Reproduce the results of the article "Relationship of gender differences in preferences to economic development and gender equality"

Introduction

This study reproduces the results of the article Relationship of gender differences in preferences to economic development and gender equality (DOI: 10.1126/science.aas9899) and partially its supplementary material.

The following two relevant papers have to be also cited in all publications that make use of or refer in any kind to GPS dataset:

- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics*, 133 (4), 1645–1692.
- Falk, A., Becker, A., Dohmen, T. J., Huffman, D., & Sunde, U. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. IZA Discussion Paper No. 9674.

Below we describe how we collected, clean and standardized the the data, as well the the whole analysis pipeline. The output files allow the reader to reproduce the figures directly, bypassing the pipeline. For more details, especially on the background study and the meaning of the variables, refer to the main paper.

Preparation of the data

Data Collection, Cleaning, and Standardization

The data used by the authors is not fully available because of two reasons:

1. **Data paywall:** Some part of the data is not available for free. It requires to pay a fee to the Gallup to access them. This is the case for the additional data set that is used in the article, for instance, the one that contains the education level and the household income quintile. Check the website of the briq - Institute on Behavior & Inequality for more information on it.
2. **Data used in study is not available online:** This is what happened for the LogGDP p/c calculated in 2005 US dollars (which is not directly available online). We decided to calculate the LogGDP p/c in 2010 US dollars because it was easily available, which should not change the main findings of the article.

The procedure for cleaning is described for each data set, in the corresponding section below. After manually cleaning the data set, we standardized the names of the countries and merged the datasets into one within the function PrepareData.r.

Global Preferences Survey This data is protected by copyright and cannott be given to third parties.

To download the GPS data set, go to the website of the Global Preferences Survey in the section "downloads". There, choose the "Dataset" form and after filling it, we can download the data set.

Hint: The organisation can be also "private".

GDP per capita From the website of the World Bank, one can access the data about the GDP per capita on a certain set of years. We took the GDP per capita (constant 2010 US$), made an average of the data from 2003 until 2012 for all the available countries, and matched the names of the countries with the ones from the GPS data set.

Gender Equality Index The Gender Equality Index is composed of four main data sets.

- **Time since women's suffrage:** Taken from the Inter-Parliamentary Union Website. We prepared the data in the following way. For several countries more than one date where provided (for example, the right to be elected and the right to vote). We use the last date when both vote and stand for election right were granted, with no other restrictions commented. Some counties were a colony or within union of the countries (for instance, Kazakhstan in Soviet Union). For these countries, the rights to vote and be elected might be technically granted two times within union and as independent state. In this case we kept the first date. It was difficult to decide on South Africa because its history shows the racism part very entangled with women's rights. We kept the latest date when also Black women could vote. For Nigeria, considered the distinctions between North and South, we decided to keep only the North data because, again, it was showing the completeness of the country and it was the last date. Note: USA data doesn't take into account that also up to 1964 black women couldn't vote (in general, Blacks couldn't vote up to that year). We didn't keep this date, because it was not explicitly mentioned in the original data set. This is in contrast with other choices made, but it is important to reproduce exactly the results of the publication, and the USA is often easy to spot on the plots.
- **UN Gender Inequality Index:** Taken from the Human Development Report 2015. We kept only the table called "Gender Inequality Index".
- **WEF Global Gender Gap:** WEF Global Gender Gap Index Taken from the World Economic Forum Global Gender Gap Report 2015. For countries where data were missing, data was added from the World Economic Forum Global Gender Gap Report 2006. We modified some of the country names directly on the csv file, that is why we provide this as an input file.
- **Ratio of female and male labour force participation:** Average International Labour Organization estimates from 2003 to 2012 taken from the World Bank database (http://data.worldbank.org/indicator/SL.TLF.CACT.FM.ZS). Values were inverted to create an index of equality. We took the average for the period between 2004 and 2013.

About Missing Data

Main issue During the reproduction of the article, we found that the authors didn't write in details how they handled missing data in the indicators.

They mention on page 14 of the Supplementary Material, that (quoting): "For countries where data were missing data were added from the World Economic Forum Global Gender Gap Report 2006 (http://www3.weforum.org/docs/WEF_GenderGap_Report_2006.pdf)."

However, there are two problems here:

- regarding the year when women received the right to vote in a specific country. The missing values here are the ones coming from the United Arab Emirates and Saudi Arabia, that neither in 2006 (when the WEF Global Gender Gap Report that the authors quote as a reference for the missing values) nor now (in 2021) have guaranteed yet the right to vote for women.
- there are missing data also in the other sources that the authors quote. So a quick search for the missing countries of the WEF report of 2015, shows us that these countries can't be found in the report of 2006 either.

These two unclear points, even though in our understanding not crucial for the replication of the analysis, are not desirable.

