Assessment and improvement of initial reproduction package

Among the many possible replication options we have chosen to focus on eassessment and improving the replication package provided by the author.

I. Assessment

In this evaluation step we analyse the current structure of the reproduction package. This includes describing the data sources and raw data, analysis data, code scripts and outputs.

.

I.1 Description of data sources and raw data

The study of the structure of the raw data is structured around five main points.

First, we look for the existence of raw data sources and present how they were written. Then we indicate the pages where they are cited in the body of the article or in the appendix. The third point is to mention the raw data files provided by the author. The location in the reproduction package of the existing files is indicated in a fourth point. The fifth point is to indicate whether the data source is included in the original reproduction package. Finally a sixth point mentions whether the existing data sources have been explicitly mentioned in the article.

All this information is recorded in a summary table.

> Table 1: Information on raw data

Data.Source	Page	Data.Files	Location	Provided	Cited
2009-2012 American Community Survey	3	Not available	Main file	Yes	Yes
2008,2010 National Survey of Recent College Graduates (NSRCG)	8	Not available	Main file	Yes	Yes
2009-2012 Student and Exchange Visitor Information System (SEVIS)	9	Not available	Main file	Yes	Yes
1977-2015 Integrated Postsecondary Education Data System Completion Surveys (IPEDS)	5	Not available	Main file	Yes	Yes

.

I.2 Description of the analysis data

In this section we highlight three main pieces of information about the analysis data. Firstly we list the names of the existing files, secondly we indicate their location in the reproduction package provided by the author. A

short description of each file is provided in a third and final point.

As a reminder, the analysis data are the data used as final inputs in a workflow to produce statistics and figures. to produce statistics and figures. The table summarises the main information collected on the analysis data provided by the author.

> Table 2 : Information on analysis data

Analytic.Data	Location	Description
Demirci_CJE_2020_figure1.dta	Main file	Generates figure 1
Demirci_CJE_2020_figure2.dta	Main file	Generates figure 2
Demirci_CJE_2020.dta	Main file	Generates all results except figures

.

I.3 Description of code scripts

In this section we highlight six main points about the codes provided in the initial package.

We first give the file name of the code, its location in the original reproduction folder, the inputs that go into the code roll. Then we give the names of the outputs that are produced, we describe them and finally we state the nature of the code (analysis or cleaning code).

> Table 3 : Code files information

File name	Location	Inputs	Outputs	Description	Primary type
Demirci_CJE_2020.do	Main file	Demirci_CJE_2020_figure1.dta	FIGURE 1	Produces FIGURE 1	Analysis code
Demirci_CJE_2020.do	Main file	Demirci_CJE_2020_figure2.dta	FIGURE 2	Produces FIGURE 2	Analysis code
Demirci_CJE_2020.do	Main file	Demirci_CJE_2020.dta	TABLE 1-6	Generates all results except figures	Analysis code

.

I.4 Connecting display items with inputs

In this section we construct, using the information from the previous sections, a diagram that visually presents the combination of components needed to reproduce a specific display item.

Thus, the following diagram describes how each figure and table in the paper was obtained using the codes and data provided in the original reproduction package.

The "X" in front of "TABLE" is a dummy variable which represents the different table numbers from 1 to 6.

```
FIGURE 1
└── Demirci_CJE_2020.do
    └── Demirci_CJE_2020_figure1.dta

FIGURE 2
└── Demirci_CJE_2020.do
    └── Demirci_CJE_2020_figure2.dta

TABLE X
└── Demirci_CJE_2020.do
    └── Demirci_CJE_2020.dta
```

Unused data sources:

```
Not available
```

Unused analytic data:

```
None - all analytic data files were used
```

.

I.5 Reproducibility score

In this section we assign, as objectively as possible, a score reflecting the level of reproducibility of each display item.

Based on the reproduction package provided by the author, we think that the reproducibility level is 5 for all display items; using the ACRe Reproducibility Scale.

In practical terms, this means that analytic data sets and analysis code are available and they produce the same results as presented in the paper (CRA). The reproducibility package may be improved by obtaining the original raw data.

.

I.6 Summary of assessment section

After a detailed analysis of the article and the reproduction elements provided by the author in the original package, we found the following :

- ☑ All display items are reproducible from the codes and analysis data provided
- ☑ No raw data files is provided
- ☑ No cleaning code file is provided for the operations leading to the analysis data
- ☑ The README file does not give enough details on how to get the raw data. It does not indicate whether site registrations, waiting periods or restrictions on access to certain confidential data are required.

