Data Analytic Replication of a research claim from Anderson (2011) in *American Economic Journal: Applied Economics*

Reproduction Team: Nathaniel D. Porter, Anirudh Tagat, & Jing Geng

Independent Reviewer(s)
(add name below when you initiate review, comment “DONE” on your name when you finish and notify the reproduction team):

Action Editor: Tabaré Capitán

As necessary:

Reviewer #1: Daniel Dunleavy

Reviewer #2: Hansika Kapoor

Review Period: April 26th, 2022 - - May 3rd, 2022

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

Privacy Statement: Other teams are making predictions about the outcomes of many different studies, not knowing which studies have been selected for reproduction. 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 reproduction and all aspects of the discussion about these reproduction designs.

Preregistration of Anderson_AmEcoJourn_2011_bLe8

Existing Data Replication

(An example of a filled-in preregistration form is available here.)

Study Information

1. Title (provided by SCORE)

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

Replication of a research claim from Anderson (2011) in *American Economic Journal: Applied Economics*.

2. Authors and affiliations

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

Dr. Nathaniel D. Porter[1]
Dr. Anirudh Tagat[2]
Jing Geng[3]

1 Virginia Tech, University Libraries & Sociology
2 Monk Prayogshala, Economics
3 Virginia Tech, Sociology

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 Anderson (2011) is that agricultural yields are systematically higher for low-caste households residing in villages dominated by lower castes (BACs), in terms of total land ownership, compared to villages dominated by upper castes. This reflects the following statement from the paper's abstract: "The key finding is that income is substantially higher for low-caste households residing in villages dominated by a low caste." Table 3 reports the main estimations results from an ordinary least squares estimation of equation (1). The sample in the estimations are the lower castes (BAC - lower backward agricultural castes, OBC - other backward castes, SC - schedule casts). Dependent variable is “crop income per acre of total land of a household”. Focal independent variable is “low-caste

villages". The portion of Table 4 selected is Column (1). The estimation results from Table 3 confirm the robustness of the positive relationship between agricultural income and residing in a low-caste dominated village. [The coefficient for "Low-caste villages" is 566.5 with robust standard errors of 209, significant at the 1 percent level.]

All public data, code, and study documentation can be found in the OSF project at https://osf.io/t95k2/.

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 low-caste households, residing in villages dominated by lower castes is associated with greater agricultural income compared to residing in villages dominated by upper castes.

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.*

There is no blinding involved.

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 data collected in the original study is from two northern states in India: Uttar Pradesh (UP) and Bihar, and were collected between 1997-98. The dataset covered 12 districts in Uttar Pradesh and 13 districts in Bihar. A total of 120 villages (57 in Bihar; 63 in UP), with an overall sample size of 2,250 households.

For replication, we will use the 1999 wave of the Additional Rural Incomes Survey (ARIS-REDS) collected by the National Council of Applied Economic Research (NCAER), New Delhi. The dataset covers village-level data on land ownership, caste composition, and other characteristics for 253 villages in 17 states. It also contains extensive census (listing) data and detailed household-level information on land ownership patterns. The unit of analysis is the household. The main deviation between the two datasets is that the replication study includes all Indian states, not just Bihar and Uttar Pradesh. The other potential deviation is the identification of Backward Agricultural Castes (BACs), which are important for testing the focal claim in the paper. There is no ready classification for this in ARIS-REDS, and will need to be coded by hand or manually against a reference that delineates caste identity into the groups specified in the paper. This is possible because ARIS-REDS contains detailed data on caste and sub-caste at the household level.

Apart from this sample issue, there are no other foreseeable deviations, since the ARIS-REDS datasets have substantial demographic information, particularly related to caste composition and land ownership, which are the two main variables of interest.

9. Randomization (free response)

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

No randomization.

Sampling Plan

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

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

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

11. Explanation of existing data

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

The replication data consists of the ARIS-REDS 1999 data that requires human-subjects research approval prior to access cross-walk identifiers (e.g. district or state identifiers). The replication sample will only include members of the lower castes (BAC, OBC, SC and ST) who own land.

