kindergarten?
- Notes: Of the 1741 respondents who did not attend school ('No' response to dl04 above), 1733 have a response of 'No' to this question, while 7 have a 'Yes' response and 1 is 'Don't Know.'
    - For an unknown reason, despite 'Kindergarten' being a valid response to dl06 below, the 7 respondents who report attending kindergarten in dl05b are NA in dl06.
    - The 7 respondents who report attending kindergarten in dl05b are 1:Unschooled in ar16.

Variable name: dl07

- File name: b3a_dl1.dta
- Folder: hh14_b3a_dta
- Description: What is the highest grade completed at that school?
- Notes: Though the codebook (https://osf.io/ydc83/?view_only=3decbb703a024302b35df2108a04a0eb) doesn't make this explicit, this is presumably a reference to the response from dl06.
- Additional notes: The valid responses run from 0 [Did not complete first grade at that level] to 7 [Graduated], but the values of 1-6 are unlabeled.

Additional education variables

Notes: The variables above should be sufficient for the purposes of the replication analysis, so the additional items below are probably not necessary. Still, they are provided to the data analyst in case they are of use in the replication analysis.

Variable name: dl07a

- File name: b3a_dl1.dta
- Folder: hh14_b3a_dta
- Description: Are you currently attending school?

Variable name: dl2type

- File name: b3a_dl2.dta
- Folder: hh14_b3a_dta
- Description: School level ["Highest level school attended/or are attending?"]
- Notes: If the education variable below (dl16xa) is used, dl2type can identify the level of school being referred to in the particular response.

Variable name: dl16xa

- File name: b3a_dl2.dta
- Folder: hh14_b3a_dta
- Description: Currently in school at this level?
- Notes: Valid responses of Yes or No

Variable name: dl10

- File name: b3a_dl2.dta
- Folder: hh14_b3a_dta
- Description: What is the school level you attended or you are still attending?
- Notes: This variable can have up to 4 responses per individual. From the structure of the question and answer format, it seems as if the respondent can provide a response for each of the four different school levels in dl2type (Elementary, Junior High, Senior High, and 'D1, D2, D3//University').

    - For the **vast majority** of respondents, the maximum value of the dl10 responses matches the value of dl06.
    - In the cases where those values don't match, it's often when high values of dl10 don't correspond to a high level of education. For example, 72, 73, and 74 on dl10 correspond to Madrasah Elementary, Madrasah Junior High School, and Madrasah Senior High School, respectively, but all of those are a lower level of education than the values corresponding to a university education (13 and 60-63).

Age/age-squared

Note: The replication dataset constructed below does **not** contain an age-squared variable. That will need to be created by the data analyst.

Variable name: age

- File name: b3a_cov.dta
- Folder: hh14_b3a_dta
- Description: Age (in years)

Variable name: ar09

- File name: bk_ar1.dta
- Folder: hh14_bk_dta
- Description: Age now [provided during listing of household members]
- Notes: When compared directly to the 'Age' variable listed above, these two variables are very similar, though not equivalent.
    - Of the 36,385 non-NA values of 'Age,' 35,867 have the exact same value on ar09. The vast majority of the differences are only off by one year.
    - There's no clear pattern to the differences, so it's difficult to say that one is more suitable than the other as a value for age.
- When merged with the dataset containing the dependent variable (cd05), 'ar09' presents no missing values and one value of 998. By contrast, 'Age' has 6 missing values and one value of 998 [same respondent -- unique_id = 276020011]. With no values of NA, ar09 is probably the better selection.

Risk preference

Variable names: si01, si02, si03, si04, si05, si11, si12, si13, si14, si15, random_si

- File name: b3a_si.dta
- Folder: hh14_b3a_dta
- Description: Series of connected and branching questions that collectively are used to measure risk preferences.
- Additional notes: The survey logic is documented in the original questionnaire [https://osf.io/gfkh6/?view_only=3decbb703a024302b35df2108a04a0eb], and depends

in part on which version of the survey ordering the respondent saw (that is, random_si = A or random_si = B). The relevant pages of the questionnaire have been excerpted out for easier reference [https://osf.io/wc726/?view_only=3decbb703a024302b35df2108a04a0eb].

Time preference

Variable names: si21a, si21b, si21c, si21d, si21e, si22a, si22b, si22c, si22d, si22e, random_si

- File name: b3a_si.dta
- Folder: hh14_b3a_dta
- Additional notes: The survey logic is documented in the original questionnaire [https://osf.io/gfkh6/?view_only=3decbb703a024302b35df2108a04a0eb], and depends in part on which version of the survey ordering the respondent saw (that is, random_si = A or random_si = B). The relevant pages of the questionnaire have been excerpted out for easier reference [https://osf.io/wc726/?view_only=3decbb703a024302b35df2108a04a0eb].

Distance to health center

Note: This variable should probably be considered ***unavailable for the replication***, since the only approximate items have serious limitations (see below for details). If the data analyst can identify alternative measures in the IFLS documentation, they are encouraged to add it to the dataset or consult the data finders about adding it on their behalf.

Variable name: rj11

- File: b3b_rj2.dta
- Folder: hh14_b3b_dta
- Description: Distance to medical facility
- Additional notes: This was only asked of respondents who had visited a medical provider in the last four weeks. Further, only respondents who knew the distance have a value for this variable. Finally, this is only a proxy for 'distance to nearest health center,' since the medical provider that respondents visited in the past 4 weeks may not be the nearest one. ***For all of these reasons, it is recommended that the data analyst not include this variable in the replication analysis.***

Variable name: rj11x

- File: b3b_rj2.dta
- Folder: hh14_b3b_dta
- Description: Distance to medical facility known/unknown [How far is it from the medical facility to your residence? (able to answer)]
- Additional notes: Only respondents who had visited a medical provider in the last four weeks have a non-missing value for this variable. **As mentioned above, *it is***

recommended that the data analyst not include this variable in the replication analysis.

Household per capita expenditures (PCE)

Weekly household spending

Variable name: ks02

- File: b1_ks1.dta
- Folder: hh14_b1_dta
- Description: During the past week, what was the total expenditure to purchase [...]?
- Notes: As indicated below, the original data includes multiple lines per household, corresponding to weekly expenditures for different items.

Variable name: ks1type

- File: b1_ks1.dta
- Folder: hh14_b1_dta
- Description: Type of food item
- Notes: As indicated below, the original data includes multiple lines per household, corresponding to weekly expenditures for different items.

Variable name: ks02x

- File: b1_ks1.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS02
- Notes: The **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks02 (volume of expenditure) is 0, the value of ks02x is 3 (not given). This suggests that the 'able to answer question ks02' variable description is misleading, since at least *some* of the responses of 0 to ks02 represent a household reporting no expenditures for that particular item.

<u>Replication dataset conversions</u>

To facilitate the later data merges more easily, the expenditure items above were reshaped to a wide format, so that each household's weekly spending was listed in a single row. The following variables were created in the process. Each corresponds to a different weekly expenditure item, and each is documented in the data dictionary linked below:

- **Expenditure amount (by item):** ks02_ks1type_C; ks02_ks1type_M; ks02_ks1type_OA; ks02_ks1type_IB; ks02_ks1type_N; ks02_ks1type_OB; ks02_ks1type_HA; ks02_ks1type_FA; ks02_ks1type_P; ks02_ks1type_S; ks02_ks1type_T; ks02_ks1type_A; ks02_ks1type_K; ks02_ks1type_EA; ks02_ks1type_H; ks02_ks1type_BA; ks02_ks1type_Z; ks02_ks1type_I; ks02_ks1type_J; ks02_ks1type_W; ks02_ks1type_IA; ks02_ks1type_AA; ks02_ks1type_F;

ks02_ks1type_E; ks02_ks1type_GA; ks02_ks1type_CA; ks02_ks1type_L; ks02_ks1type_B; ks02_ks1type_U; ks02_ks1type_V; ks02_ks1type_G; ks02_ks1type_R; ks02_ks1type_DA; ks02_ks1type_D; ks02_ks1type_Y; ks02_ks1type_Q; ks02_ks1type_X

- **Able to answer (by item)**: ks02x_ks1type_C; ks02x_ks1type_M; ks02x_ks1type_OA; ks02x_ks1type_IB; ks02x_ks1type_N; ks02x_ks1type_OB; ks02x_ks1type_HA; ks02x_ks1type_FA; ks02x_ks1type_P; ks02x_ks1type_S; ks02x_ks1type_T; ks02x_ks1type_A; ks02x_ks1type_K; ks02x_ks1type_EA; ks02x_ks1type_H; ks02x_ks1type_BA; ks02x_ks1type_Z; ks02x_ks1type_I; ks02x_ks1type_J; ks02x_ks1type_W; ks02x_ks1type_IA; ks02x_ks1type_AA; ks02x_ks1type_F; ks02x_ks1type_E; ks02x_ks1type_GA; ks02x_ks1type_CA; ks02x_ks1type_L; ks02x_ks1type_B; ks02x_ks1type_U; ks02x_ks1type_V; ks02x_ks1type_G; ks02x_ks1type_R; ks02x_ks1type_DA; ks02x_ks1type_D; ks02x_ks1type_Y; ks02x_ks1type_Q; ks02x_ks1type_X

Monthly household spending

Variable name: ks06

- File: b1_ks2.dta
- Folder: hh14_b1_dta
- Description: What were the total expenditures by all household members for [...] during the past month, namely since date [...] one month ago?
- Notes: As indicated below, the original data includes multiple lines per household, corresponding to monthly expenditures for different items.

