# The SISTER Method: Robustness Tests and an Annotated Guide to Implementation Commands for STATA 12

**ONLINE SUPPLEMENT**

**to**

## Capturing Culture: A New Method to Estimate Exogenous Cultural Effects using Migrant Populations

Recommended citation: Polavieja, Javier G. 2015. "The SISTER Method: Robustness Tests and an Annotated Guide to Implementation Commands for STATA 12. Online Supplement to "Capturing Culture: A New Method to Estimate Exogenous Cultural Effects using Migrant Populations." *American Sociological Review*, 80(1): 166-191."

## Index

# I.

# Robustness Tests

## 1. SECTION 1- Are Results Robust to Using Different Imputation Methods and Probability Models?

In this section I test the robustness of my findings to alternative imputation methods. I also test whether results are robust to using linear probability models (LPM) instead of IV-probit models in the second stage. LPMs treat the outcome variable as a linear function of the explanatory variables using standard OLS estimation, as explained in my ASR paper (in what follows referred to as "this study").

There are several imputation methods that are based on multiple regression (see e.g. Longford 2005; McKnight et al. 2007). The standard regression method applied in this study replaces missing values with predicted values that are calculated using multiple regression. Yet one shortcoming of imputation based on standard multiple regression is that it underestimates uncertainty due to sample variance. I have not judged this problem as severe because the SISTER method uses imputed (i.e. synthetic) values only as instruments for observed traits and hence observed variance (at destination) still plays an essential role in the estimation of cultural effects in the second stage (i.e. observed variance affects instrument relevance in the first-stage of the IV-regression). Yet there are more sophisticated imputation methods that account for imputation uncertainty due to sample variance.

Imputation uncertainty can be augmented by adding a stochastic component to the computation of synthetic values. Two different methods are commonly used to this end. The first and simplest method consists of adding an extraction of a normal distribution with mean equal to the predicted value and standard error equal to the root of the mean standard error (RSME) of the imputation regression. This method comes with one caveat, however: because the uncertainty added around imputed values is calculated from a random draw, this method will inevitably yield different IV-estimates for the effect of traditionalism at each draw. Estimates calculated on the basis of one single extraction can be unreliable.

A much more developed method to deal with uncertainty in the imputation of missing values is multiple imputation, MI (see e.g. Longford 2005:62-63; StataCorp. 2011). MI creates *m* sets of imputations for the missing values using an imputation process with a stochastic component. This results in *m* full data sets, each containing somewhat different values due to randomness. Each completed data set is then analyzed (in this case, *m* IV regressions are fitted for each *m* data set) and this yields *m* sets of parameter estimates, which will differ slightly because the data sets are not identical. Finally, results are combined and the variation in parameter estimates is calculated by accounting for both within-*m* variance and between-*m* variance. It must be noted that one problem with MI in the SISTER framework is that it can be computationally intensive and difficult to implement using standard software (implementation details for all three imputation methods are presented below).

Table S1 shows SISTER estimates for the effect of traditionalism on female labor-force participation (FLFP) calculated using both IV-probit and IV-LPM specifications and for the three different imputation methods discussed above (i.e. standard multiple regression, multiple regression with random extraction of RMSE, and multiple imputation). Reassuringly, results hold across the board.

Table S1. The Full Impact of Cultural Traditionalism on FLFP: Testing for the Robustness of SISTER Estimates to Different Specifications and Different Imputation Methods

| *Imputation method* | Standard Multiple Regression | | Multiple Regression with Random RMSE | | Multiple Imputation (*m*=10) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Estimation method* | IV Probit | IV LPM | IV Probit | IV LPM | IV Probit | IV LPM |
| Traditionalism | -0.240*** | -0.083*** | -0.339*** | -0.126** | -0.198** | -0.067** |
| | [0.071] | [0.026] | [0.128] | [0.054] | [0.079] | [0.027] |
| Constant | 0.770** | 0.780*** | 1.188** | 0.958*** | 0.591† | 0.712*** |
| Observations | 2,915 | 2,915 | 2,915 | 2,915 | 29,150 | 29,150 |
| R-squared | | 0.107 | | 0.052 | | 0.120 |
| 1st stage effect T' on T (z): | 20.57**** | 16.88**** | 8.00**** | 7.60**** | 9.41**** | 19.69**** |
| Exogeneity tests: | 7.32*** | 7.34** | 4.23** | 3.61† | 3.48† | 3.39† |

Legend: † p<0.1; ** p<0.05; *** p<0.01; ****p<0.0000

Notes: S.E. Clustered at Destination. All models control for age, age$^2$, years of schooling, generational status, language assimilation and type of destination location. Exogeneity tests are Wald's test for IV-Probit models and Wooldridge's test for IV-LPM models.

Source: Calculated by the author from European Social Survey, Rounds 1, 2 and 3 combined, Restricted Migrants Sample.