Compensating for the missing elements of the initial reproduction package would considerably improve the reproducibility of the paper. This is what we attempt to do in the following section.

.

II. Improvements

In this section, we present the various improvements we made to the initial reproduction package to enhance paper's results replicability level.

.

II.1 Adding raw data : missing files

Given that the initial replication package contains only analytical data, we worked to provide raw data from which analytical data have been obtained.

After contacting the author of the article, he gave us the essential missing raw data files. Concerning data that he was not allowed to share with us for confidentiality reasons, the author provided us the necessary directions to obtain them.

Here are the raw data files obtained :

- ☑ figure1_raw.dta : IPEDS raw data for Figure 1
- ☑ SEVIS_aggregates.dta : aggregated raw data on international students labor supply
- ☑ Demirci_CJE_2020.dta : ACS raw data for econometric estimation
- ☑ 2004-2013SEVIS.dta : SEVIS raw data for Figure 2

.

II.2 Adding missing data cleaning code

Having obtained the missing raw data, we present here the codes that allow their cleaning or processing to obtain the analysis data.

After contacting the author, we obtained the following code files :

- ☑ figure1_clean.do : code that prepares IPEDS raw data for Figure 1
- ☑ figure2_clean.do : code that prepares the raw SEVIS data for Figure 2
- ☑ SEVIS_aggregates_clean.do : code that generates the SEVIS_aggregates.dta

.

II.3 New diagramm for onnecting display items with inputs

After gathering raw data and cleaning codes missing files, we analysed analysis data and codes initially provided by the author and we propose a more detailed scheme of display items creation.

The new diagram looks like this :

```
FIGURE 1
└── Demirci_CJE_2020.do
    └── Demirci_CJE_2020_figure1.dta
        └── figure1_clean.do
            └── figure1_raw.dta

FIGURE 2
└── Demirci_CJE_2020.do
    └── Demirci_CJE_2020_figure2.dta
        └── figure2_clean.do
            └── 2004-2013SEVIS.dta

TABLE X
└── Demirci_CJE_2020.do
    └── Demirci_CJE_2020.dta
        └── MERGE
            ├── Demirci_CJE_2020_1.dta
            │   └── SEVIS_aggregates_clean.do
            │       └── Demirci_CJE_2020_0.dta
            └── SEVIS_aggregates.dta
                └── SEVIS_aggregates_clean.do
                    └── 2004-2013SEVIS.dta
```

Unused data sources:

```
Not available
```

Unused analytic data:

```
None - all analytic data files were used
```

NB : Demirci_CJE_2020.dta_0 and Demirci_CJE_2020.dta_1 files are only the initial and cleaned version of Demirci_CJE_2020.dta analysis data file.

From this diagram we can clearly see the transformation of raw data into analysis data, and cleaning and analysis codes that have been used. Also we can distinguish the data files merge that produced the final analysis data used by the author for econometric estimations.

.

II.4 Adding information on how to access confidential/proprietary data

The confidential data that the author could not share with us are the raw SEVIS data that he obtained through a FOIA request. This student data can be obtained from U.S. Immigration and Customs Enforcement via a Freedom of Information Act (FOIA) request on their website.

For more details on access to this confidential data, please see the Data Availability Statement by clicking here. The adding of this document fulfils the AEA's requirements about data and code availability policy.

By adding new files and providing detailed informations on access to confidential data, we have been able to improve display items reproducibility score. While initially all display items had a reproducibility score of 5;

TABLE 3, TABLE 4 and Figure 2 have now achieved a score of 9/10. Figure 1 has reached the ACRe maximum level as all necessary documents are available and produce the expected result.

.

II.5 Documenting the improvements using version control

In order to facilitate the replicability of the results of the paper by future replicators, I used the GitHub version control software to document all the improvements I made to the original replication package provided by the author. These changes can be seen by visiting my GitHub account.

The main code "Demirci_CJE_2020.do" contains commands that are incompatible with Stata 17 and probably with other versions of Stata. To solve this problem I added the syntax "version 16" at the beginning of the main code: this allows to use the functionalities of Stata 16 which is the most adapted version.

One of the changes I have made is to improve the original README file provided by the author. I have added more detailed and accurate informations on access to raw data opened to the public as well as confidential data. The final README will also include the new digram which better presents the combination of inputs to obtain the outputs displayed in the paper.

In addition to various improvements mentioned above, we believe that many other actions could enrich the analysis of the paper we have studied. For example, a next step could be to rewrite all the initial code in another statistical software like R and do a robustness check of the paper's results.

Another version of this report can be obtained by consulting my account on Social Science Reproduction Platform (SSRP).