Variable name: ks2type

- File: b1_ks2.dta
- Folder: hh14_b1_dta
- Description: Type of non-food items
- Notes: As indicated below, the original data includes multiple lines per household, corresponding to monthly expenditures for different items.

Variable name: ks06x

- File: b1_ks2.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS06
- Notes: As with the weekly spending variable above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks06 (volume of expenditure) is 0, the value of ks06x is 3 (not given). This suggests that the ‘able to answer question ks06’ variable description is misleading, since at least *some* of the responses of 0 to ks06 represent a household reporting no expenditures for that particular item.

Replication dataset conversions

To facilitate the later data merges more easily, the expenditure items above were reshaped to a wide format, so that each household's monthly spending was listed in a single row. The following variables were created in the process. Each corresponds to a different monthly expenditure item, and each is documented in the data dictionary linked below:

- **Expenditure amount (by item):** ks06_ks2type_F1; ks06_ks2type_B; ks06_ks2type_A1; ks06_ks2type_C; ks06_ks2type_A2; ks06_ks2type_A3; ks06_ks2type_F2; ks06_ks2type_E; ks06_ks2type_A4; ks06_ks2type_D; ks06_ks2type_G; ks06_ks2type_C1; ks06_ks2type_
- **Able to answer (by item)**: ks06x_ks2type_F1; ks06x_ks2type_B; ks06x_ks2type_A1; ks06x_ks2type_C; ks06x_ks2type_A2; ks06x_ks2type_A3; ks06x_ks2type_F2; ks06x_ks2type_E; ks06x_ks2type_A4; ks06x_ks2type_D; ks06x_ks2type_G; ks06x_ks2type_C1; ks06x_ks2type_

Yearly household spending

Variable name: ks08

- File: b1_ks3.dta
- Folder: hh14_b1_dta
- Description: What were the total expenditures by all household members for [...] during the past year, namely since the month of [...] last year?
- Notes: As indicated below, the original data includes multiple lines per household, corresponding to yearly expenditures for different items.

Variable name: ks3type

- File: b1_ks3.dta
- Folder: hh14_b1_dta
- Description: Type of non-food items
- Notes: As indicated below, the original data includes multiple lines per household, corresponding to yearly expenditures for different items.

Variable name: ks08x

- File: b1_ks3.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS08
- Notes: As with the weekly and monthly spending variables above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks08 (volume of expenditure) is 0, the value of ks08x is 3 (not given). This suggests that the 'able to answer question ks08' variable description is misleading, since at least *some* of the responses of 0 to ks08 represent a household reporting no expenditures for that particular item.

Replication dataset conversions

To facilitate the later data merges more easily, the expenditure items above were reshaped to a wide format, so that each household's yearly spending was listed in a single row. The following variables were created in the process. Each corresponds to a different yearly expenditure item, and each is documented in the data dictionary linked below:

- **Expenditure amount (by item):** ks08_ks3type_C; ks08_ks3type_A; ks08_ks3type_E; ks08_ks3type_F; ks08_ks3type_G; ks08_ks3type_D; ks08_ks3type_B
- **Able to answer (by item)**: ks08x_ks3type_C; ks08x_ks3type_A; ks08x_ks3type_E; ks08x_ks3type_F; ks08x_ks3type_G; ks08x_ks3type_D; ks08x_ks3type_B

Schooling expenses

Note: Unlike the expenditure items above, each row in the IFLS 5 dataset used below (b1_ks0.dta) already uniquely identifies households, so there is no need to reshape the format.

Variable name: ks10aa

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Total expenditures for school fees for children/family members inside the household during the past year

Variable name: ks10aax

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS10AA
- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks10aa (volume of expenditure) is 0, the value of ks10aax is 3 (not given). This suggests that the 'able to answer question ks10aa' variable description is misleading, since at least *some* of the responses of 0 to ks10aa represent a household reporting no expenditures for that particular item.

Variable name: ks10ab

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Amount spent for school fees for children/family members outside the household during the past year

Variable name: ks10abx

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question ks10ab

- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks10ab (volume of expenditure) is 0, the value of ks10abx is 3 (not given). This suggests that the 'able to answer question ks10ab' variable description is misleading, since at least *some* of the responses of 0 to ks10ab represent a household reporting no expenditures for that particular item.

Variable name: ks11aa
- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Amount spent for schooling needs for children/family members inside the household during the past year

Variable name: ks11aax
- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS11AA
- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks11aa (volume of expenditure) is 0, the value of ks11aax is 3 (not given). This suggests that the 'able to answer question ks11aa' variable description is misleading, since at least *some* of the responses of 0 to ks11aa represent a household reporting no expenditures for that particular item.

Variable name: ks11ab
- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Amount spent for schooling needs for children/family members outside the household during the past year

Variable name: ks11abx
- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS11AB
- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks11ab (volume of expenditure) is 0, the value of ks11abx is 3 (not given). This suggests that the 'able to answer question ks11ab' variable description is misleading, since at least *some* of the responses of 0 to ks11ab represent a household reporting no expenditures for that particular item.

Variable name: ks12aa
- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Amount spent on school transportation and pocket money for children/family members in the household during the past year

Variable name: ks12aax

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS12AA
- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks12aa (volume of expenditure) is 0, the value of ks12aax is 3 (not given). This suggests that the 'able to answer question ks12aa variable description is misleading, since at least *some* of the responses of 0 to ks12aa represent a household reporting no expenditures for that particular item.

Variable name: ks12ab

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Amount spent on school transportation and pocket money for children/family members outside the household during the past year

Variable name: ks12abx

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS12AB
- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks12ab (volume of expenditure) is 0, the value of ks12abx is 3 (not given). This suggests that the 'able to answer question ks12ab variable description is misleading, since at least *some* of the responses of 0 to ks12ab represent a household reporting no expenditures for that particular item.

Variable name: ks12bb

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Amount spent on food and boarding/rent for children/family members outside the household during the past year

Variable name: ks12bbx

- File: b1_ks0.dta
- Folder: hh14_b1_dta
- Description: Able to answer question KS12BB
- Notes: As above, the **vast majority** of responses to this variable are either 1 (given) or 3 (not given). However, in ***every case*** where the value of ks12bb (volume of expenditure) is 0, the value of ks12bbx is 3 (not given). This suggests that the 'able to answer question ks12bb variable description is misleading, since at least *some* of the responses of 0 to ks12bb represent a household reporting no expenditures for that particular item.

Household size

Variable name: ar01a

- File: bk_ar1.dta
- Folder: hh14_bk_dta
- Description: Does HHM still live in this household
- Notes: This variable is used to create a measure of household size in the data preparation steps below, since household size is needed to construct a measure of household per capita expenditures. The ar01a variable takes on the following values:
    - 0. Died
    - 1. Yes, HHM is still in HH
    - 2. Yes, HHM was in other IFLS HH in previous wave
    - 3. No
    - 5. New HHM
    - 11. HHM returns in current wave
    - 6 is a value in the data but it does not appear in the questionnaire (see more below).
- As documented below, responses 0, 3, and 6 were first removed from the dataset. Then for each household the number of rows in the data were summed together in a new variable (hh_size) that represents the number of household members as of IFLS 5.
    - The IFLS 6 documentation does not make clear whether individuals with a value of 6 for ar01a are in the household or not. The data validation exercise (explained below) demonstrated that the hh_size variable more closely matches the alternative measure below when observations with a value of 6 are *not* included in the hh_size calculation, though the difference is negligible.

Alternative measure (for data validation purposes)

The derived variable documented above (hh_size) is the closest approximation to a measure of household size that appears in the codebook, but it is still an indirect measure.

To validate that it's a sensible proxy for household size, an alternative measure was constructed by adding up the number of rows per household in the blood pressure data file (bus_us.dta; see below). Since each row of the bus_us.dta file is a different individual, the total rows per household should also proxy the household size. Joining this alternative measure to the hh_size variable above and taking the difference revealed the following:

- 11,248 households had the exact same value between the two measures.
- 2,142 households had a difference of 1 between the two measures
- 1,386 households had a difference of 2-4
- 190 households had a difference of 5-10
- 4 households had a difference of more than 11
- 911 households had a difference of NA (no value from the household size proxy based on the blood pressure data)

In all but 3 households, the measure based on the ar01a variable is larger than the measure based on the blood pressure data, which suggests that the measure based on the blood pressure data is undercounting members of the household. Therefore, the data finders recommend that the data analyst use the hh_size variable documented above (based on ar01a in the original data) as the proxy for household size.

Additional notes about household expenditures

The collection of household spending items documented above (weekly, monthly, yearly, and school expenses) is comprehensive, though not exhaustive of all potential expenditure items in the IFLS 5. For example, there are additional items about the quantity and price of certain staple goods purchased in the past month (see variables ks13-ks16 in the questionnaire for details), but it's unlikely that these are useful for a measure of household expenditures per capita.