## 2. SECTION 2- Are Results Driven by Particular Immigrant Groups?

One source of concern is the possibility that the results of this are driven by immigrants from particular origin groups. This could happen if 1) a few immigrant groups drive the overall relevance of the instrument; or if 2) a few immigrant groups (e.g. women coming from very traditional countries) drive the overall effect of synthetic traits on FLFP. Note that the first source of bias can either stem from genuine differences in assimilation/resilience across origins (which would make some immigrant groups more alike their non-migrating counterparts than others), or be produced by differences in sampling (affecting the ratio of acceptors to donors) that result in differences in the accuracy of imputations.

In order to check these two plausible sources of bias, and more generally to test for the robustness of SISTER estimates to unobserved differences across origin groups, I have performed several tests (see Table S2). These tests include: 1) clustering standard errors at origin and accounting for destination fixed effects; 2) clustering standard errors at each observed combination of destination and origin; 3) introducing a control for the ratio acceptors-to-donors in the structural equations; 4) estimating the structural model separately for respondents coming from synthetic samples with acceptors-to-donor ratios above and below the median; 5) dropping all migrants whose countries of origin are not present in the 3 rounds of the ESS combined dataset; 6) estimating the structural model separately for respondents coming from synthetic samples with high and low gross instrument relevance (measured as the gross correlation between *T* and *T'* at each synthetic sample); 7) removing the 8 poorest origins; 8) removing the 4 richest and the 4 poorest origins; 9) removing the 8 most traditional origins; and, finally, 10) removing the 4 least and the 4 most traditional origins (note that 8 origins amounts to one-third of the origins in the analytical sample).

It must be noted that SISTER estimates are robust to all these tests, even when many of them imply a very drastic reduction of sample size. This strongly suggest 1) that estimates are not an artifact of differences in sample size; 2) that they are not driven by unobserved differences in ethnic resilience by cultural origin; 3) that they are not biased (at least not in any serious way) by any of these two potential problems; 4) that they are not driven by differences in instrument relevance by origin; and 5) that they are not driven by national-origin outliers as defined by traditionalism or GDP. All these tests hold regardless of the model specification and they are also robust to the imputation method (available on request).

Table S2. Testing for the Robustness of SISTER Estimates to Different Clustering, Controlling for Ratio Acceptors to Donors and Different Sample Splits and Restrictions.

| | (1) Country FE & SE clustered by Origin | (2) SE clustered by Origin x Destina-tion | (3) Control for Ratio A/D | (4) Only origins below median Ratio A/D | (5) Only origins above median Ratio A/D | (6) Only origins in all 3 ESS rounds + Ratio A/D | (7) Only origins below median gross instru-ment relevance[a] | (8) Only origins above median gross instru-ment relevance | (9) Remo-ving the 8 poorest national origins[b] | (10) Remo-ving the 4 poorest and the 4 richest national origins[c] | (11) Remo-ving the 8 most tradition-nal origins[d] | (12) Remo-ving the 4 most and the 4 less tradition-nal origins |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Traditionalism. | -0.184*** [0.071] | -0.240*** [0.071] | -0.201*** [0.073] | -0.335*** [0.094] | -0.183** [0.084] | -0.168** [0.077] | -0.322*** [0.095] | -0.158** [0.076] | -0.149† [0.085] | -0.185† [0.095] | -0.168** [0.079] | -0.169** [0.071] |
| Observations | 2,915 | 2,915 | 2,915 | 1,466 | 1,449 | 2,139 | 1,325 | 1,590 | 2,017 | 2,177 | 2,209 | 2,185 |
| 1st stage effect of T' on T(z): | 10.63**** | 14.18**** | 20.09**** | 8.41**** | 15.36**** | 18.07**** | 6.53**** | 23.18**** | 15.37**** | 13.43**** | 13.18**** | 14.95**** |
| Wald test: | 5.34** | 6.77*** | 4.65** | 10.66*** | 1.29 | 4.46** | 8.08*** | 2.10 | 1.0 | 2.27 | 2.07 | 2.86† |

Legend: † p<0.1; ** p<0.05; *** p<0.01; ****p<0.0000

Notes: S.E. clustered at destination unless otherwise indicated. All models control for age, age2, years of schooling, generational status, language assimilation and type of destination location. Constant not shown. [a]Gross instrument relevance measured as the correlation between T and T' as in Table 1. [b]The 8 poorest national origins are those with gross GDP per capita below 26,000 US dollars. [c]The 4 richest national origins have gross GDP per capita above 43,500 US dollars; the 4 poorest national origins have gross GDP per capita bellow 22,500 US dollars (data from the World Bank, 2012: http://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD); [d]Average traditionalism is measured using observed traits (T) as in Table 1; [e] Immigrant groups whose FLFP rate is different from origin in more than 5 absolute percentage points.

