

README for the Reproducibility Package for Validation of a Cross-Linguistic Pictorial Socioeconomic Status Instrument for Illiterate and Diverse Global Populations in the Technological Age

Overview

The code in this replication package prepares the analysis files and executes the statistical modeling to support the manuscript titled "SES Validation Manuscript." The workflow relies on data gathered from three linguistic groups (Bangla, French, and Zulu). The data cleaning and descriptive analyses, including the generation of two tables and an inter-item correlation statistic, are performed using R. Confirmatory Factor Analysis (CFA) and Measurement Invariance (MI) testing, as well as Structural Equation Modeling (SEM) to establish nomological validity, are executed using standalone Mplus software. The main scripts in the `02_Code/` directory systematically clean the raw datasets, generate formatting for Tables 2 and 3, and provide the syntax to run the necessary models. These outputs confirm the measurement model fit and validate the invariance of the instrument across language groups. The expected runtime for the entire package is approximately 5 to 10 minutes on a standard computer.

Data Availability Statement

1. Pictorial SES Surveys from 2022-2023

- **Source:** Databases from Curious Learning
- **Year:** 2024
- **URL:** Forthcoming (<https://reproducibility.worldbank.org>)
- **Access Date (Month-Year):** Data collection period: 2022-12-01 to 2024-03-31
- **Note:** File locations: `01_Data/Bangla_2022-12-01_2023-08-01.csv`, `01_Data/French_2022-12-01_2023-08-01.csv`, and `01_Data/Zulu_2022-09-01_2024-03-31.csv`; Variables used include country, language, SES survey items (BQ1 through BQ8), and IRT scores for early literacy skills. These data will be published as part of the World Bank reproducibility archive.
- **Access Type:** Forthcoming - Data collected by Curious Learning and will be publicly deposited at <https://reproducibility.worldbank.org> upon publication acceptance.
- **License:** To be determined upon publication
- **License URL:** To be determined upon publication
- **Citation:** Curious Learning. 2024. "Pictorial SES Surveys from 2022-2023" [dataset]. World Bank Reproducibility Archive (forthcoming). <https://reproducibility.worldbank.org>

Statement about Rights

I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Instructions for Replicators

To produce the exhibits and analyses presented in the manuscript, a new user should follow the steps below in sequence.

1. Install Required Software and Packages

Ensure you have recent versions of R, RStudio, and Mplus installed on your computer. The main script will automatically check for and install all required R packages from CRAN (whatever the current CRAN version is at the time you run it). Required packages are:

- dplyr ($\geq 1.0.0$)
- tidyr ($\geq 1.2.0$)
- psych ($\geq 2.0.0$)
- knitr (≥ 1.30)
- kableExtra ($\geq 1.3.0$)
- rmarkdown ($\geq 2.0.0$)

Note: `renv.lock` records the exact package versions used to produce the results shipped in `03\_Outputs/` (dplyr 1.0.7, tidyr 1.2.0, psych 2.2.3, knitr 1.38, kableExtra 1.3.4, rmarkdown 2.14), for reference or for replicators who want to reproduce those exact versions themselves via `renv::restore()`. It is not applied automatically by `main.R`.

2. Extract the Replication Package

Unzip the replication package to a local directory on your machine. Do not alter the folder structure (e.g., `01\_Data`, `02\_Code`, `03\_Outputs`). The scripts are programmed with relative paths and will automatically locate the necessary files as long as the folders remain in their original arrangement.

3. Run the R Data Preparation and Analysis Scripts

One-click execution using the main script (recommended):

Open RStudio and navigate to the package directory. Open `02\_Code/main.R` and:

1. Edit line 21 to set `project_root` to your local package directory (e.g., `"C:/Users/YourName/Desktop/Reproducibility Package_SES Validation"`)
2. Run the entire script (Ctrl+A, then Ctrl+Enter)
3. The main script will run all R analyses in sequence (~2-3 minutes):
 - `00\_Data Cleaning.Rmd`: Processes raw data and generates `Combined.csv` and `Combined\_Mplus.csv`
 - `01\_Inter-item\_Correlation.Rmd`: Calculates internal consistency (outputs `01\_Inter-item\_Correlation.txt`)
 - `02\_Table 2.Rmd`: Generates Table 2 summary (outputs `02\_Table2\_Formatted.html` and `.csv`)
 - `03\_Table 3.Rmd`: Generates Table 3 statistics (outputs `03\_Table3\_Formatted.html` and `.csv`)

4. Run the Mplus Models

Open the standalone Mplus application and sequentially run the following syntax files located in the `02\_Code` folder. These scripts pull data directly from the `01\_Data` folder:

- **04_01_CFA_6Items.inp**: Runs the Confirmatory Factor Analysis.
- **04_02_MI_Testing.inp**: Conducts the Measurement Invariance testing (Configural and Scalar models).
- **05_SEM Correlation.inp**: Runs the Structural Equation Model to test nomological validity.