Sex

Variable name: Sex

- File name: b3a_cov.dta
- Folder: hh14_b3a_dta
- Description: Sex (as filled out by interviewer) (Male/Female)

Variable name: ar07

- File name: bk_ar1.dta
- Folder: hh14_bk_dta
- Description: Sex [provided during listing of household members]
- Notes: When compared directly to the Sex variable listed above, these two variable are nearly equivalent [just one substantive difference], so either could be a suitable selection.
- When merged with the dataset containing the dependent variable (cd05), ar07 presents no missing values and 'Sex' has 6 missing values, so the ar07 variable is probably the better choice.

Additional variables

Variable name: unique_id

- Description: Individual-level identifier
- Notes: This variable was created during the data preparation process by concatenating 'hhid14' [2014 Household ID] and 'pid14' [2014 Person ID within household] from the original datasets. This variable is used to merge variables measured at the individual level.

- Additional notes: The rationale for this variable is based on the IFLS Wave 5 User Guide [https://osf.io/bn3a4/?view_only=3decbb703a024302b35df2108a04a0eb]: “If the level of observation is the individual, both HHID14 and PID14 are required to uniquely identify a person, unless PIDLINK and AR01a are used. [footnote 6: Within IFLS5 files, use HHID14 and PID14 to identify individuals.]” (p. 9)

Variable name: hhid14

- Description: Household identifier [2014 Household ID]
- Additional notes: This variable appears in each of the individual data files. It is used to merge variables measured at the household level.

Variable name: ar00x

- File name: bk_ar0.dta
- Folder: hh14_bk_dta
- Description: Household status in the the panel
- Additional notes: This indicator for whether the household is an existing panel household or a new household does not have a clear use in the replication analysis, but it is being included in the replication dataset in case the data analyst has a need for it.

Variable name: rspndnt

- File name: b3a_cov.dta
- Folder: hh14_b3a_dta
- Description: Respondent household member type
- Notes: Valid responses are 1:Head of household, 2:Spouse of head, 3:Other HH member. As with the ar00x above, it’s not immediately clear that it has a purpose in the replication analysis, but it is available in case the data analyst would like to use it.

<u>Proxy variables</u>

The IFLS5 data contains a subset of individuals for whom at least some of the questions were answered by a proxy, instead of directly by that individual. The IFLS5 user guides and codebooks document these processes in a few different places, and relevant passages are below:

- “Because obtaining interviews with all household members is difficult, IFLS5, like earlier waves, included a proxy book that was used for collecting more limited information (from other household members) about individuals who could not be interviewed in person.” (p. 14, Vol. 1: https://osf.io/bmc6j/?view_only=3decbb703a024302b35df2108a04a0eb)
- “Anticipating the impossibility of interviewing all the adult respondents from whom we wanted information, we used a proxy book (Book Proxy), first introduced in IFLS2, to obtain a subset of information from someone who could answer for a respondent. The proxy book contained many of the modules from books 3A, 3B, and 4, but most modules asked for considerably less information than the ‘main’ books...Table 2.1 indicates the

differences in information obtained from Book Proxy and corresponding main books in IFLS5." (p. 5, Vol. 2: https://osf.io/84whx/?view_only=3decbb703a024302b35df2108a04a0eb)

- "Book 3 is administered to all adults age 15 and older. It is split into two parts, A and B, because the book is very long and few respondents were willing or able to complete all modules in a single sitting. By splitting the books, we had more control over when breaks were taken. All adult respondents are supposed to complete Book 3. If the target respondent could not be interviewed after at least 3 attempts to find him or her, a proxy respondent was asked to complete a subset of the questions in Book 3. Those questions are in Book 3P. IT IS IMPORTANT TO COMBINE INFORMATION IN BOOK 3 (A or B) WITH THE PROXY INFORMATION IN BOOK 3P TO OBTAIN A COMPLETE SET OF INDIVIDUAL RESPONSES. THE INDIVIDUAL WEIGHTS ARE BASED ON THE ASSUMPTION THAT PROXY RESPONSES ARE INCLUDED IN THE ANALYSES." (p. 2, Book 3a codebook: https://osf.io/ydc83/?view_only=3decbb703a024302b35df2108a04a0eb)

What is not entirely clear from the documentation above is whether all responses for an individual who is recorded as a 'Book Proxy' respondent in one of the modules (e.g. Book 3a) were answered by proxy, or whether only the responses in that specific module are answered by proxy.

Among the modules included in this replication dataset, only Book 3a and Book 3b have variables recording whether an individual's responses were provided by proxy.

- This could suggest that those same individuals' responses in other modules were provided in person, rather than by proxy. Some corresponding evidence is found for this in the documentation: "Also, if a prime aged, healthy person had not been found, so a proxy book used to acquire information, an interviewer was sometimes sent back to attempt to find and interview that person" (p. 32, Vol. 1: https://osf.io/bmc6j/?view_only=3decbb703a024302b35df2108a04a0eb)

Variable name: proxy

- File name: b3a_cov.dta
- Description: Proxy Book? [1: Book Proxy, 3: Not Book Proxy]
- Notes: This is presumably the variable in Book 3a that identifies whether an individual's responses were provided by proxy. As mentioned above, when compared to an individual's value for the equivalent variable in Book 3b (which is not included in the replication dataset), it is either the same value or can't be compared because the value for Book 3b is NA. There are six cases where an individual is NA in both versions, and no cases where the individual's proxy value is NA in Book 3a but available in Book 3b.

Variable name: relatprox

- File name: b3a_cov.dta
- Description: Relationship with the respondent

- Notes: Identifies the relationship between the proxy and the individual. Most often it's a family member. The value is always NA if the response to proxy is anything except '1: Book Proxy.'

Variable name: reasprox

- File name: b3a_cov.dta
- Description: Reason for proxy
- Notes: As with relatprox above, the value is always NA if the response to proxy is anything except '1: Book Proxy.'

Despite that the "proxy respondents" are relevant in the process of constructing the replication dataset, they do not matter to test the SCORE claim, **H***. Following the covariates considered in the original study (page 18), the probit specification proposed to test the focal claim should include measures for risk and time preference (see the "Inference criteria" section). All subjects that answered the questions necessary to construct the risk preference or time preference measures are not proxy respondents. Consequently, the sample used to test the focal claim will exclude the proxy respondents.

Lines 430 and 438 of the Stata do-file "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" are included to verify the proxy classification of the respondents used in the probit regressions (i.e., 1:Book Proxy or 3:Not Book Proxy).

Weighting variables

The user's guide for IFLS5 [https://osf.io/bn3a4/?view_only=3decbb703a024302b35df2108a04a0eb] contains an extensive discussion of the weighting variables available, and the considerations behind how each one was constructed and should be used. A relevant excerpt for this project is: "While IFLS is a longitudinal survey, there will be some analyses that treat IFLS5 as a cross-section. We have attempted to construct weights so that estimates based on IFLS5 will be representative of the Indonesian population living in the 13 IFLS provinces in 2014...An analogous strategy has been adopted to construct cross-section analysis weights at the household level." (p. 20-21)

Since the replication study only relies on the fifth wave of the IFLS, only the IFLS5 cross-sectional weights for individuals and households have been included in the replication dataset. Those variables are below. All quotes below are references from the user's guide linked above [https://osf.io/bn3a4/?view_only=3decbb703a024302b35df2108a04a0eb].

Individual weight variables

Variable name: pwt14usxa

- File name: ptrack.dta

- Folder: hh14_trk_dta
- Description: IFLS5 cross-section US weight w/ attrition correction
- Notes: "...Similar weights have been constructed for use with the health assessments. PWT14USXa was constructed by raking IFLS5 for persons who had US measurements, to the 2014 SUSENAS, first taking into account attrition from 1993 to 2014 (from the IFLS1 roster to who was measured in IFLS5)."

Variable name: pwt14usx_
- File name: ptrack.dta
- Folder: hh14_trk_dta
- Description: IFLS5 cross-section US weight w/o attrition correction
- Notes: "...Similarly, PWT14USX_ constructs the US weight without attrition adjustments."

Variable name: pwt14xa
- File name: ptrack.dta
- Folder: hh14_trk_dta
- Description: IFLS5 X-section member weight w/ attrition correction
- Notes: "The IFLS5 cross-section analysis person weights are the ratio of the 2014 SUSENAS proportion to the IFLS5 proportion in each cell...The resulting weight is called PWT14Xa and is included in PTRACK."

Variable name: pwt14x_
- File name: ptrack.dta
- Folder: hh14_trk_dta
- Description: IFLS5 X-section member weight w/o attrition correction
- Notes: "As for the household cross-section weights, we also report the weights without attrition corrections, PWT14X_."

Variable name: res14us
- File name: ptrack.dta
- Folder: hh14_trk_dta
- Description: 'result 14: Book US completed by resp?'
- Notes: Book US refers to the blood pressure data, so only respondents who are coded as 1 ('complete') on this measure have weights included in the replication dataset.
- Additional notes: The number of respondents who are coded as 'complete' on this measure (48,139) matches the number of observations in the blood pressure dataset, but it's not a one-to-one relationship.
    - There are five ID strings in the weights data that are not in the blood pressure data; and there are nine ID strings in the blood pressure data that are not in the weights data. This mismatch in number is likely due to the repeated ID strings in the weights data (see below for details).