Source: Calculated by the author from European Social Survey, Rounds 1, 2 and 3 combined, Restricted Migrants Sample.

## 3. SECTION 3- Are SISTER Estimates Relevant for Non-Migrating Populations?

As other epidemiological methods, the SISTER method uses migration as a key source of identification of exogenous cultural effects. This identification strategy rests on the (mostly testable) assumption that migrant acceptors and non-migrant donors are observational equivalents embedded in different social environments. It is on the basis of this *equivalence* condition, that SISTER estimates can be interpreted as having *general* plausibility, that is, as *empirically* valid for immigrant acceptors and *theoretically* valid for non-migrating donors.

Note that this equivalence condition *could be* violated 1) if immigrants were selected on the trait of interest; 2) if immigrants were selected on the outcome of interest (even if they are not selected on the trait); and/or 3) if immigrants were selected on endogenous characteristics associated with the trait and/or the outcome (e.g. if the family structure of first generation migrants varied between more and less traditional societies so that immigrants were selected on family types affecting FLFP).

In this study I have argued that instrument relevance already deals with the first potential problem, selection on the trait, since if immigrants were particularly selected on the trait of interest, synthetic traits would be poorly correlated with observed traits at destination. Moreover, in Section 2 of this online supplement I have shown that my estimates are robust to removing immigrants from the most and the least traditional origins, as well as to removing immigrant groups below and above median instrument relevance (see models 7, 8, 11 and 12 on Table S2). These robustness checks can be interpreted as tests for selection on the trait bias.

Selection-on-the outcome bias would also constitute a major violation of the equivalence assumption if women migrated for economic reasons to such an extent so as to render the estimated effect of immigrants' traditionalism on FLFP unreliable as a general causal model. In this case, SISTER estimates would provide a biased theoretical extrapolation for non-migrating populations. Yet, as noted above, most immigrants in the ESS data come from, and reside in, rich countries, and this rules out the possibility of extreme forms of economic migration. Furthermore, in the previous section I have shown that results hold after removing the poorest (and the richest) origins from the analytical sample (see models 9 and

10 on Table S2). This already suggests that selection-on-the outcome is not a major source of bias in this study.

Below I perform two further tests for selection bias. The first test consists of removing immigrant groups with large FLFP gaps (with respect to country-of origin FLFP) from the analytical sample. The logic is simple: if SISTER estimates are driven by a few groups with the potential of being selected on FLFP, the theoretical extrapolation of SISTER estimates to non-migrating populations could be at stake. Note that this test makes the very improbable assumption that large FLFP gaps between immigrants and their countries of origin reflect selection on the outcome (rather than differences between the origin and destination environments) and hence constitutes a very conservative test. Reassuringly, I find that SISTER estimates are fully robust to this rigorous test (see models 1 through 4 on Table S3).

The second and final test is the stringiest. It consists of removing all first-generation immigrants from the analytical sample, thus restricting the analysis to second-generation and generation 1.5 respondents only. Obviously, the family and labor-market behavior of women born and/or raised in the country of destination cannot possibly have affected the original migration decisions of their parents. Hence this test should take care of a*ll possible forms of selection-into-migration bias*. Two issues make this test particularly demanding, however: First, removing first generation migrants from the analysis cuts the number of observations in half and this inevitably implies a great loss of statistical power; and second, both 1.5 and second-generation women have longer (and arguably more intense) exposure to the destination environment and this could weaken the instrument relevance of synthetic traits (note that testing whether this is the case is in itself a sociologically relevant question). Both issues should in principle reduce the odds of finding significant effects for this restricted group of respondents given the existing sample sizes.

Model 5 in Table S3 shows, however, that the SISTER estimate for the exogenous impact of traditionalism on FLFP obtained when restricting the sample to 1.5 and second-generation women has roughly the same size as the estimates found for the entire immigrant population, although standard errors are logically larger and the statistical significance of the parameter does not quite reach the 95 per cent level threshold in this case (P|z|=0.090). Most probably the increase in estimation uncertainty provoked by the drastic reduction in sample size imposed by this stringent test also explains why the Wald test fails to reach statistical significance in this restricted sample. Yet it must be noted that the model yields a large and clearly significant z-value for synthetic traditionalism in the 1st stage, the magnitude of which comfortably surpasses conventional thresholds for instrument strength. This means that synthetic values of traditionalism computed using donors from the countries of ancestry also seem to be strong predictors of traditionalism amongst the second and the 1.5 generations of European-origin immigrants. This findings is substantively relevant, for it speaks of cultural persistence, and constitutes a very strong indication that SISTER estimates are not driven by selection bias.

In sum, all the tests summarized in Table S3 (together with many of the tests presented in the previous section) strongly suggest that SISTER estimates are not an artifact of immigrants' selection bias and hence can be interpreted as having *theoretical* validity—i.e. as theoretically plausible for non-migrating populations.

