Overview

The codes in this replication package 1) construct the dyadic analysis file from confidential individual student survey data, 2) construct a robustness IV based on the alphabetical order of student names in additional cohorts besides that used for the main analysis, 3) run the analysis and produce the results presented in the published version of the paper. The codes are Stata codes and meant to be run on Stata 17.0 or later. Two main files prepare the data for the analysis, one file create the alphabetical order robustness IV, two files define new functions used to compute cluster corrected standard errors and perform permutation tests, and one file produces the results.

Data Availability and Provenance Statements

The data used in the paper come from different sources 1) two surveys that the authors conducted to gather information from students at Sciences Po; 2) student administrative data obtained from the administration of Sciences Po, which were used to complement the information collected with the survey.

The survey's organization and questionnaire are described in the main paper. The survey's data are anonymized to omit all information on students' names or any names of associations. The anonymized data are built into a dyadic dataset that contain all pairs of individuals retained in the main sample. The selection of this main sample is detailed in Appendix Table A2 of the paper.

The administrative data are obtained from the administration of Sciences Po. They contain information on individual students' attributes. Those attributes are described in Appendix Table A1 of the paper. Because of the protection of students' privacy, **this dataset is not publicly shared**.

Based on the administrative data, the authors have constructed a set of dyadic variables that pertain to all pairs of individuals retained in the main sample. The selection of this main sample is detailed in Appendix Table A2 of the paper. The dyadic variables are described in Appendix Table A1 of the paper. The dyadic data are further anonymized, so that there is no information on students' names or any names of associations, groups, nationality. After anonymization, **it is not possible to identify individual students from the dyadic data. This dyadic dataset can thus be publicly shared**.

The publicly shared dyadic dataset thus contains both the survey and the administrative data. All names are anonymized, so that individuals, associations, and nationality cannot be identified from the data.

In the package, the raw data by individual students are stored in \Data\Raw\. Those data are not to be shared publicly, and only serve as basis for replication. The intermediate data created from the raw data are stored in \Data\Intermediate\. Those data are not to be

shared publicly, and only serve as basis for replication. Last, the dyadic data for most of the analysis are stored in \Data\Analysis\. Those data are fully anonymized to the extent that individuals cannot be identified, and they will be shared publicly.

Statement about Rights

- We certify that the authors of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Summary of Availability

- Only **some data files** in the subfolder \Data\Analysis will be made publicly available.

Details on each Data File

1. **Data in the subfolder \Data\Analysis**
   a. **Publicly shared data**

Work_data_public.dta

This is the main dataset that contains the dyadic data use in the main part of the analysis. It is constructed from the administrative data and the survey data contained in the subfolder \Data\Raw.

It is built for all pairs of individuals retained in the main sample. The selection of this main sample is detailed in the main paper and its Appendix Table A2. The dyadic variables are described in Appendix Table A1 of the paper. The dyadic data are anonymized, so that there is no information on students' names or any names of associations, groups, nationality. After anonymization, it is not possible to identify individual students from the dyadic data.

Recall_data.dta

This file contains data that serves to analyze the recall bias, as shown in Appendix Tables A11 and A12.

   b. **Confidential data**

Work_data_confidential.dta

All of the analyses are designed to run on this file. It is not publicly available, because it is built upon confidential data and contains confidential information about students' identity that cannot be revealed to the public.

2. **Intermediate data in the subfolder \Data\Intermediate**

Those datasets are provided for replication purposes only. They cannot be shared publicly, because students' identity can be recovered from those datasets.

Map.dta

This file contains information to map students' real ID to their generated ID.

Map_asso.dta

This file contains information to map associations' name to their generated code.

Name_ranking_full.dta

This file contains information to map students' family name to their alphabetical rank, as detailed in the paper.

Name_ranking.dta

This file contains information to map students' real ID to their alphabetical rank across several cohorts, as detailed in the paper.