Household weight variables

Variable name: hwt14xa

- File name: htrack.dta
- Folder: hh14_trk_dta
- Description: IFLS5 HH X-section weight w/ attrition adj.
- Notes: “All households in the IFLS5 sample have been stratified by province and urban-rural sector. For each cell, the ratio of the proportion of households in the 2014 SUSENAS sample (in IFLS provinces) to the IFLS5 sample proportion, multiplied by the attrition-weight provides the IFLS5 cross-section analysis household weight, HWT14Xa...Estimates that are weighted with HWT14Xa should be representative of all households living in the IFLS provinces in Indonesia in 2014.”

Variable name: hwt14x_

- File name: htrack.dta
- Folder: hh14_trk_dta
- Description: IFLS5 HH X-section weight w/o attrition adj.
- Notes: “...A second weight, HWT14X_, does not use the attrition correction.”

Variable name: result14

- File name: htrack.dta
- Folder: hh14_trk_dta
- Description: 14 result of HH Ivw
- Notes: This item documents whether the household was interviewed for ILFS 5. It is included just as a verification check that all weights included in the replication dataset are only for households interviewed in the wave used for this replication.

(d) Data Creation

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

- *If the data can be freely shared and posted to OSF, please post it in your OSF project and provide a link to the completed dataset below.*
- *If any part of the dataset cannot be shared between researchers or posted to the OSF, please leave the final dataset off the OSF. Instead, include either below or in your script (commented out at the bottom) two pieces of information that will help an independent team verify they have created the dataset according to your instructions:*
    - *The dimensions of the final dataset(s) you’ve created (# of rows, # of columns)*

    - *A summary of 8-10 variables in the replication dataset. For numeric variables, the summary should include the mean, standard deviation, and count of NAs. For categorical variables, the summary should include each level present in the data and its count, as well as a count of NAs. If multiple datasets are submitted as part of your work, at least one variable should be included from each dataset.*

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

Please also use this section to describe:

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

The replication data is split across three files. The main replication dataset contains the core variables needed for the replication analysis. There are two additional datasets, as well:

Alternative education items

Three of the alternative educational items documented above (dl2type, dl16xa, and dl10) are contained in a dta file where the rows do not uniquely identify individuals. To preserve the structure of the main dataset -- and because these items are not recommended for use in the analysis anyways -- the items have not been merged into the main dataset. Instead, they're located in a separate dataset (alt_education).

Weights

A single weights dataset was created by merging the household weights items with the individual weights items, after filtering the individual weights dataset so that only respondents who were 1 ('complete') on res14us were retained. The result is a weights dataset where the rows *almost* uniquely identify individuals. There are 4 ID strings (unique_id) that repeat in the dataset (2 rows for each value of unique_ID). Because of this, the weights dataset has not been merged into the main dataset, so that rows in the main dataset can continue to uniquely identify individuals. The weighting items are kept in a separate dataset (full_weights).

All data cleaning and merging was performed in R. Please see the R code linked below for details on packages and versions.

The following steps create the replication datasets:

- Load the 16 data files from the ILFS 5 that have been documented above. Those files are listed in this readme file for reference: https://osf.io/mt7bc/?view_only=3decbb703a024302b35df2108a04a0eb

- Create the following datasets that are indexed by a derived variable (unique_id) that identifies each individual in the dataset [see above for details on how unique_id was constructed]. Select the variables from each dataset that are needed for the replication analysis. Those datasets are:
    - hypertension
        - Note: the ILFS5 dataset this is drawn from (b3b_cd3.dta) contains individual responses spread out over multiple rows. The hypertension dataset only retains observations with a value of 'A' on the cdtype variable (indicating that the specific response is about hypertension), such that in the resulting dataset, each row uniquely identifies an individual.
    - b_pressure
    - health
    - distance
    - education
    - alt_education
    - demographics
    - time_risk
    - added_demographics
    - indiv_weights

- Create the following datasets that are indexed by a household identifier (hhid14) that identifies each household in the IFLS5. Select the variables from each dataset that are needed for the replication analysis:
    - consumption1
        - Reshape from long to wide such that each row is a separate household
    - consumption2
        - Reshape from long to wide such that each row is a separate household
    - consumption3
        - Reshape from long to wide such that each row is a separate household
    - consumption4
    - panel_dummy
    - household_weights
- Create a single household consumption dataset by merging consumption1, consumption2, consumption3, and consumption4 using hhid14.

- Create an hh_size dataset containing an approximate measure of household size by summing, at the household level, the number of new household members and household members from the previous wave who are still members of the household.

- Create a weights dataset (full_weights) with both individual-level and household-level weights variables.

- Finally, merge all datasets except alt_education and full_weights. This merged dataset (replication_data) is created in the following way:
    - full join for most of the join operations on the individual-level data in order to preserve as many observations as possible
    - left join for merging in the 'added_demographics' individual-level data because a full join drastically increases the number of rows without matches on the 'left side'
    - left join for merging all of the household level data, since there's no purpose to adding household data without any matching individual-level observations

The datasets for the replication are as follows:

- replication_data
- alt_education
- full_weights

The R code to produce the replication datasets is found here: https://osf.io/5c93n/?view_only=3decbb703a024302b35df2108a04a0eb. Please consult this file for details on its use: https://osf.io/mt7bc/?view_only=3decbb703a024302b35df2108a04a0eb

- The R code also includes a function that summarizes the variables in each dataset with a set of key descriptives, in order to verify that the code is producing the intended dataset. The reference files to compare the variable summaries to are as follows:
    - Replication dataset verification file: https://osf.io/s8dae/?view_only=3decbb703a024302b35df2108a04a0eb
    - Weights dataset verification file: https://osf.io/arwyb/?view_only=3decbb703a024302b35df2108a04a0eb
    - Alternative education measures verification file: https://osf.io/6qkeu/?view_only=3decbb703a024302b35df2108a04a0eb

(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.*

There are three data dictionaries, corresponding to the three datasets created in Section 12d:

- Main replication dataset dictionary: https://osf.io/jq4e5/?view_only=3decbb703a024302b35df2108a04a0eb
- Weights file dictionary: https://osf.io/bh2qj/?view_only=3decbb703a024302b35df2108a04a0eb
- Alternative education measures dictionary: https://osf.io/k3psa/?view_only=3decbb703a024302b35df2108a04a0eb

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 main replication dataset created above (replication_data) contains 51,731 unique individuals. After limiting to the respondents who are in poor general health (kk01 equal to 3 or 4); who have a valid Yes or No response to the hypertension diagnosis measure (cd05); and who received a second and third blood pressure reading (us07bx and us07cx both equal to 1), that number drops to 7,027 unique respondents.

This sample size will be further reduced after the data analyst limits the sample to respondents who qualify as hypertensive according to WHO's guidelines ("Following WHO standards, a person is considered hypertensive if his systolic is greater than 140 or his diastolic is greater than 90." p. 18). Depending on how the data analyst decides to address missing data, there could be further reductions in the sample size as well.

Required sample size [to be filled out by the SCORE team]: The primary unit of analysis is the survey respondent. An estimate of the minimum viable sample size for the data analytic replication is: 970. For comparison, the stage1 required sample size would be: 4,529 and the stage2 sample size would be: 10,091.

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/sv7gz/?view_only=94275403f3fd472b8e10bac849dcb857

15. Stopping rule (provided by SCORE)

All observations will be used in a single analysis for this replication.

Variables

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

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

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

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

16. Manipulated variables

RR TEAM INSTRUCTIONS: *Manipulated variables in this preregistration refer specifically to variables that have been randomly assigned in an experiment. The use of data from an experiment should be rare in existing data replications. If your existing data replication relies on experimental data, please document each manipulated variable as a measured variable, and use the codebook to indicate what each level of the variable corresponds to (e.g. participants*

assigned to the treatment condition = 1; participants assigned to the control condition = 0). The default language in bold below has been copied into all existing data replication preregistrations.

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

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?*

Note: Lines 8 to 420 of the Stata do-file "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" generate the variables described below.

UNDER_DIAG

- Use in the analysis: Dependent variable.

- Description from the original study: "Our dependent variable is a dummy equal to one for those respondents who were found to be hypertensive during the IFLS screenings, but were not previously diagnosed by a doctor." (page 18)

  ""As part of the survey, trained nurses measured respondents' blood pressure three different times. The first measurement was dropped because many people get nervous at first which can cause false high measurements. We then used the average of the other two measurements to construct the hypertension variable. Following WHO standards, a person is considered hypertensive if his systolic is greater than 140 or his diastolic is greater than 90." (page 18)

- Variables used in the replication: us07b1, us07c1, us07b2, us07c2, and cd05.

To generate the HYPERTENSION variable, first, take the average of the last two blood measurements using the variables:

1. Blood pressure systolic: us07b1 and us07c1
2. Blood pressure diastolic: us07b2 and us07c2

The HYPERTENSION is a dummy variable equal to 1 if systolic>140 or diastolic>90, and 0 otherwise.