Table S3. Testing for the Robustness of SISTER Estimates: Removing Immigrant Groups with Largest Gaps in FLFP Rates (with respect to Country of Origin) and First Generation Immigrants

| | (1) Removing 2 largest positive FLFP gaps (>+15% difference w.r.t. origin country)[a] | (2) Removing 6 largest positive FLFP gaps (>+6% difference w.r.t. origin country) [b] | (3) Removing 5 largest negative FLFP gaps (>-6% difference w.r.t. origin country) [c] | (4) Removing 11 largest absolute gaps in FLFP (>\|6%\| difference w.r.t. origin country) [d] | (5) Removing all first generation immigrants |
|---|---|---|---|---|---|
| Traditionalism | -0.183** [0.077] | -0.156** [0.070] | -0.313*** [0.078] | -0.214*** [0.072] | -0.217† [0.128] |
| Observations | 2,652 | 2,399 | 2,345 | 2,057 | 1,432 |
| 1st stage effect T' on T(z): | 16.12**** | 16.03**** | 17.31**** | 15.57**** | 12.66**** |
| Wald test: | 3.89** | 3.33† | 10.48*** | 6.45** | 1.82 |

Legend: † p<0.1; ** p<0.05; *** p<0.01; ****p<0.0000

Notes: S.E. clustered at destination. All models control for age, age2, years of schooling, generational status, language assimilation and type of destination location. Constant not shown.

[a] Immigrants from Spain and Turkey; [b] Immigrants from Spain, Turkey, the Netherlands, Ireland, Greece and Ukraine; [c] Immigrants from Britain, Denmark, France, Slovenia and Switzerland; [d] All previous origins removed.

Source: Calculated by the author from European Social Survey, Rounds 1, 2 and 3 combined, Restricted Migrants Sample.

# II.

# The SISTER Method: An Annotated Guide to Implementation Commands for STATA 12

# 1. Introduction

These lines present an annotated guide to implement the SISTER method using Stata 12 (see StataCorp. 2011a). Annotations are brief and assume previous reading of my ASR paper "Capturing Culture: A New Method to Estimate Exogenous Cultural Effects using Migrant Populations". I also assume readers are familiarized with basic data manipulation and data management commands in Stata as, for the sake of brevity, I skip some of these commands in the syntax do-files.

In my ASR paper I apply the SISTER method to estimate the impact of one particular cultural trait, traditionalism, on one particular outcome, women's labor force participation (FLFP), using one particular comparative survey, the European Social Survey. My annotated syntax commands reflect this particular research question and are prepared for a comparative survey such as the ESS.

Note, finally, that I present three different imputation methods in Step I and two different estimation methods in Step II. The main results of my ASR paper use the first of these methods in each step (i.e. imputation by multiple regression in Step I and IV-probit estimation in Step II) and use the other three methods for robustness only.

# 2. Labels for key variables

tradition→ Stands for immigrants' actual values of traditionalism (observed trait, $T_d$ ). This variable must be set to missing in the imputation step.

IVtrad(culture)(_number)→ Is the imputed (or synthetic) values of traditionalism for a given synthetic sample that are generated by means of imputation; (culture) indicates the imputation group or synthetic sample defined by national origin; and (_number) reflects the imputation method used (1=multiple regression; 2=multiple regression with stochastic component). Hence, for example, the variable IVtradAT_1 captures the predicted values of traditionalism for Austrian-origin immigrants residing in any country of the ESS but Austria. These predicted values have been generated using Austrian non-migrating women as donors by fitting a multiple regression (method 1) to the Austrian synthetic sample, after having set immigrants' values of traditionalism (tradition) temporarily to missing.

IVtrad(_number)→ This variable combines all the imputed (or synthetic) values generated for each immigrant group into one single variable, which is subsequently used in the final estimation step (step II). Again (_number) describes the imputation method used. Hence IVtrad_1 contains all the synthetic values imputed using method 1 (i.e. multiple regression) for all immigrants from the 23 different national origins present in the ESS dataset. The main findings reported in my ASR paper are obtained using this variable (IVtrad_1) , which is noted *T'* in the paper, as an instrument for immigrant's observed values of traditionalism (tradition), as I explain in detail below.

culture→ This is a dummy variable containing the 23 different synthetic samples that can be identified in the ESS dataset (see below).

INHERTITED_traditional→ This variable captures the strength of the intergenerational transmission of traditionalism at the country of origin. This parameter is included as a means to better guarantee that the exclusion restriction is met and it is noted $\bar{T}_{o_{(g-1)}}$ in my ASR paper (the computation of this variable is explained in section 6 below).

## 3. Data preparation: Synthetic samples

In this section I explain the computation of synthetic samples. For the sake of brevity, very basic data management steps are skipped.