Name_ranking_internal.dta

This file contains information to map students' real ID to their alphabetical rank within their own cohort, as detailed in the paper.

3. **Raw data in the subfolder \Data\Raw**

Those datasets are provided for replication purposes only. They cannot be shared publicly, because students' identity can be recovered from those datasets.

base_complete_1314_i1.dta, **base_complete_1314_i1.dta, base_complete_1314.dta**

Those files contain the main dataset that is combined from both administrative data and survey data. They have been anonymized, but may still contain enough identifying information. They will not be shared publicly.

name_09.dta, name_1314.dta, name_end_131415.dta, name_17.dta,

Those files contain the list of student names, and the mapping to student IDs in the cohorts studied in this paper. They will not be shared publicly.

The codebooks of those data files are included in the accompanying log files.

Computational requirements

There is no specific computational requirement.

Software Requirements

Stata (code was run with version 18.0). The following Stata commands are required, and will be installed by running Setup.do in the subfolder Codes/Create.

- ranktest
- esttab

- estpost
- ivreg2
- outreg2

Running Codes/Create/Setup.do also defines the path to the replication folder. Currently, this line is written as:

global path = "/Users/quocanhdo/Monash Uni Enterprise Dropbox/Quoc-Anh Do/Project - Network at Sciences Po/QA - Nicolo/Replication_Package/ Replication_Package_Mar2025"

This path is required for the main analysis program Codes/Analysis/Analysis.do, and is redefined in this file. They should be modified to reflect the correct path.

The global variable "confidentiality" defines whether the code runs with the confidential data (by setting confidentiality = "confidential") or with the part of the data that can be publicly accessible (by setting confidentiality = "public").

This main analysis program Codes/Analysis/Analysis.do requires two additional programs: IVdyadic.ado and Permutation.ado. They need to be placed in the same folder Codes/Analysis/Analysis. They will be called when Analysis.do is run.

The code in Analysis.do assures that IVdyadic.ado is called once and only once. Sometimes rerunning this ado file with the command below can create an error. Normally it suffices to place this ado file in the same folder Codes/Analysis/. Calling it from Analysis.do the first time just makes sure that it is properly loaded in memory.

Controlled Randomness

Controlled randomness is used to perform permutation tests. In each command, the randomness seed is reset separately. They are all included in the program file "Analysis.do".

Memory and Runtime Requirements

Summary

Approximate time needed to reproduce the analyses on a standard 2025 desktop machine: 3 hours

Details

The code was last run on a **10-core Intel-based laptop with Windows 11 Enterprise version 10.0.26100 with 32 GB of RAM**.

Description of programs/code

A. In the subfolder Codes/Create

1. The file Individual_data.do contains the code that cleans individual data. This file needs to be run on the raw data.

2. The file Create_work_data.do contains the code that creates the dyadic dataset. This file needs to be run on the raw data.

- The file Create_data_permutation_test.do serves to create the data for permutation tests.

- To recreate the main dataset, first run Individual_data.do, then run Create_work_data.do. The file Create_data_permutation_test.do is called upon internally.

B. In the subfolder Codes/Analysis

- The file Analysis.do runs all the empirical analysis shown in the paper. It needs to be run on the dyadic dataset.

- The file Permutation.ado serves to run the permutation tests. It is run once in Analysis.do, and it loads the programs used in permutation tests.

- The file IVdyadic.ado calculates the proper standard errors in dyadic regressions. It is run once by the analysis in Analysis.do by the command *do "$path/Codes/Analysis/IVdyadic.ado"*. The program IVdyadic will later be called repeatedly in Analysis.do.

C. In the subfolder Results

- This subfolder contains output files from the codes

- The tables and figures in the main text are stored in files with corresponding names

- The tables and figures in the appendix are stored in files according to the comments in the analysis' codes.