Dealing with the missing values What one can do when dealing with missing values is a matter of debate and of taste. As a first approach, we simply excluded them. In a second step, we tried to use some algorithm for the imputation of the missing values.

Remove values with NAs As a first step of the reproduction analysis, the Principal Component Analysis (PCA) has been performed only on the complete data set, leading to a cut of 7 countries from the initial data set for different missing values:

- For the **time since women's suffrage**:
  - United Arab Emirates
  - Saudi Arabia
- For the **WEF Global Gender Gap Index**:
  - Afghanistan
  - Bosnia Herzegovina
  - Haiti
  - Iraq
- For the **UN Gender Inequality Index**:
  - Nigeria

This cut our data set from 76 countries to 69, meaning a 10% less of the initial data set. As 10% is small enough not to have a strong influence on the result, still to estimate how missing values influences the results, we compared different strategies for imputation of the missing values.

Imputing NAs: the `missMDA` library With a small research on the web, we could find several proposed solutions for the imputation of NAs with the scope of performing a PCA in *R*. One of these was to install the `missMDA` package and use the `imputePCA` function on the data we needed to fill.

This function works very nicely: One first selects the columns of the data where there are missing values, then passes it to the function, and then runs the PCA on the sub-list called completeObs generated from this function. The result is a list of four columns that are exactly the input data but with the missing values that have been filled with imputed values. The data must be numeric.

Research Article

Here a quick explanation of the methods used to perform the analysis on the data and some further information about the creation of the plots.

Creation of the Models

Linear Model on Each Country for Each Preference Starting from the complete data set (meaning with removed NA rows), we wanted to reproduce the data plotted in Fig. S2. regarding the gender differences and economic development by preference and by country.

As already mentioned in the previous paragraph, part of the data to reproduce the article is missing and it can be accessed only after payment of a subscription fee to the Gallup World Poll. We decided, therefore, to continue the analysis without using two of the variables used in the model (education level and household income quintile).

We created a linear model for each country using an expression from the article, omitting the 2 missing variables:

```
preference ~ gender + age + age_2 + subj_math_skills
```

This resulted in 6 different models (one for each preference measure), having intercept and 4 weights, each of the weight being related to the variable in the formula above. The weight for the dummy variable "gender" is used as a measure of the country-level gender difference. Therefore, in total, we have 6 weights that represent the preference difference related to the gender for 76 countries.

We plotted the logarithm of the average GDP per capita versus the preference differences, for the 6 different preference measurements. When plotting this, we used a linear model to fit and extract the correlation and the p-value.

Principal Component Analysis To summarize the average gender difference among these preferences, we performed a principal component analysis on the gender preference differences from the linear model and used the first component as a summary index of average gender differences in preferences.

We then performed a linear regression on the data points, extracting the correlation and p-value of the average gender difference in preferences versus the logarithm of the average GDP per capita. The variables on y-axis were additionally transformed as (y-y_min)/(y_max-y_min) (see Fig. 1B).

We performed a PCA also on the four data sets used for Gender Equality, to extract a more general Gender Equality Index based on them. We then used this Gender-Equality Index for plotting the same average gender differences as a function of this index, performing a linear regression to calculate the p-value and the correlation value (see Fig. 1D). Note that here also the Gender Equality Index is transformed to be on a scale between 0 and 1.

Variable Conditioning For the plots in Fig. 2, a conditional analysis was performed. To plot the variables x and y residualised using the variable z:

1. We performed a linear regression of x on z, and then a linear regression of y on z.
2. We calculated the residuals of the variable x, meaning that we take the points on the x-axis and calculate the difference between these and the projection of these points on the model created from the linear regression of x on z; the same for y.
3. We plot the residuals on the corresponding axis.

This has been done for the economic development, for the Gender Equality Index, and for each of the four indicators building the Gender Equality Index. The variable used on the y-axis is, therefore, the first Principal Component of the PCA made on the gender differences on the six preferences.

Additional about the plots

- Any variable called "Std" is standardized.
- Average Gender Difference (Index) is the variable extracted from the first component of the PCA performed on the gender coefficients of the six preferences, for each country.
- Gender Equality Index is the variable extracted from the first component of the PCA performed on the four indicators building this index (WEF Global Gender Gap, UN Gender Inequality Index, Ratio Female to Male LFP, and Time since Women's Suffrage).
- The indicators UN Gender Inequality Index and Time since Women's Suffrage must be inverted to obtain the plots below, because their values suggest an inequality, while the index is measuring the equality.
- The preferences marked with a (–), that are patience, negative reciprocity, and risk-taking, show an inverted trend with respect to the others. To plot them, we inverted manually their values (for instance, in the histograms we multiply the mean value of the resulting quantile by a -1 factor for the above-mentioned preferences).
- The plots showing residuals in the main article are using the log GDP p/c and the Gender Equality Index after standardization, while in the supplementary material they are not standardized.