The SISTER method generates synthetic traits by treating all values of the observed trait (immigrants' traditionalism) as temporarily missing and then using relevant information from non-migrating equivalents to impute these values back. This requires that migrants and their non-migrating counterparts are temporarily grouped together in what I call a *synthetic sample*, where non-migrants act as imputation donors for migrants.

In a comparative survey, such as the ESS, synthetic samples are computed very straightforwardly simply by generating a variable that groups each immigrant women (aged 16 to 65) with her country-of-origin sample (i.e. non-migrating natives at country of origin aged 16 to 65) by assigning them the same value. I call this variable culture.

In order to generate synthetic groups it is thus essential that the survey includes information on: 1) migration status (i.e. whether the respondent is native, born abroad, or born from foreign-born parents); 2) country of birth of foreign-born respondents; and 3) country of birth of parents (for second-generation respondents). With such information synthetic samples can be computed using the most basic data management commands.

```
/*
NOTES ON SYNTHETIC SAMPLES

If migrant is from country X, she is grouped with all native respondents from country X that will act as imputation donors (i.e. native women aged 16 to 65 with non-missing values of traditionalism sampled at country X) by assigning both immigrant acceptors and non-migrating donors the same value in a variable comprising all national cultures in Europe. I have called this variable “culture”. Each value of culture thus defines a synthetic sample.

[generation commands skipped]

This is how  “culture” variable looks like :

cultural
groups in
Europe                  Freq.   Percent  Cum.

Spanish                 1,955   4.47     4.47
Portuguese              2,429   5.56     10.03
Italian                 859     1.97     12.00
Irish                   1,890   4.32     16.32
French                  1,925   4.40     20.72
Austrian                2,642   6.04     26.77
Polish                  2,446   5.60     32.37
Slovenian               1,502   3.44     35.80
Slovakian               1,319   3.02     38.82
Hungarian               1,950   4.46     43.28
Czhec                   1,541   3.53     46.81
Greek                   1,737   3.97     50.78
Ukranian                1,377   3.15     53.93
Belgian                 1,852   4.24     58.17
Swiss                   1,685   3.86     62.03
German                  3,469   7.94     69.96
British                 2,308   5.28     75.24
Dutch                   2,374   5.43     80.68
Swedish                 1,710   3.91     84.59
Norwegian               1,805   4.13     88.72
Finish                  2,207   5.05     93.77
Danish                  1,665   3.81     97.58
Turkish                 1,059   2.42     100.00

Total                   43,706  100.00
*/
```

## 4. Step I: The imputation regression

The imputation regression used in my ASR paper includes the following imputation predictors: age, schooling (yearsedu), parental education (parent_edu3), religious denomination (REL_denom) and the Proportion of traditional women in previous generation (INHERITED_traditional).

There are several imputation methods that are based on multiple regression. In Section 1 of this online supplement I have discussed three such methods (i.e. standard multivariate regression, multivariate regression with random extraction of RMSE, and multiple imputation). Next I present annotated syntax commands for each of them.

### 4.1. Method 1: Imputation by Multiple Regression

The standard regression method applied in my ASR study (method 1) replaces missing values with predicted values that are calculated using a multiple regression model. Annotated commands recovered from my research do-files are presented below:

```
/*
IMPUTATION METHOD 1: MULTIPLE REGRESSION
[Use ESS combined dataset (restricted to women 16 to 65 with valid values of traditionalism)]

First, I compute a variable for traditionalism (XTRAD) that sets all values for immigrants to missing:
*/
gen XTRAD=tradition
replace XTRAD=. if immigrant==1
/*
Next  I generate 23 different variables, one for each synthetic sample:
[I use vehicle registration plates as labels to rename each national culture of origin]

xi: sum i.culture
rename _Iculture_2 PT
rename _Iculture_3 IT
rename _Iculture_4 IE
rename _Iculture_5 FR
***and so on for all 23 cultures of origin in the dataset…
/*
[Note that "xi: sum i.culture" will not generate a dummy for the first value of the culture variable
(culture==1) for STATA takes this value as the reference category.  Hence we need to generate this
value directly ]
*/
gen ES=0
replace ES=1 if culture==1
ta ES
/*
Now I have generated 23 dummies (named as in vehicle plates). In each of them value 1 defines the
synthetic sample.
```

Next I fit one imputation regression for each of these 23 synthetic samples and save the predicted values in a new variable called zhatt(culture) using the following 2 loops:

```
*/
foreach var of varlist PT-ES {
            xi: quietly regress XTRAD yearsedu age i.parent_edu3 i.REL_deno INHERITED_traditional if
`var'==1
                                    predict zhatt`var'_1, xb
            }
foreach var of varlist PT-ES {
gen IVtrad`var'_1=cond(missing(XTRAD), zhatt`var'_1, XTRAD) if `var'==1
}
/*
```

[Note: to display the outcome of each imputation regression remove “quietly” from the regression command above]

/*

Note that each imputation regression is fitted to one synthetic sample only and that, because XTRAD has been previously set to missing for immigrants, each regression uses only information from non-migrating donors at country of origin. These two loops thus yield 23 different zhatt variables, one for each culture of origin, each of which contains the predicted values of traditionalism imputed for immigrants using non-migrating natives from their country of origin as imputation donors.