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.*

Dependent Variable

Crop household income per acre of total land (p.247)

- UP-Bihar LSMS World Bank (household questionnaire); n = 2250 (overall); n = 1295 (estimation)
- Data from 1997-1998
- Geographical location: Household-level information from UP and Bihar

- Crop income per acre is equal to the total value of sales of all crops over the past year divided by the total land (p.246, table 2)
- Non-landowners have missing values on this variable

Focal Independent Variable(s)

Dummy variable for village dominated by lower caste

- UP-Bihar LSMS World Bank (village-level questionnaire p.244); n = 120 villages (overall); n = 90 villages (estimation; p.245 table 1)
- Data from 1997-1998
- Geographical location: UP and Bihar

Control Variable(s)

Education

- UP-Bihar LSMS World Bank (household questionnaire); n = 2250 (overall); n = 1295 (estimation)
- Data from 1997-1998
- Geographical location: Household-level information from UP and Bihar
- No clear definition of literacy provided by authors, but coded so 1 = literate
- Education variable in original data: 1 = Illiterate; 0 and 2-11 indicate literate and/or educated.

Land ownership

- UP-Bihar LSMS World Bank (household questionnaire); n = 2250 (overall); n = 1295 (estimation)
- Data from 1997-1998
- Geographical location: Household-level information from UP and Bihar

Caste identity

- UP-Bihar LSMS World Bank (household questionnaire); n = 2250 (overall); n = 1295 (estimation)
- Data from 1997-1998
- Geographical location: Household-level information from UP and Bihar
- Either belonging to backward agricultural castes (BACs), other backward castes (OBCs), or scheduled castes (SCs).
- “The BAC group represents the traditional farming castes, and the OBC group represents the traditional artisan castes. The BAC group is ranked higher than OBC. The SC are the lowest in the caste ranking, formerly known as the untouchable castes.” (footnote 2, p.240)

State fixed effects

- Indicator of State (Bihar or Uttar Pradesh)

Sample Parameters

"The sample in the estimations are the lower castes (BAC, OBC, SC)" (p.248, table 3). Additionally, the sample is limited to landowners, due to the inclusion of land area in the denominator of the dependent variable.

Sample size of original analysis has 1295 observations (households).

(b) Data Access

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

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

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

The ARIS-REDS data can be used to replicate this focal claim. The dataset is available from Andrew Foster's homepage at Brown University, but comes with little to no technical support (readme). The dataset is part of household survey data collected by NCAER, and is available for 1999. Instructions on accessing the dataset(s) are available here. Following IRB approval, the authors of the replication study gained access to cross-walk identifiers to merge household data with village data on a secure computer.

Final replication data does not include restricted data, due to changes in how the variables were defined and data are included in the OSF repository for replication.

(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.***

Crop Income per Acre (Dependent Variable)

Source data

- rd99015.dta (household data aggregated from members)
    - q15: “Area under crop”
    - q50: “Total receipts”
- rd99043 (household)
    - q51 (Total miscellaneous receipts from farming)

Crop income per acre = Agricultural Receipts/Area under cultivation = (rd99015.q50 + rd99043.q51)/rd99015.q15

Village dominated by lower caste (key Independent Variable)

- Datasets: rd99001, rd99010 and rd99012
- Description: The land owned by low-caste households is calculated by summing land_owned (see below) across all low-caste respondents. Land owned by high-caste households is calculated in the same way across remaining respondents. The village is considered dominated by lower castes (1) if low-castes own more land than other castes (0) in total.
- This variable is calculated before removing high caste respondents from the sample.
- See variables below (caste and land ownership) for construction of caste and ownership

Literacy/Education (control)

- Dataset: “arisreds_data/public99/hhecon/rd99002.dta”
    - Variable q15: ‘Education’
- The original variable is a categorical variable, but is rank ordered, i.e. higher value means more education. It takes values between 1 (“Illiterate”) and 20 (“Professional and Technical Post-graduate level degree such as M.E., M.Ed etc.”)
- *The recoded variable is coded to 1 if any household member has formal education or reports being literate despite no formal education.*

Land ownership (control)

- Two sources:

    - Dataset: “arisreds_data/public99/hhecon/rd99010.dta”
        - Variable q8: “Land owned (in acres)” (excludes inheritance)
        - Variable q11: “Land inherited”
    - Dataset: “arisreds_data/public99/hhecon/rd99012.dta”
        - Variable q35: “Total land owned at the end of RP in acres”
- Coding:
    - Land ownership is by the individual respondent (not summed across all brothers)
    - Land owned is equal to rd99012.q35 if valid. Missing values are replaced using the sum of rd99010.q8 and rd99010.q11 when available.