Variable Conditioning

For the plots in Fig. S5 and S6, we were following the same approach described above in the Main Article section regarding the Variable Conditioning.

Regarding the x-axis, we took the same variables used already in Fig. S2A (for plot S5) and S2B (for plot S6), while for the y-axis we took the gender coefficient for each country selecting for the specific preference, made a linear regression on the same z variable used to residualize the x-axis, and then the residuals.

Preferences Standardized at Global Level

To build Fig. S8, what we have done is simply to standardize the preferences on a global level instead of at the country level. Then, the creation of the models has been done in the same way as in the main article. The plot shows the gender coefficient extracted from the model versus the log GDP per capita, for all the six preferences.

Alternative Model

To build the alternative model without control variables, we started again from the complete data set and created a linear model for each country using simply:

```
preference ~ gender
```

This resulted in 6 different models (one for each preference measure), having intercept and only 1 weight related to the gender. This weight is used as a measure of the country-level gender difference for the alternative model.

We plotted the logarithm of the average GDP per capita versus the preference differences, for the 6 different preference measurements. When plotting this, we used a linear model to fit and extract the correlation and the p-value (see Fig. S9).

Output of the Analysis

The result of the analysis is written into five csv files (two for the main analysis, three for the supplementary material), and these files can be used to reproduce the plots and for comparison to any other analysis based on this method.

The files are:

- **main_data_for_histograms.csv** This file contains the data for reproducing the plot in Fig. 1 (A and C), the distribution of the gender differences within poorer/less gender-equal (corresponding to 1 in the data) and richer/more gender-equal countries (corresponding to 4) among the six preferences.The data consists of 4 variables: preference, GDPquant, GEIquant, meanGenderGDP, meanGenderGEI.
  - *preference* is a character and can be one of the 6 economic preferences (patience, risk-taking, altruism, negative and positive reciprocity, and trust).
  - *GDPquant* is a numeric, from 1 to 4, where 1 represents the lowest quantile, meaning the poorest countries of the dataset, and 4 represents the highest quantile, that is the richest countries from the dataset.
  - *GEIquant* is a numeric, from 1 to 4, representing the equality of the countries in terms of gender opportunities, where 1 is the lowest quantile (the less gender-equal countries), and 4 is the highest (the more gender-equal).
  - *meanGenderGDP* is the average of the gender difference coefficient by preference by GDP quantile
  - *meanGenderGEI* is the average of the gender difference coefficient by preference by Gender Equality Index quantile

- **main_data_aggregatedByCountry_preferencePCA_genderIndexPCA.csv** Data aggregated by country containing the Average Gender Differences (the first component of the PCA made on the six preferences for the gender coefficient), all the indicators of the economic development and gender equality, plus their Standardization, and the residuals that can be used to build Fig. 2. The data consists of 33 variables:
  - *country* is a character
  - *avgGenderDiff* is numeric and it is the result of the PCA on the gender coefficients at country level
  - *isocode* is a character
  - *logAvgGDPpc* is numeric and it is the logarithm of the average GDP of the country in the period from 2003 to 2012, in 2010 US dollars
  - *Date* is integer and corresponds to the time of women's suffrage
  - *ScoreWEF* is numeric and corresponds to a score extracted from the WEF Global Gender Gap Index
  - *avgRatioLabor* is numeric and is the average of the labor of females divided by the labor of males in the country
  - *ValueUN* is numeric and corresponds to a score extracted from the UNDP Gender Inequality Index
  - *region* is a character and corresponds to the region of the world in which the country belongs
  - *telephone* is logic, and it is TRUE if the survey has been performed by telephone in that country
  - *personal* is logic, and it is TRUE if the survey has been performed face-to-face in that country
  - *avgGenderDiffRescaled* is the same measurement as avgGenderDiff, but rescaled using the min-max method.
  - *GenderIndex* is numeric and it is the first component extracted from the PCA performed on the four measurements of Gender Equality Indexes used in FH article (*Date*, *ScoreWEF*, *ValueUN*, and *avgRatioLabor*)
  - *logAvgGDPpcStd* same, but standardized
  - *ScoreWEFStd* same, but standardized
  - *ValueUNStd* same, but standardized
  - *DateStd* same, but standardized
  - *avgRatioLaborStd* same, but standardized
  - *GenderIndexStd* same, but standardized
  - *GenderIndexRescaled* same, but rescaled using a min-max method
  - *residualsavgGenderDiffStd_GDP* is the variable created from the *avgGenderDiffStd* residualised using *logAvgGDPpcStd*
  - *residualslogAvgGDPpcStd_GEI* is the variable created using the *logAvgGDPpcStd* residualised using *GenderIndexStd*
  - *residualsavgGenderDiffStd_GEI* is the variable created from the *avgGenderDiffStd* residualised using *GenderIndexStd*
  - *residualsGenderIndexStd* is the variable created from *GenderIndexStd* residualised using *logAvgGDPpcStd*
  - *residualslogAvgGDPpcStd_WEF* is the variable created using the *logAvgGDPpcStd* residualised using *ScoreWEFStd*