Finally, I merge all these 23 different variables into one single variable (IVtrad_1) using the following loop

```
*/
gen IVtrad_1=. if XTRAD==.
foreach var of varlist PT-ES {
replace IVtrad_1=IVtrad`var'_1 if `var'==1
}
/*
```

The resulting variable, IVtrad_1 , thus comprises all the predicted values obtained from fitting 23 different imputation regressions (one for each synthetic sample). This variable can now be used in step II as an exogenous instrument for immigrants’ actual values of traditionalism.

*/

## 4.2. Method 2: Adding a stochastic component (robustness)

As explained above (see Section 1), imputation uncertainty can be easily augmented by adding an extraction of a normal distribution with mean equal to the predicted value and standard error equal to the root of the mean standard error (RSME) of the imputation regression. Below I present annotated commands retrieved from my research do-files that explain how to implement this imputation method (method 2) in Stata 12:

```
/*IMPUTATION METHOD 2: ADDING A STOCHASTIC COMPONENT TO SYNTHETIC TRAITS

Recall that I fitted one imputation regression for each of the 23 synthetic samples and saved the
predicted values in 23 new variables called zhatt(culture)_1 . The imputation regression I used was:

foreach var of varlist PT-ES {
		xi: quietly regress XTRAD yearsedu age i.parent_edu3 i.REL_deno INHERITED_traditional if
`var'==1
					predict zhatt`var'_1, xb
		}

Now for each of these 23 sets of imputed values, I add an extraction of a normal distribution, where
the predicted value is the mean for each synthetic sample and the standard error is the squared root of
the mean standard error (RMSE)

*/
foreach var of varlist zhattPT_1-zhattES_1 {
set seed 654854
gen rand`var'=cond(missing(XTRAD), rnormal(`var', e(rmse)), XTRAD)
}
/*
[Note that for replication purposes, I have identified the random draws using Stata's "set seed"
command followed by an ID number of my own choosing]

Now I have 23 new different variables called randzhatt(culture)_1,  each of which contains imputed
values calculated by adding  a stochastic component to the predicted values of each of the 23
imputation regression fitted to each synthetic sample.

Finally, I merge all these 23 different variables (one from each synthetic sample) into one single
variable called IVtrad_2.

*/
gen IVtrad_2=. if XTRAD==.
foreach var of varlist PT-ES {
replace IVtrad_2=randzhatt`var'_1 if `var'==1
}
/*

The resulting variable, IVtrad_2, contains stochastically-enhanced predicted values for all immigrant
women in the dataset. It has been used as a robustness test (see Section 1 of this online supplement).
*/
```

## 4.3. Method 3: Multiple Imputation (robustness)

As explained above (see Section1), a much more developed method to deal with uncertainty in the imputation of missing values is multiple imputation, MI (see e.g. Longford 2005:62-63; StataCorp. 2011a). It must be noted, however, that one problem of MI in the context of the SISTER method is that it is a very computationally-intensive method (i.e. it might take literally hours for a standard PC to generate the desired number of imputed datasets for each synthetic sample). Also MI might be difficult to implement using other standard statistical software. Below I present syntax commands for Stata 12:

```
/*IMPUTATION METHOD 3: MULTIPLE IMPUTATION, MI

[Use ESS combined dataset (restricted to women 16 to 65 with valid values of traditionalism)]

MI explained step by step:
1.
First, I compute a variable for traditionalism (IVTRAD_3) that sets all values for immigrants to missing
(as I did for the previous two imputation methods):
*/
gen IVtrad_3=tradition
replace IVtrad_3=. if migrant>0 & migrant<.
/*
2.
Stata requires that the data is set and identified as MI data before they can be used with the other mi
commands.  This implies registering the dataset as suitable for MI and identifying the variables as
imputable or regular
*/
mi set flong
mi register imputed IVtrad_3
mi register regular yearsedu age culture parent_edu3 REL_deno INHERITED_traditional tradition
/*
3.
Using MI to impute "m" imputation sets for 23 different synthetic samples becomes almost
computationally unmanageable. Yet, in a comparative survey such as the ESS, this procedure can be
greatly facilitated by using a single interacted imputation regression (rather than 23) as a
computational shortcut (see below).

[Note that for replication purposes, I also identify the draw via a seed ID]
*/
set more off
set seed 654854
set mat 1000
 /*

For computational efficiency,  I rewrite the 23 different imputation regressions as one single interacted
model with (23-1=)22 interacted parameters:
*/
xi: mi impute regress IVtrad_3 i.culture*yearsedu i.culture*age i.culture*i.parent_edu3  ///
i.culture*i.REL_deno i.culture*INHERITED_traditional, force add(10)
/*

[Command "add(n)" specifies the number of imputations to add to mi data. I note that generating 10
imputed datasets already takes quite a long time]

IVtrad_3 now contains imputed values for traditionalism for each immigrant women in the dataset that
have been predicted using information from non-migrating donors. These imputed values have been
computed by generating 10 different imputation sets with a stochastic component and then combining
the predictions of these sets accounting for both within-m variance and between-m variance. I have
used IVtrad_3 as an alternative synthetic trait in Step II for robustness (see Section 1 in this online
supplement).
*/
```