Besides the HYPERTENSION variable, use the variable cd05: “Have a doctor/paramedic ever told you that you had hypertension? (=3 for no)

- UNDER_DIAG is a dummy equal to 1 for those respondents who were found to be hypertensive during the IFLS screenings, but were not previously diagnosed by a doctor, and 0 otherwise.

- No deviations between the original study and the replication study.

YRS_SCHOOL

- Use in the analysis: Focal independent variable.

- Description from the original study: "Explanatory variables include respondents' education (measured in years of formal education)" (Page 18)

- Variables used in the replication: dl04, dl06 and dl07 .

First, using the variable dl04 (Have you ever attended/are you attending school?) it is possible to identify respondents with zero years of education.

Second, variable dl06 provides the highest level of education attended. And variable dl07 provides school achievement within that level of education. For the cases when dl07=7 (Graduated), information on the education in Indonesia provides an overview of how many years are needed to complete a particular level of education. The information on the education in Indonesia is gathered from the website “Education in Indonesia” (March 21, 2019), accessed on August 8th, 2020, and available at:

https://wenr.wes.org/2019/03/education-in-indonesia-2#:~:text=Elementary%20education%20(pendidikan%20dasar)%20lasts,%2C%20arts%2C%20and%20physical%20education.

The information can be double-checked with the Wikipedia entry of Education in Indonesia (accessed on August 8th, 2020):

https://en.wikipedia.org/wiki/Education_in_Indonesia

Using the categorization of school types (dl2type) listed under dl10 in the questionnaire and the information about education in Indonesia, it is possible to conclude that:

1. Elementary education lasts for six years. This includes dl06 = 2 (Elementary school), and dl06 = 72 (Islamic Elementary School (Madrasah Elementary)).
2. Junior High lasts for three years. This includes dl06 = 3 (Junior high general), dl06 = 4 (Junior high vocational), and dl06 = 73 (Islamic Junior/High School (Madrasah Senior High School).
3. Senior High lasts for three years. This includes dl06 =5 (Senior high general), dl06 =6 (Senior high vocational), dl06 =74 (Islamic Junior/High School (Madrasah Senior High School))

For D1, D2, D3//University, according to the values for dl07, it is possible to conclude:

1. College (D1, D2, D3) lasts for three years (dl06 =60)
2. University (BA) lasts for four years (dl06=61)
3. University (MA) lasts for three years (dl06=62)
4. University (PhD) lasts for five years (dl06=63)
5. Open University lasts for six years (dl06=13)

Finally, Adult Education A lasts for one year (dl06 =11), Adult Education B lasts for four years (dl06 =12), and Adult Education C for three years (dl06 =15).

- YRS_SCHOOL correspond to respondents' years of education.

- It is not clear from the study how to handle cases when dl06=14 (Islamic School (pesantren)), dl06=17 (School for Disabled), dl06=90 (Kindergarten), dl06=95 (Other), dl06=98 (Don't Know), and dl06=99 (MISSING)

According to the the categorization of school types (dl2type) listed under dl10 in the questionnaire (page 82 of the pdf), dl06=17 (School for Disabled) can correspond to Elementary, Junior High, or Senior High, while dl06=95 (Other) can correspond to any category of school type. Given the identification problem, YRS_SCHOOL is considered as a missing value for those cases.

dl06=90 (Kindergarten) is considered as zero years of schooling.

When dl06=14 (Islamic School (pesantren)), dl06=98 (Don't Know), and dl06=99 (MISSING), YRS_SCHOOL is considered as a missing value.

AGE

- Use in the analysis: Control variable.

- Description from the original study: “Explanatory variables include [...] respondents' age and age squared (to allow for possible non-linear effects)” (page 18)

- Variables used in the replication: ar09.

- AGE corresponds to respondents' age.
- For one respondent, ar09=998. In this case AGE is considered as a missing value.

AGESQRT

- Use in the analysis: Control variable.
- Description from the original study: “Explanatory variables include [...] respondents' age and age squared (to allow for possible non-linear effects)” (page 18)
- Variables used in the replication: ar09.
- AGE corresponds to respondents' age squared.
- For one respondent, ar09=998. In this case AGESQRT is considered as a missing value.

RISK_PREFERENCE

- Use in the analysis: Control variable.
- Description from the original study: “Explanatory variables include [...] respondents' individual risk and time preferences” (page 18)

“For the time and risk preference parameters, we follow Ng (2013) and group respondents in four distinct groups from the most patient to the most impatient, respectively from the least risk averse to the most risk averse”. (page.18)

- Variables used in the replication: si01, si02, si03, si04, si05, si11, si12, si13, si14, si15.

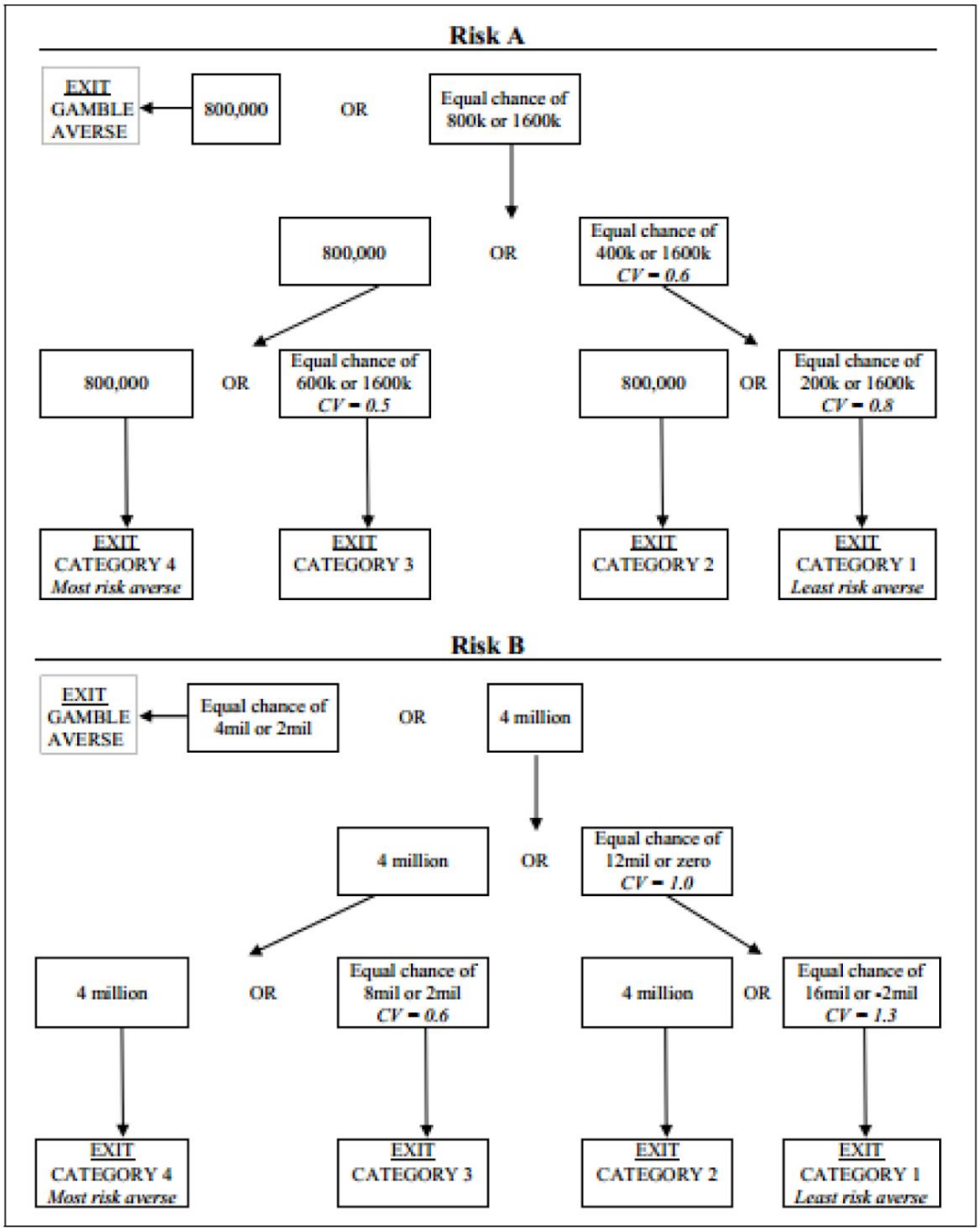
The methodology to construct the risk preference parameter is detailed in Ng "Risk and Time Preferences in Indonesia: The Role of Demographics, Cognition, and Interviewers." Although Kim & Radoias cite Ng (2013), the available online version is from 2012. The 2012 version also contains a risk and time elicitation process from the Indonesian Family Life Survey.

In Section 2.2, Ng explains how to construct a risk aversion measure based on respondents' certainty equivalent at the termination of the lottery questions.