Note: Running Mplus interactively will generate **.out** log files in the **02_Code** folder. For convenience and package completeness, the final generated **.out** files are already provided for you in the **03_Outputs** folder.

5. Review Manually Constructed Exhibits

Because Table 4 is assembled in a word processor rather than programmatically, you can verify its construction by opening the document **04_04_Table 4 with Evidence.docx** in the **03_Outputs** folder. This document maps the final table directly to the specific outputs generated by the Mplus **.out** files.

Software Requirements

To successfully execute this replication package, the following software and libraries are required.

1. R and RStudio

- **R Software:** Version 4.0.0 or higher is recommended.
- **RStudio:** A recent version is required to natively compile and render the RMarkdown files.
- **R Packages (Dependencies):** The following libraries must be installed prior to running the scripts:
 - dplyr
 - tidyr
 - psych
 - knitr
 - kableExtra
 - rmarkdown

2. Mplus

- **Mplus Software:** Version 8.10 or higher. The **.out** files shipped in **03_Outputs/** were generated using Mplus Version 8.10 for Mac.

Memory and Storage Requirements

5-10 minutes

List of Exhibits

The tables and in-text statistics presented in the manuscript correspond to the generated output files as follows:

- **Table 1:** This table does not contain analysis results generated by the code and is constructed manually to describe the SES questions with collapsed responses.
- **Table 2 (Count and Percentage of Users by Country):** Generated programmatically by **02_Table 2.Rmd**. The final output is **03_Outputs/02_Table2_Formatted.html**.

- **Table 3 (Descriptive Statistics of the Responses from the SES Instrument):** Generated programmatically by `03\_Table 3.Rmd`. The final output is `03\_Outputs/03\_Table3\_Formatted.html`.
- **Table 4 (Results of the CFA Measurement Model and the Measurement Invariance Testing):** This table involves manual assembly and formatting. It is constructed using the statistical results found in the Mplus outputs `04\_01\_CFA\_6Items.out` and `04\_02\_MI\_Testing.out`. A mapping document, "04_04_Table 4 with Evidence.docx", is provided in the `03\_Outputs` folder to show exactly how the raw Mplus text output aligns with the final formatted table.
- **In-text Statistic (Internal Consistency):** The mean inter-item tetrachoric correlation of .23 is generated by `01\_Inter-item\_Correlation.Rmd` and saved in `03\_Outputs/01\_Inter-item\_Correlation.txt`.
- **In-text Statistic (Nomological Validity):** The correlation with early literacy skills of .34 is produced by the Mplus script `05\_SEM Correlation.inp`. The result can be found in the output log `03\_Outputs/05\_SEM Correlation.out`.
- **Figure 1 (Illustrations of Questions from the Pictorial Socioeconomic Survey):** Not generated by code. These are static illustrations created by artists during instrument development, not a data-driven exhibit.

Folder Structure

The replication package is neatly organized into four main directories. This structure separates the raw inputs, the processing scripts, the generated results, and the final manuscript.

- **01_Data/:** Contains all data files used in the project. The files with dates in their names (e.g., `Bangla\_2022-12-01\_2023-08-01.csv`) are the original, raw datasets. The cleaned individual language files (e.g., `Bangla.csv`) and the merged analysis files (`Combined.csv` and `Combined\_Mplus.csv`) are generated by the cleaning script.
- **02_Code/:** Contains all RMarkdown (`.Rmd`) and Mplus syntax (`.inp`) files. The files are numbered sequentially from 00 to 05 to guide the replicator through the exact order of execution.
- **03_Outputs/:** Stores all the results generated by the scripts in the `02\_Code/` folder. This includes formatted HTML tables, raw CSV data summaries, text files for in-text statistics, and Mplus output logs (`.out`). It also includes the document `04\_04\_Table 4 with Evidence.docx`, which maps the raw Mplus output to the manually constructed Table 4.
- **04_Draft/:** Contains the submitted manuscript document, `SES Validation Manuscript.docx`.