## 5. Step II: IV estimation

Synthetic traits can now be used as instruments for immigrants' observed traits using standard IV-estimation techniques. The final analysis is logically restricted to immigrant women only. Because the dependent variable of interest is in this case a binary outcome variable capturing women's labor-force participation, in my ASR paper I use IV-probit (estimation method 1). Yet I also check for the robustness of my findings to the use of Linear Probability Models, LPMs (estimation method 2).

Before displaying syntax commands for each estimation method, a few notes on labelling of new variables seem in order:

Variables (not described earlier): active is the dependent variable (0=out of the labor force but not in full time education; 1= in the labor force); agec (age centered); agec2 (age centered squared); G_s (immigration status: 1= first generation; 1.5= generation 1.5; 2= second generation); lang_assim3 (whether respondents speak the language of the host country at their homes); city (whether respondents live in: 1=big city; 2=a town or small city; or 3= in the countryside)

### 5.1. Method 1: IV-Probit estimation

Below I present annotated commands retrieved from my research do-files:

```
/*ESTIMATION METHOD 1: IV-PROBIT
[Run the following syntax command using the analytical sample only (immigrant women aged 16 to
65)]

SISTER estimates can now be calculated using synthetic traits (IVtrad_1) as instruments for immigrants'
observed traits (tradition) in IV-probit estimation, as follows:
*/
xi: ivprobit active (tradition =IVtrad_1) agec agec2 yearsedu i.G_s i.lang_assim3  ///
  i.city, first vce(cluster country)
/*
By default, ivprobit in Stata 12 uses MLE and reports the Wald test of exogeneity but it does not show
the relationship between the endogenous variables and the instruments. By adding "first" at the end of
the regression command I requests Stata to display the parameters for the reduced-form equations
showing this relationship. This allows me to check for instrument relevance.

[Note that in this particular specification standard errors are clustered at destination country using the
"vce" command. Other possible specifications could include clustering standard errors at country of
origin and accounting for destination fixed effects, or clustering standard errors at each observed
combination of destination and origin countries (see Section 1of this online supplement)]

For more information on IVprobit estimation in Stata 12 the reader should consult StataCorp. (2011b).
*/
```

## 5.2. Method 2: Linear Probability Model estimation (robustness)

Albeit mostly discredited in sociology and political science, the use of linear probability models LPM for binary outcome variables is widespread in applied economics (see e.g. Angrist and Pischke 2008; Friedman 2012). For robustness, I calculate SISTER estimates using two-stage least squares (2SLS) linear probability models, where the first stage is identical as in IVprobit models but in the second stage the DV is treated directly as a linear function of the explanatory variables using standard OLS regression estimation. The syntax for LPMs is thus identical to the syntax one should use if the outcome variable of interest was (approximately) continuous.

Below I present annotated commands retrieved from my do-files:

```
/*
ESTIMATION METHOD 2: LPM USING TWO-STAGE LEAST SQUARES (2SLS) ESTIMATION (reporting first-stage results)

[Run the following syntax commands using the analytical sample only (immigrant women aged 16 to 65)]

*/
xi: ivregress 2sls active4 (tradition =IVtrad_1) agec agec2 yearsedu i.G_s i.lang_assim3  ///
 i.city,  first vce(cluster country)
/*

In this case we need to ask Stata to perform Wald's exogeneity test explicitly by typing:

*/
estat endog
/*

[Again standard errors are clustered at destination country using the "vce" command. Other specifications are possible, including clustering standard errors at country of origin and accounting for destination fixed effects, or clustering standard errors at each observed combination of destination and origin countries (see robustness tests in Section 1 of this online supplement)]
*/
```

## 6. Addendum: Generating parameter $\bar{\boldsymbol{T}}_{\boldsymbol{o}_{(\boldsymbol{g}-\mathbf{1})}}$

The imputation regression for synthetic traditionalism includes parameter $\bar{T}_{o_{(g-1)}}$, which captures the proportion of traditional women in the preceding generation for each of the countries of origin. This parameter is used as an estimate of the strength of the intergenerational transmission of traditionalism at the country of origin and has been included in the imputation as a further guarantee that the exclusion restriction is met. Yet it is important to note that I do not judge the inclusion of this particular parameter as an essential condition for the SISTER method. In my view, Step I can produce valid instruments as long as the following two conditions are met: 1) that the imputation predictors that clearly violate the exclusion restriction (e.g. age and schooling) are also included as a predictors in Step II; and 2) that the imputation regression includes at least one parameter that is arguably implicated in cultural transmission and hence orthogonal to the error term in the second stage of the IV estimation (e.g. parental education).