The elicitation process is best illustrated in a flowchart, so Figure 1 from Ng (2012) is reproduced below. “The most risk averse respondents will exit the interview at the terminal node represented in the lower left corner of Figure 1; the least averse exit in the lower right node. The terminal nodes therefore represent an ordinal ranking of risk

aversion among the respondents. Respondents with risk aversion = 4 are the most risk averse, and those with risk aversion = 1 are least risk averse." (Ng, 2012, p. 9)

Figure 1: Flowcharts illustrating elicitation of risk aversion in IFLS-4



Source: Ng (2012), page 28.

Ng constructs two measures of risk aversion, that he calls Risk A and Risk B. Each measure depends on the two sets of questions that were asked. However, the results from Table 2 in Kim & Radoias suggests that they use a single risk preference measure. To combine Risk A and Risk B, an average of the two measures is taken. Moreover, because they “group respondents in four distinct groups [...] from the least risk averse to the most risk averse”. (page.18), the average is rounded up to the nearest integer.

- RISK_PREFERENCE corresponds to respondents' risk preferences. It is a categorical variable ranging from 1 to 4, increasing in risk aversion. Therefore:

RISK_PREFERENCE = 1 is the least risk averse

RISK_PREFERENCE = 2

RISK_PREFERENCE = 3

RISK_PREFERENCE = 4 is the most risk averse

- It is not clear from the study how to handle cases when si01=8 (Don’t know), si02=8 (Don’t know), si03=8 (Don’t know), si04=8 (Don’t know), si05=8 (Don’t know), si11=8 (Don’t know), si12=8 (Don’t know), si13=8 (Don’t know), si14=8 (Don’t know), or si15=8 (Don’t know). In those cases RISK_PREFERENCE is considered as a missing value.

TIME_PREFERENCE

- Use in the analysis: Control variable.

- Description from the original study: “Explanatory variables include [...] respondents' individual risk and time preferences” (page 18)

“For the time and risk preference parameters, we follow Ng (2013) and group respondents in four distinct groups from the most patient to the most impatient, respectively from the least risk averse to the most risk averse”. (page.18)

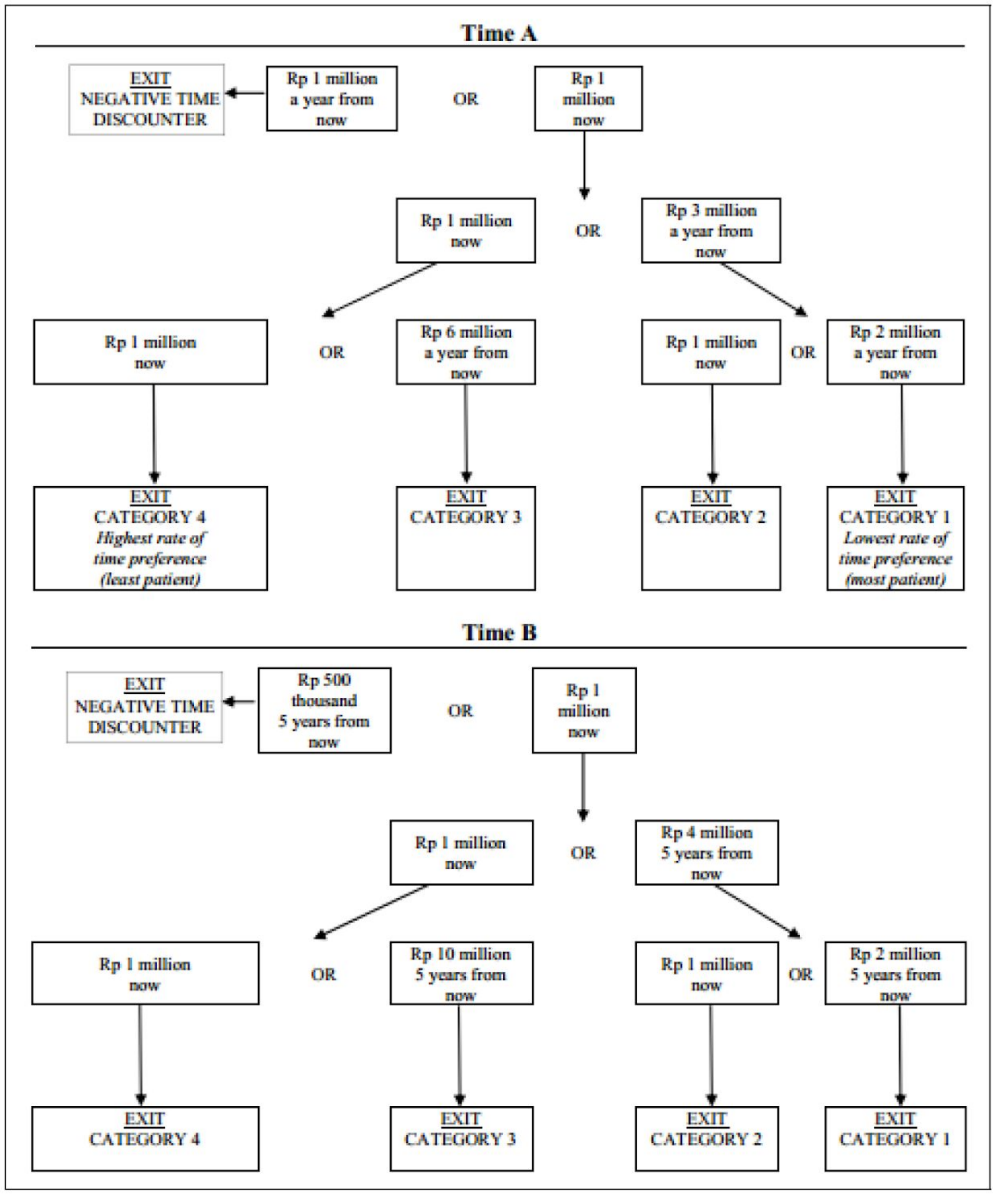
- Variables used in the replication: si21a, si21e, si21b, si21c, si21d, si22a, si22e, si22b, si22c, si22d

The methodology to construct the risk preference parameter is detailed in James Ng "Risk and Time Preferences in Indonesia: The Role of Demographics, Cognition, and Interviewers." Although Kim & Radoias cite Ng (2013), the available online version is from 2012. The 2012 version also contains a risk and time elicitation process from the Indonesian Family Life Survey.

In Section 2.3, Ng explains how to construct a time preference measure based on respondents' choices at the termination of the intertemporal choice questions.

The elicitation process is best illustrated in a flowchart, so Figure 2 from Ng (2012) is reproduced below. “Respondents who exit at Category 1 have the lowest time preference, and so are the most patient. At the other end, those who exit at Category 4 are the most impatient as they have the highest time preference” (Ng, 2012, p. 10)

Figure 2: Flowcharts illustrating elicitation of time preference in IFLS-4



Source: Ng (2012), page 29.

Ng constructs two measures of time preferences, that he calls Time A and Time B. Each measure depends on the two sets of questions that were asked. However, the results from Table 2 in Kim & Radoias suggests that they use a single time preference measure. To combine Time A and Time B, an average of the two measures is taken. Moreover, because they “group respondents in four distinct groups from the most patient to the most impatient”. (page.18), the average is rounded up to the nearest integer.

- TIME_PREFERENCE corresponds to respondents' time preferences. It is a categorical variable ranging from 1 to 4, decreasing in the level of patient. Therefore:

TIME_PREFERENCE = 1 is the most patient

TIME_PREFERENCE = 2

TIME_PREFERENCE = 3

TIME_PREFERENCE = 4 is the least patient

- It is not clear from the study how to handle cases when si21a=9 (Don’t know), si21e=9 (Don’t know), si21b=9 (Don’t know), si21c=9 (Don’t know), si21d=9 (Don’t know), si22a=9 (Don’t know), si22e=9 (Don’t know), si22b=9 (Don’t know), si22c=9 (Don’t know), or si22d=9 (Don’t know). In those cases TIME_PREFERENCE is considered as a missing value.

DISTANCE

- Use in the analysis: Control variable.

- Description from the original study: “Explanatory variables include [...] the distance from the closest health center (to proxy for the ease of access to medical care)” (page 18)

- Variables used in the replication: rj11.

- DISTANCE corresponds to the distance to medical facility.

- As explained in Section 12, the variable rj11 contains serious limitations. For example, this was only asked of respondents who had visited a medical provider in the last four weeks. Further, only respondents who knew the distance have a value for this variable. Finally, this is only a proxy for ‘distance to nearest health center,’ since the medical provider that respondents visited in the past 4 weeks may not be the nearest one.

  Also, in the replication, DISTANCE is considered as missing if rj11x=8 (Don't know)

LOG_PCE

- Use in the analysis: Control variable.