- For all Tables 1-12, A1-A31 and all Figures 1-3, A1-A4: The corresponding codes are included in the file Analysis.do (with comments that describe each panel for each table)

D. Organizations of output files – How to find the paper's numbers:

1. In permutation tests: The numbers are shown in the log files (*.smcl)
    - The actual value is shown under the column T(obs)
    - The correct p-value is shown under the column "p", row "lower"

    Below is an example: The actual value is 2.788 and the p-value is .897.

```
Monte Carlo permutation results                 Number of observations = 323
Permutation variable: IGno_i1                   Number of permutations = 300

      Command: W_B_Cat Codepostal_adressePR_i1 IGno_i1
Within_Betw~s: r(W_B_Cat)
```

						Monte Carlo error		
T	T(obs)	Test	c	n	p	SE(p)	[95%	CI(p)]
Within_Bet~s	2.788416	lower	269	300	.8967	.0176	.8565	.9287
		upper	31	300	.1033	.0176	.0713	.1435
		two-sided			.2067	.0234	.1608	.2525

```
Notes: For lower one-sided test, c = #{T <= T(obs)} and p = p_lower = c/n.
       For upper one-sided test, c = #{T >= T(obs)} and p = p_upper = c/n.
       For two-sided test, p = 2*min(p_lower, p_upper); SE and CI approximate.
```

2. Main estimates:
    i. The main estimates are just the first coefficient from ivreg2 outputs. They are shown as coefficients in the .tex output files.
    ii. Some .tex files also contain first-stage regressions. In this case, the first-stage estimate is shown at the row that corresponds to the variable name of the instrumental variable. For example, in Table8.tex the last column represents the first-stage regression, and its main coefficient is shown on the row "IG".
3. Standard errors: Most standard errors are estimated separately with the procedure IVdyadic. The standard errors are then written in the .tex output file on separate rows indicated by StdErr IG for estimates of the effect of IG, StdErr LUL for estimates of the effect of friendship LUL, and StdErr IV for estimates of IVs.
4. Other statistics shown in .tex output files include R squared, Kleibergen-Paap F-statistics, Mean(Dependent variable), Standard Error(Dependent varable), and the number of observations.
    i. The Kleibergen-Paap F-stat is calculated by the IVdyadic program from the first stage regression to take into account the specific dyadic standard error correction. They are shown on the columns of the corresponding first stage regressions.
    ii. The correct p-values are available in the log files, from the outputs of IVdyadic.
5. All statistics are stored in log files (*.smcl). Here is the structure of the outputs from IVdyadic commands:
    i. A display of the Left Hand Side variable and the Right Hand Side variables, including the covariates.

ii. The clusters: In this paper, they are always IGno_i1 and IGno_i2

iii. The number of observations

iv. The number of variables on the Right Hand Side, plus the constant.

v. "Diagonal of V": The diagonal of the Variance-Covariance matrix of the estimators Betas (in the order of the RHS variables).

    1. We do not show the estimates of the coefficients Betas, which are the same as the estimates from ivreg2.

vi. eValues: the eigenvalues of the Variance-Covariance matrix of the Betas. They help determine in case we need to adjust this VCV matrix to avoid the standard errors of the Betas being negative.

vii. Adjustment factor: the factor we use to adjust the VCV matrix so that the smallest standard error is just nonnegative. Throughout the paper, this factor is mostly 1, meaning no adjustment is needed. If it is different from 1 (smaller than 1), it means that we need to adjust upwards the estimated standard errors to make sure none of them are negative.

viii. Standard errors: Those are the standard errors of the Betas, as calculated from the Diagonal of V. They follow the same order of the RHS variables.

ix. p-values: Those p-values correspond to the t-test of each coefficient being zero. They follow the same order of the RHS variables.

x. F-stat for all: The F-stat of the hypothesis that all coefficients of the RHS variables are zero, taking into account clustering.

xi. F-stat for first coefficient: The F-stat of the hypothesis that the first coefficient of the RHS is zero, taking into account clustering. In the first stage regression, this corresponds to the Kleibergen-Paap cluster-adjusted F-stat that we report in the tables.