$\bar{T}_{o_{(g-1)}}$, has been computed by considering three large cohorts of roughly similar sample size: those born before 1946 (g=1); those born between 1945 and 1969 (g=2) and those born after 1969 (g=3) (note that larger sample sizes would have allowed for a larger number of cohorts capturing shorter time-intervals). I have chosen 1945 and 1969 as the cutoff points, for such dates mark two crucial historical events, the end of World War II and the demise of the so-called Spring Revolutions in France and Czechoslovakia. In any event, robustness tests show that considering different cutoff points in the definition of cohorts does not alter my results (available on request).

Below I present annotated commands recovered from my research do-files that explain how this parameter has been generated:

```
/*
GENERATING A VARIABLE THAT MEASURES THE P OF TRADITIONAL WOMEN IN PRECEEDING
GENERATIONS AT COUNTRY OF ORIGIN

 [Use full combined sample]

3 cohorts are considered
/*
gen cohort=3 if yearbirth<1946
replace cohort=2 if yearbirth>1945 & yearbirth<1970
replace cohort=1 if yearbirth>1969 & yearbirth<.
/*
```

```
I generate a dummy variable for traditional women (traditional).  Women with values 4 and 5 in the
traditionalism scale are considered “traditional” (1), else is non-traditional (0)
/*
gen traditional=tradition
recode traditional . 1 2 3=0 4 5=1
/*
Now I generate a variable for the P of traditional women in each cohort considered. This variable is
restricted only to non-migrating native women (migrant==0)
*/
gen nattraditionalc1=traditional if migrant==0 & cohort==1
gen nattraditionalc2=traditional if migrant==0 & cohort==2
gen nattraditionalc3=traditional if migrant==0 & cohort==3
/*
Next I calculate the proportion of traditional women amongst non-migrating natives in each cohort and
for each culture of origin (culture)
*/
egen traditional_c1= mean(nattraditionalc1), by(culture)
egen traditional_c2= mean(nattraditionalc2), by(culture)
egen traditional_c3= mean(nattraditionalc3), by(culture)
/*
Finally, I compute the P of traditional women in the preceding generation for each of the countries of
origin in the following two steps:

First, for generations 1 (born after 1969) and 2 (born btw 1945-1969), I use the previous generation
(cohorts 2 and 3, respectively)
*/
gen INHERITED_traditional=.
replace INHERITED_traditional=traditional_c2 if cohort==1
replace INHERITED_traditional=traditional_c3 if cohort==2
/*
Second, for individuals born before 1946, I consider only respondents under 66, and use respondents
over 65 as the previous generation (for this I have to compute a new cohort comprising  respondents
older than 65 and a new variable capturing the P of traditional women in this older cohort)
*/
gen cohort2=4 if age>65
gen nattraditionalc4=traditional if migrant==0 & cohort2==4
egen traditional_c4= mean(nattraditionalc4), by(culture)
replace INHERITED_traditional=traditional_c4 if cohort==3 & age<66
/*
The resulting variable, INHERITED_traditional, now captures the P of traditional women in the
preceding generation using information for non-migrating natives only.
*/
```

# III. References

Angrist, Joshua David, and Jörn-Steffen Pischke. 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press.

Friedman, Jed. 2012. "Whether to Probit or to Probe It: In Defense of the Linear Probability Model." Development Impact (blog). From blogs.worldbank.org. http://blogs.worldbank.org/impactevaluations/whether-to-probit-or-to-probe-it-in-defense-of-the-linear-probability-model.

Longford, Nicholas T. 2005. *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York: Springer.

McKnight, Patrick E., Katherine M. McKnight, Souraya Sidani, and Aurelio J. Figueredo. 2007. *Missing Data: A Gentle Introduction.* New York: Guilford Press.StataCorp. 2011a. *Stata Statistical Software: Release 12.* College Station, TX: StataCorp LP.

StataCorp. 2011a. *Stata Multiple Imputation Reference Manual. Release 12*. College Station, TX: StataCorp LP.

—. 2011b. *Stata Multiple Imputation Reference Manual*. Release 12. College Station, TX: StataCorp LP.