- Description from the original study: “Explanatory variables include [...] household per capita expenditures (PCE)” (page 18)

- Variables used in the replication: ks02_ks1type_A, ks02x_ks1type_A, ks02_ks1type_B, ks02x_ks1type_B, ks02_ks1type_C, ks02x_ks1type_C, ks02_ks1type_D , ks02x_ks1type_D, ks02_ks1type_E, ks02x_ks1type_E, ks02_ks1type_F, ks02x_ks1type_F, ks02_ks1type_G, ks02x_ks1type_G, ks02_ks1type_H, ks02x_ks1type_H, ks02_ks1type_I, ks02x_ks1type_I, ks02_ks1type_J, ks02x_ks1type_J, ks02_ks1type_K, ks02x_ks1type_K, ks02_ks1type_L, ks02x_ks1type_L, ks02_ks1type_M, ks02x_ks1type_M, ks02_ks1type_N, ks02x_ks1type_N, ks02_ks1type_OA , ks02x_ks1type_OA, ks02_ks1type_OB, ks02x_ks1type_OB, ks02_ks1type_P, ks02x_ks1type_P, ks02_ks1type_Q, ks02x_ks1type_Q, ks02_ks1type_R, ks02x_ks1type_R, ks02_ks1type_S, ks02x_ks1type_S, ks02_ks1type_T, ks02x_ks1type_T, ks02_ks1type_U, ks02x_ks1type_U, ks02x_ks1type_V, ks02x_ks1type_V, ks02_ks1type_W, ks02x_ks1type_W, ks02_ks1type_X, ks02x_ks1type_X, ks02_ks1type_Y, ks02x_ks1type_Y, ks02_ks1type_Z,ks02x_ks1type_Z, ks02_ks1type_AA, ks02x_ks1type_AA, ks02x_ks1type_BA, ks02x_ks1type_BA, ks02_ks1type_CA, ks02x_ks1type_CA, ks02_ks1type_DA, ks02x_ks1type_DA, ks02_ks1type_EA, ks02x_ks1type_EA, ks02_ks1type_FA, ks02x_ks1type_FA, ks02_ks1type_GA, ks02x_ks1type_GA, ks02_ks1type_HA, ks02x_ks1type_HA, ks02_ks1type_IA, ks02x_ks1type_IA, ks02_ks1type_IB, ks02x_ks1type_IB, ks06_ks2type_A1, ks06x_ks2type_A1, ks06_ks2type_A2, ks06x_ks2type_A2, ks06_ks2type_A3, ks06x_ks2type_A3, ks06_ks2type_A4, ks06x_ks2type_A4, ks06_ks2type_B, ks06x_ks2type_B, ks06_ks2type_C, ks06x_ks2type_C, ks06_ks2type_C1, ks06x_ks2type_C1, ks06_ks2type_D, ks06x_ks2type_D, ks06_ks2type_E, ks06x_ks2type_E, ks06_ks2type_F1, ks06x_ks2type_F1, ks06_ks2type_F2, ks06x_ks2type_F2, ks08_ks3type_A, ks08x_ks3type_A, ks08_ks3type_B, ks08x_ks3type_B, ks08_ks3type_C, ks08x_ks3type_C, ks08_ks3type_D, ks08x_ks3type_D, ks08_ks3type_E, ks08x_ks3type_E, ks08_ks3type_F, ks08x_ks3type_F, hh_size

Use the categorization of consumption items listed under ks in the questionnaire (page 32 of the pdf).

First, consider the items that were consumed by all the members of this household during the past week:

1. Stable Foods (Variable: STAPLE_FOOD):
    a. ks02_ks1type_A (Hulled, uncooked rice)

    b. ks02_ks1type_B (Corn)
    c. ks02_ks1type_C (Sago/flour)
    d. ks02_ks1type_D (Cassava, tapioca, dried cassava)
    e. ks02_ks1type_E (Other staple foods, like sweet potatoes, potatoes, yams)
2. Vegetables (Variable: VEGETABLES)
    a. ks02_ks1type_F (Kangkung, cucumber, spinach, mustard greens, tomatoes, cabbage, katuk, green beans, string beans and the like)
    b. ks02_ks1type_G (Beans like mung-beans, peanuts, soya-beans, and the like.)
    c. ks02_ks1type_H (Fruits like papaya, mango, banana and the like.)
3. Dried Foods (Variables: DRIED)
    a. ks02_ks1type_I (Noodles, rice noodles, macaroni, shrimp chips, other chips, and the like)
    b. ks02_ks1type_J (Cookies, breads, crackers)
4. Meat and fish (Variable: MEAT_FISH)
    a. ks02_ks1type_K (Beef, mutton, water buffalo meat and the like)
    b. ks02_ks1type_L (Chicken, duck and the like)
    c. ks02_ks1type_M (Fresh fish, oysters, shrimp, squid and the like)
    d. ks02_ks1type_N (Salted fish, smoked fish)
5. Other dishes, like (Variable: OTHER_DISHES):
    a. ks02_ks1type_OA (Jerky, shredded beef, canned meat, sardine and the like)
    b. ks02_ks1type_OB (Tofu, tempe, other side dishes)
6. Milk/Eggs (Variable: MILK_EGGS):
    a. ks02_ks1type_P (Eggs)
    b. ks02_ks1type_Q (Fresh milk, canned milk, powdered milk and the like)
7. Spices (Variable: SPICES):
    a. ks02_ks1type_R (Sweet and salty soy sauce)
    b. ks02_ks1type_S (Salt)
    c. ks02_ks1type_T (Shrimp paste)
    d. ks02_ks1type_U (Chili sauce, tomato sauce, and the like)
    e. ks02_ks1type_V (Shallot, garlic, chili, candle nuts, corriander, MSG and the like)
    f. ks02_ks1type_W (Javanese (brown) sugar)
    g. ks02_ks1type_X (Butter)
    h. ks02_ks1type_Y (Cooking oil like coconut oil, peanut oil, corn oil, palm oil and the like)
8. Beverages and other drinks/consumer products (Variable: BEVERAGES):
    a. ks02_ks1type_Z (Drinking water)
    b. ks02_ks1type_AA (Granulated sugar)
    c. ks02_ks1type_BA (Coffee)
    d. ks02_ks1type_CA (Tea)

    e. ks02_ks1type_DA (Cocoa)
    f. ks02_ks1type_EA (Soft drinks like Fanta, Sprite, etc.)
    g. ks02_ks1type_FA (Alcoholic beverages like beer, palm wine, rice wine, etc.)
    h. ks02_ks1type_GA (Betel nut (for chewing, traditional drug, others))
    i. ks02_ks1type_HA (Cigarettes, tobacco)
    j. ks02_ks1type_IA (Prepared food (eaten at home))
    k. ks02_ks1type_IB (Prepared food (away from home))

Second, consider the items that were consumed by all the members of this household during the past month:

1. Electricity (Variable: ELECTRICITY)
    a. ks06_ks2type_A1 (Electricity)
2. Water (Variable: WATER)
    a. ks06_ks2type_A2 (Water)
3. Fuel (Variable: FUEL)
    a. ks06_ks2type_A3 (Fuel)
4. Telephone (including vouchers and mobile starter pack) (Variable: TELEPHONE)
    a. ks06_ks2type_A4
5. Personal toiletries (Including soap, shaving supplies, cosmetics and the like) (Variable: TOILETRIES)
    a. ks06_ks2type_B
6. Household items (Including laundry soap, cleaning supplies, anti-mosquitoes and the like) (Variable: HH_ITEMS)
    a. ks06_ks2type_C
7. Domestic services and servants' wages (Variable: DOMESTIC_SERV)
    a. ks06_ks2type_C1
8. Recreation and Entertainment (Including movies, theater, outings, sport equipment, newspapers, magazines and the like) (Variable: RECREATION)
    a. ks06_ks2type_D
9. Transportation (Including bus fare, cab fare, vehicle repair costs, gasoline and the like) (Variable: TRANSPORTATION)
    a. ks06_ks2type_E
10. Sweepstakes and the like (Variable: SWEEPSTAKES)
    a. ks06_ks2type_F1
11. Arisan (Variable: ARISAN)
    a. ks06_ks2type_F2

Third, consider the items that were consumed by all the members of this household during the past one year:

1. Clothing for children and adults (Including shoes, hats, shirts, pants, children clothing and the like) (Variable: CLOTHING)

    a. ks08_ks3type_A
2. Household supplies and furniture (Including tables, chairs, kitchen tools, bed sheets, towels and the like) (Variable: HH_SUPPLIES)
    a. ks08_ks3type_B
3. Medical costs (Including hospitalization costs, clinic charges, physician's fee, traditional healer's fee, medicines and the like) (Variable: MEDICAL_COSTS)
    a. ks08_ks3type_C
4. Ritual ceremonies, charities and gifts (Including weddings, circumcisions, tithe, charities, gifts and the like) (Variable: RITUAL)
    a. ks08_ks3type_D
5. Taxes (Including property tax, vehicle tax, income tax, sales tax and the like) (Variable: Taxes)
    a. ks08_ks3type_E
6. Other expenditures not specified above (Including the purchase of cars, house, television sets, handphones, beds, livestock and the like) (Variable: OTHER_EXP)
    a. ks08x_ks3type_F

The study is not clear at what frequency household per capita expenditure is estimated. In the replication, monthly expenditure will be used. To do so:

1. Multiply by 4.34524 (weeks in a month) the value of items consumed weekly, i.e., STAPLE_FOOD, VEGETABLES, DRIED, MEAT_FISH, OTHER_DISHES, MILK_EGGS, SPICES, and BEVERAGES.
2. Divide the value of items consumed yearly by 12 (months in a year), i.e, CLOTHING, HH_SUPPLIES, MEDICAL_COSTS, RITUAL, TAXES, OTHER_EXP.