Lower Caste (selection)

- Two sources
    - Dataset: “arisreds_data/public99/hhecon/rd99001.dta”
        - Variable q11: “Caste (annex 2)”
        - Variable q51: “Caste group”
- Lower caste is defined as:
    - SC (3), ST (4) or BC (5) in q51 OR
    - BAC (Yadav) when present in a state (q11) OR
    - Castes labeled as SC/ST/BC/OBC in q11 OR
    - State-specific OBC lists (for Bihar and Uttar Pradesh only) in q11 OR
    - Scheduled Castes or Scheduled Tribes = State-specific SC lists (for Bihar and Uttar Pradesh only) in q11
- Dummy coding 1=lowercaste, 0 = other caste
- Lower caste is used in “village dominated by lower caste” variable and in selection (final sample is only lower-caste respondents)

State and Caste identity (fixed effects)

- Dataset: “arisreds_data/public99/hhecon/rd99001.dta”
    - Variable q8: “State Code”
    - Variable q11: “Caste”
- States: Bihar and Uttar Pradesh
- Because caste lists vary by state (e.g. code values and castes present differ for each state - see REDS99-Appendix) these are treated as an interaction such that a fixed effect is included for each caste in each state, regardless of whether a caste of the same name exists in a different state

(d) Data Creation

RR TEAM INSTRUCTIONS: *Create a dataset using the data sources and variables listed above. Provide a detailed narrative describing how the various datasets were cleaned and merged into a final replication dataset. Provide a view-only link to a clearly commented script on the OSF that produces the replication data as described in the narrative. Our preference is that*

this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible.

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

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

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

Please also use this section to describe:

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

Data were downloaded from the ARIS-REDS webpage (public) Data were cleaned using the script anderson_2011_replication_data_cleaning.do in Stata MP-4 17.0 for Windows.

(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.

Data dictionary: https://osf.io/gsvuz/ in OSF folder

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.***

Final sample size is 2,537 households from 216 villages across all 16 states (with between 2 and 30 villages per state)

SAMPLE DERIVATION (AFTER EACH STEP):
Original ARIS-REDS 1999 sample: 7,474
Subset of low-caste households: 4,244
Subset of land owners: 3,166
Subset of households who farmed at least 1 acre of their land: 2,537
No additional case loss due to missing values for the following variables: literacy, state, caste, village, village dominated by low-caste

---

Required sample size [to be filled out by the SCORE team]: The primary unit of analysis is the household clustered in villages. An estimate of the minimum viable sample size for the data analytic replication is: 678. For comparison, the stage1 required sample size would be: 3293 and the stage2 sample size would be: 7409. [**Note:** Given the clustering of standard errors at the village level, all else being equal, a replication dataset that reaches the target N by having more different villages and fewer responses per village will have more power than one that reaches the same target N with fewer villages but more responses per village.]

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.

15. Stopping rule (provided by SCORE)

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

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

Variables

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

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

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

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

16. Manipulated variables

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

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

17. Measured variables

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

attempt. In cases where you are adding a variable that was not present in the original study, please explicitly state that you are doing so, and explain how, if at all, you expect it will affect the replication attempt.

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

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

NOTE: THIS SECTION DUPLICATES THE VARIABLES REQUIRED SECTION ABOVE

Crop Income per Acre (Dependent Variable)

Source data

- rd99015.dta (household data aggregated from members)
    - q15: “Area under crop”
    - q50: “Total receipts”
- rd99043 (household)
    - q51 (Total miscellaneous receipts from farming)

Crop income per acre = Agricultural Receipts/Area under cultivation = (rd99015.q50 + rd99043.q51)/rd99015.q15

Village dominated by lower caste (key Independent Variable)

- Datasets: rd99001, rd99010 and rd99012
- Description: The land owned by low-caste households is calculated by summing land_owned (see below) across all low-caste respondents. Land owned by high-caste households is calculated in the same way across remaining respondents. The village is considered dominated by lower castes (1) if low-castes own more land than other castes (0) in total.
- This variable is calculated before removing high caste respondents from the sample.
- See variables below (caste and land ownership) for construction of caste and ownership

Literacy/Education (control)

- Dataset: “arisreds_data/public99/hhecon/rd99002.dta”
    - Variable q15: ‘Education’
- The original variable is a categorical variable, but is rank ordered, i.e. higher value means more education. It takes values between 1 (“Illiterate”) and 20 (“Professional and Technical Post-graduate level degree such as M.E., M.Ed etc.”)
- *The recoded variable is coded to 1 if any household member has formal education or reports being literate despite no formal education.*