- *residualsavgGenderDiffStd_WEF* is the variable created using the *avgGenderDiffStd* residualised using *ScoreWEFStd*
  - *residualsScoreWEFStd* is the variable created from the *ScoreWEFStd* residualised using *logAvgGDPpcStd*
  - *residualslogAvgGDPpcStd_UN* is the variable created using the *logAvgGDPpcStd* residualised using *ValueUNStd*
  - *residualsavgGenderDiffStd_UN* is the variable created using the *avgGenderDiffStd* residualised using *ValueUNStd*
  - *residualsValueUNStd* is the variable created from the *ValueUNStd* residualised using *logAvgGDPpcStd*
  - *residualsavgRatioLaborStd* is the variable created from the *avgRatioLaborStd* residualised using *logAvgGDPpcStd*
  - *residualsDateStd* is the variable created from the *DateStd* residualised using *logAvgGDPpcStd*
- **supplementary_data_aggregatedByCountry_singlePreference_genderCoefficients.csv** Data aggregated by country but separating each of the six preferences gender difference values.
  - *country*, *isocode*, *logAvgGDPpc*, *GenderIndex* are the same variables as described above
  - *gender* is the coefficient related to the linear regression on the preference
  - *preference* is the same as described above, and in this dataset we kept the six preferences distinguished and not combined into a PCA
  - *residualsgenderGEI_trust*, *residualsgenderGEI_altruism*, *residualsgenderGEI_negrecip*, *residualsgenderGEI_posrecip*, *residualsgenderGEI_risktaking*, *residualsgenderGEI_patience* are built performing a linear regression of the gender coefficient of the specific preference and the Gender Equality Index, and then calculating the residuals from it
  - *residualsgenderGDP_trust*, *residualsgenderGDP_altruism*, *residualsgenderGDP_negrecip*, *residualsgenderGDP_posrecip*, *residualsgenderGDP_risktaking*, *residualsgenderGDP_patience* are built performing a linear regression of the gender coefficient of the specific preference and the log GDP p/c, and then calculating the residuals from it
  - *residualslogAvgGDPpc_trust* is the variable created from *GenderIndex* residualised using *logAvgGDPpc*, and *residualsGenderIndex_trust* is the variable created from *logAvgGDPpc* residualised using *GenderIndex*, selecting the data set on the specific preference "trust". The other ten variables (*residualslogAvgGDPpc_altruism*, *residualslogAvgGDPpc_negrecip*, *residualslogAvgGDPpc_posrecip*, *residualslogAvgGDPpc_risktaking*, *residualslogAvgGDPpc_patience*, and *residualsGenderIndex_altruism*, *residualsGenderIndex_negrecip*, *residualsGenderIndex_posrecip*, *residualsGenderIndex_risktaking*, *residualsGenderIndex_patience*) are created in the same way for each of the corresponding preference.
  - *meanGender* and *stdGender* are the variables indicating the average of the gender differences by preferences, and the 95% confidence interval of the gender differences by preferences
- **supplementary_data_aggregatedByCountry_singlePreference_genderCoefficientsGlobal.csv** Data aggregated by country, separating each of the single preferences and standardize them at a global level.
  - *country*, *isocode*, and *logAvgGDPpc* are the same variables as described above
  - *gender* is the coefficient for the differences between men and women extracted from linear regression on the preference
  - *age*, *age_2*, and *subj_math_skills* are the coefficients (control variables) extracted from the linear regression on the preference

  - *genderOrig* is the original value of the *gender* coefficient, before inverting for those preferences requiring it (negative reciprocity, risk taking, and patience)
  - *preference* is the same as described above, except that here we standardize them at a global level
  - *(Intercept)* is a numeric variable that is not needed for the analysis

- **supplementary_data_aggregatedByCountry_singlePreference_genderCoefficients_alternativeModel.csv** Data created from the alternative model where only the gender is kept as variable to regress the preference, aggregated by country and including the economic development variable.
  - *country*, *isocode*, *preference*, and *logAvgGDPpc* are the same variables as described above
  - *gender* is the coefficient calculated from the linear regression when using the alternative model without control variables
  - *preference* is the same as described above, and in this dataset we kept the six preferences distinguished and not combined into a PCA
  - *(Intercept)* is a numeric variable that is not needed for the analysis
  - *genderOrig* is the original value of the *gender* coefficient, before inverting for those preferences requiring it (negative reciprocity, risk taking, and patience)