After adding all the items, divide by the number of household members (hh_size) to obtain per capita expenditure. Finally, take log of the household per capita expenditure.

- LOG_PCE corresponds to the log of household per capita expenditure.

- As explained in Section 12, most responses to the variables "Able to answer (by item)" are either 1 (given) or 3 (not given). However, in every case where the value of the volume of expenditure is 0, the value of "Able to answer" is 3 (not given). This suggests that the "Able to answer" question variable description is misleading since at least some of the responses of 0 of the volume of expenditure represent a household reporting no expenditures for that item. For the replication, all values of "Able to answer" equal to 3 are considered as a household reporting zero expenditures for that item. The complete list of variable where this happens is: ks02x_ks1type_A, ks02x_ks1type_B,ks02x_ks1type_C, ks02x_ks1type_D, ks02x_ks1type_E, ks02x_ks1type_F, ks02x_ks1type_G, ks02x_ks1type_H, ks02x_ks1type_I, ks02x_ks1type_J, ks02x_ks1type_K, ks02x_ks1type_L,

ks02x_ks1type_M, ks02x_ks1type_N, ks02x_ks1type_OA, ks02x_ks1type_OB, ks02x_ks1type_P, ks02x_ks1type_Q, ks02x_ks1type_R, ks02_ks1type_S, ks02x_ks1type_T, ks02x_ks1type_U, ks02x_ks1type_V, ks02x_ks1type_W, ks02x_ks1type_X, ks02x_ks1type_Y,ks02x_ks1type_Z, ks02x_ks1type_AA, ks02x_ks1type_BA, ks02x_ks1type_CA, ks02x_ks1type_DA, ks02x_ks1type_EA, ks02x_ks1type_FA, ks02x_ks1type_GA, ks02x_ks1type_HA, ks02x_ks1type_IA, ks02x_ks1type_IB, ks06x_ks2type_A1, ks06x_ks2type_A2, ks06x_ks2type_A3, ks06x_ks2type_A4, ks06x_ks2type_B, ks06x_ks2type_C, ks06x_ks2type_C1, ks06x_ks2type_D, ks06x_ks2type_E, ks06x_ks2type_F1, ks06x_ks2type_F2, ks08x_ks3type_A, ks08x_ks3type_B, ks08x_ks3type_C, ks08x_ks3type_D, ks08x_ks3type_E, ks08x_ks3type_F

Moreover, the expenditures categories for transfers given to other parties outside the household are not included in the estimation of total household expenditures. The reason is that it is not clear that it captures additional spending over and above the other expenditures categories. This corresponds to the variables ks06_ks2type_G, ks06x_ks2type_G, ks08_ks3type_G, and ks08x_ks3type_G.

FEMALE

- Use in the analysis: Control variable.
- Description from the original study: “Explanatory variables include [...] a sex dummy” (page 18)
- Variables used in the replication: ar07.
- FEMALE corresponds to a dummy variable equal to one if the respondent is a female and zero otherwise.
- No deviations between the original study and the replication study.

POOR_HEALTH

- Use in the analysis: Sample parameter variable.
- Description from the original study: “we estimate these effects for two separate subsamples of individuals: those in good general health and those in poor general health. Respondents were asked to evaluate their general health status (GHS) on a scale from 1 to 4. Depending on the answers provided, we split the sample in two groups: a healthy group containing respondents who characterized their general health status as being either “very healthy” or “somewhat healthy”, and an unhealthy group containing respondents who claimed they were either ”unhealthy” or “somewhat unhealthy” (page 18)

- Variables used in the replication: kk01.

  Poor health is defined as kk01=3 (Somewhat unhealthy) or kk01=4 (Very unhealthy).

  - POOR_HEALTH corresponds to a dummy variable equal to 1 if respondent is in poor general health and 0 otherwise.

- No deviations between the original study and the replication study.

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?*

List of composite measure variables in the replication:

- UNDER_DIAG, the dependent variable. It is constructed using the "hypertension" and the "Diagnosed with chronic condition" (cd05) variables. At the same time, "hypertension" is obtained after taking an average of the "systolic" and "diastolic", and then checking if for a person his systolic is greater than 140 or his diastolic is greater than 90. Lines 8 to 31 of the Stata do-file "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" construct the UNDER_DIAG variable.

- RISK_PREFERENCES, a control variable. It is constructed following Ng "Risk and Time Preferences in Indonesia: The Role of Demographics, Cognition, and Interviewers." (2012), where risk preferences are elicited through the method of hypothetical lottery-choice questions. Lines 170 to 207 of the Stata do-file "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" construct the RISK_PREFERENCES variable.

- TIME_PREFERENCES, a control variable. It is constructed following Ng "Risk and Time Preferences in Indonesia: The Role of Demographics, Cognition, and Interviewers." (2012), where time preferences are elicited through the experimental method of asking respondents to choose between a smaller, immediate payoff and a larger, delayed payoff method of hypothetical lottery-choice questions. Lines 209 to 243 of the Stata

do-file "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" construct the RISK_PREFERENCES variable.

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?*

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

Lines 4 to 6 of the Stata do-file "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" generate a random sample that contains only 5% of the data.

To replicate the analysis a probit regression model is estimated. The dependent variable is a dummy equal to 1 for those respondents who were found to be hypertensive during the IFLS screenings, but were not previously diagnosed by a doctor, and 0 otherwise. The focal independent variable is the respondents' years of formal education. As control variables, it is included age, age squared, risk preferences, time preferences, distance to medical facility, household expenditures, and a sex dummy. After estimating the probit regression, the marginal effect of the respondents' years of formal education is obtained. Due to the concerns with the distance variable, a probit regression excluding that control variable is also proposed. The software used is Stata 15.1

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?*

According to Table 2, the log of Household per Capita expenditure should be taken as the control variable.

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***).

To test the SCORE claim, **H*,** that among the sample of respondents in poor general health who were found to be hypertensive during a screening, the probability of being undiagnosed decreases with education, a probit regression is estimated. The specification takes the form:

$$P(UNDER\_DIAG = 1 \mid YRS\_SCHOOL, X) = \Phi(\alpha + \gamma YRS\_SCHOOL + \beta X) \ \ if\ POOR\_HEALTH = 1$$

Where:

1. UNDER_DIAG: a dummy equal to 1 for those respondents who were found to be hypertensive during the IFLS screenings, but were not previously diagnosed by a doctor, and 0 otherwise. It is the dependent variable.
2. YRS_SCHOOL: correspond to respondents' years of formal education. It is the focal independent variable. The coefficient of interest is γ, which indicates how the probability of being under-diagnosed varies with education.
3. α : constant term
4. X: vector of control variables. According to page 18, the probit regression model for being under-diagnosed contains eight covariates, detailed in Table 4. One control variable is YRS_SCHOOL, which is the focal independent variable in the replication. Therefore, the remaining seven covariates are:
    1. AGE
    2. AGESQRT
    3. RISK_PREFERENCE
    4. TIME_PREFERENCE
    5. DISTANCE
    6. LOG_PCE
    7. FEMALE

    NOTE: Due to the issues with the DISTANCE variable (described in detail in Sections 12 and 17), two probit regressions will be estimated. The difference is that in one of them the vector X will exclude the DISTANCE variable as a control variable.

The replication study will prioritize the probit regression without the DISTANCE variable as the final test for the focal hypothesis. The reason is to reduce the limitations arising from the measurement of the distance variable.

5. β : vector of coefficients for X
6. Φ is the cumulative standard distribution.
7. POOR_HEALTH: dummy variable equal to 1 if respondent is in poor general health and 0 otherwise.

Then, using the results from the probit regression, the marginal effect of YRS_SCHOOL is obtained to test the SCORE claim (**H***). This is in line with Kim & Radoias (2016) page 18: “For all our estimations, we use a probit model and report the marginal effects of the explanatory variables.”

22. Data exclusion

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

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

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

The only reason to exclude observations in the analysis is due to missing values, as will be described in Section 23.

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?*

For some respondents in the sample, some variables are reported as "Don't know" or "MISSING." (more details in Section 17) In those cases, the variable value is recorded as missing. Consequently, the respondent is excluded from the probit regression in Section 22.

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.*

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?*

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?

      The concern could be the quality of the distance from the closest health center variable. As was detailed in Sections 12 and 17, the variable in the dataset contains several limitations. Aside from that variable, all information seems of high quality for the replication. Two probit regressions are proposed to reduce the issues related to the quality of the distance variable. One probit regression includes the distance as a regressor, but not the other one.

    - 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?

      Yes. All the steps are detailed in this document and the Stata do-files.

- 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?

      Yes.

    - Has the data analyst demonstrated that the analysis code works as expected on a random 5% of the final replication dataset?

      Yes. The analysis code is "Kim & Radoias 2016 - Replication Analysis 5% random sample.do" A log file with the Stata session demonstrates that this code works using a random 5% of the final replication dataset (Kim-Radoias_Replication_5_Random_Sample.pdf)

- I have reviewed all sections of this pre-registration, and I believe it represents a good-faith replication attempt of the original focal claim.

  Yes.