Land ownership (control)

- Two sources:
    - Dataset: “arisreds_data/public99/hhecon/rd99010.dta”
        - Variable q8: “Land owned (in acres)” (excludes inheritance)
        - Variable q11: “Land inherited”
    - Dataset: “arisreds_data/public99/hhecon/rd99012.dta”
        - Variable q35: “Total land owned at the end of RP in acres”
- Coding:
    - Land ownership is by the individual respondent (not summed across all brothers)
    - Land owned is equal to rd99012.q35 if valid. Missing values are replaced using the sum of rd99010.q8 and rd99010.q11 when available.

Lower Caste (selection)

- Two sources
    - Dataset: “arisreds_data/public99/hhecon/rd99001.dta”
        - Variable q11: “Caste (annex 2)”
        - Variable q51: “Caste group”
- Lower caste is defined as:
    - SC (3), ST (4) or BC (5) in q51 OR
    - BAC (Yadav) when present in a state (q11) OR
    - Castes labeled as SC/ST/BC/OBC in q11 OR
    - State-specific OBC lists (for Bihar and Uttar Pradesh only) in q11 OR
    - Scheduled Castes or Scheduled Tribes = State-specific SC lists (for Bihar and Uttar Pradesh only) in q11
- Dummy coding 1=lowercaste, 0 = other caste
- Lower caste is used in “village dominated by lower caste” variable and in selection (final sample is only lower-caste respondents)

State and Caste identity (fixed effects)

- Dataset: “arisreds_data/public99/hhecon/rd99001.dta”
    - Variable q8: “State Code”
    - Variable q11: “Caste”
- States: Bihar and Uttar Pradesh
- Because caste lists vary by state (e.g. code values and castes present differ for each state - see REDS99-Appendix) these are treated as an interaction such that a fixed effect is included for each caste in each state, regardless of whether a caste of the same name exists in a different state

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?*

Crop Income per Acre, Village Dominated by Lower Caste, and Low Caste are constructed as described in section 17 and reflected in anderson_2011_replication_data_preparation.do.

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).

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?*

Analysis is conducted in Stata to match the format of the original datafiles and simplify combination of multiple files.Code is in the script anderson_2011_replication_data_analysis.do, with sampled analysis active and commented lines for full analysis (not run at this time).

The analysis will consist of OLS regression on *Household Crop Income* (per acre), with the predictors *Low-caste village* (focal IV), *Literate* and *Land Owned*. Additional fixed effects are included for State and Caste using a factorial interaction to reflect caste differences between states but not reported. Standard errors will be clustered by village.

Regression code and output from 5% sample are included below, with coefficients for fixed effects suppressed and the coefficient and p-value for the focal test in bold.

	raw_inc_per_acre
Literate_hh (b)	3.349
SE	(16.470)
Test statistic	0.20
P	0.84
land_owned	-0.007
	(0.007)
	-0.98
	0.33
locaste_land_v	**-46.102**
	(101.327)
	-0.45
	0.65
Number of observations	127

```
R-squared                          0.52
Adjusted R-squared                -0.24
AIC                             1521.22
BIC                             1589.48
Log likelihood                  -736.61
--------------------------------------
```

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

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?*

No transformations.

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.

H* will be considered confirmed if the coefficient for locaste_land_v (Low-caste village) is positive and significant at the p<0.05 level.

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?)*

No additional exclusion.

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?*

Missing data will be removed using listwise deletion of cases with missing values for land owned and missing or zero values for crop area (the dependent variable cannot be calculated with a denominator of zero) by explicitly dropping cases prior to saving the dataset for analysis. All other variables are complete in the sample. Low caste has no missing data, but the sample size is reduced because the analytic sample is only low caste households.

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.*

To assess whether the decision to use net vs raw income (not clearly specified by original authors) is consequential, the primary analysis will be conducted a second time with new agricultural income (by subtracting crop, labor, etc costs before dividing by area under cultivation).

Additionally, a separate analysis using raw income will be conducted of cases exclusively in the two states used in the original analysis (Bihar and Uttar Pradesh), although as a consequence, its power will be limited by sample size.

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?*

All materials included in pre-registration and final analysis will be made public through an OSF repository or open-access data repository.

Final review checklist

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

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