Replication Project for the paper Immigration and Crime: An International Perspective

Yu-Chiao Tseng and Tawanda Nhundu

This is a report of a replication study. Each group can freely choose a research paper and try to interpret the code provided by the author and replicate the charts with R.The replication process can be assisted by ChatGPT and the file needs to be published on GitHub.This report will cover all the steps to complete the report, the codes used and the problems encountered during the replication. Although the process was not always smooth, we managed to complete this report.

Introduction

The goal of this project is to conduct a replication study. Our group selected the paper from Olivier Marie and Paolo Pinotti (2024) which explored the association between immigration and crime in the Journal of Economic Perspectives. They found that an increase in the share of immigrants was not accompanied by an increase in the global murder rate.

We will discuss more detail the process of finding papers, how we used R to complete this replication study, the challenges we encountered, and how we published our report on GitHub. Lastly, we will summarize the learning process.

Motivation for Replication

Academic research nowadays values transparency and credibility, and replication studies are an important way to verify these.There are some researchers manipulate data to achieve the results they want, and replication studies can confirm that the author's research results are not accidental or the result of incorrect analysis method. The importance of research in science and academic knowledge is self-evident. According to Bouter & ter Riet (2021) replication is divided into three different levels. First, reanalysis of the original study by re-running their exact code and dataset, to verify that the published figures and tables can be fully reproduced. The second one is direct replication which means replicate with new data but the same research protocol. The third one is conceptual replication, researchers use new data with the modified research protocol and same research question. In the project I think we did re-analysis, we used the data provided by the authors and the same data analysis plan.Using the world bank data as weight can be a conceptual replication. Bouter & ter Riet mentioned that by varying data sources or methods, conceptual replications can reveal whether a finding can be use in different context and proof it validity and generalizability. In contrast, reanalyses or direct replications which rely on the same dataset or same procedures might reproduce the flaws that were present in the original study.

(Bouter & Riet, 2021)

Finding a paper

We selected two papers, "Inside the Box" and "Immigration and Crime" at the beginning. The first one was written by an author who used R to do most of his work, so the professor suggested that we choose another one. In the article "Immigration and Crime", the author used Stata for coding. The article explored the relationship between immigration and crime. Living abroad as students, we felt that the increase in immigration in many countries in recent years has brought benefits such as filling the labor gap in developed countries, but it has also caused many conflicts between immigrants and locals. Many people believe that immigrants cause public security problems and have low tolerance for them. Seeing this trend, enhanced our interest in this topic, and we hope to learn more about the content of the article and how the author conducts statistical analysis and explores this issue.

The authors Olivier Marie and Paolo Pinotti cleaned, merged, and statistically analyzed the United Nations' immigration and crime data to explore whether there is a significant correlation between immigration and crime. They also referred to other studies to further explore whether obtaining legal status reduces the tendency to commit crimes. We chose Figure 2 and Figure 4 for replication. First, we read the author's README file and Stata code to confirm the feasibility and verify whether the data set can be accessed from the websites they provided.

Figure 2 Top Replication Process

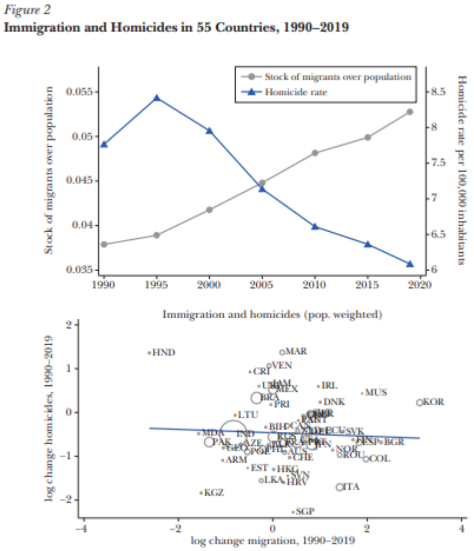
Figure 2 is composed of two figures, the first one is a double y-axis graph, showing the long-term trend of the "proportion of immigrants to the total population" gray line and the

Correspondence concerning this article should be addressed to Yu-Chiao Tseng, Email: tseng.yu-chiao@stud.hs-fresenius.deTawanda Nhundu, Email: nhundu.tawanda@stud.hs-fresenius.de

"murder rate" blue line in 55 countries (1990-2019). The second figure is a scatter plot, showing how much the "number of immigrants" in each country changed from 1990 to 2019 and how much the "murder rate" changed during the same period (calculated in logarithms).

Figure 1.

Original graph of Figure 2

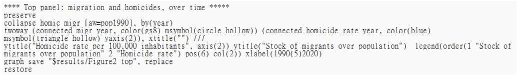


(Marie & Pinotti, 2024)

Below is the original Stata code for Figure 2 top

Figure 2.

Original Stata code Figure 2 top



```
**** Top panel: migration and homicides, over time *****
preserve
collapse homic migr [aw=pop1990], by(year)
twoway (connected migr year, color(gs8) msymbol(circle hollow)) (connected homicide rate year, color(blue)
msymbol(triangle hollow) yaxis(2)), xtitle("") ///
ytitle("Homicide rate per 100,000 inhabitants", axis(2)) ytitle("Stock of migrants over population")  legend(order(1 "Stock of
migrants over population" 2 "Homicide rate") pos(6) col(2)) xlabel(1990(5)2020)
graph save "$results/Figure2 top", replace
restore
```

This part I relied on chatgpt's explanation to understand the meaning of each code collapse homic migr [aw=pop1990], by(year) is grouped by year, and the variables homic and migr are weighted averaged (weight is pop1990), Two-way connected plot is a two-line graph, connected migr year, color(gs8) msymbol(circle_hollow) draws the first line (left Y axis), connected homicide_rate year, color(blue) msymbol(triangle_hollow) yaxis(2) draws the second line (right Y axis). Then I created a new file in R and saved the data provided by the author. ChatGPT guided me step by step to reproduce this grpah in R.

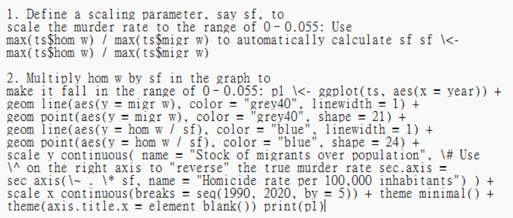
(OpenAI, 2025)

Here is the first problem. As we can see, the gray line on the graph made according to the above code is almost attached to the bottom of the graph. I asked Chatgpt, and it said that the gray "immigrant proportion" line is always close to 0. This is because the dual Y axes of ggplot2 are not truly "independently scaled", but share a set of numerical spaces. hom_w (murder rate) is about 6 to 9, and migr_w (immigrant proportion) is about 0.03 to 0.055. After looking at the graph, I also noticed that the scales on both sides of my graph are 0-8, while the scale on the left side of the author's graph is 0.035-0.055.

Then ChatGPT Offered Solution:

Figure 3.

Solution for flat gray line



```
1. Define a scaling parameter, say sf, to
scale the murder rate to the range of 0 - 0.055: Use
max(ts$hom w) / max(ts$migr w) to automatically calculate sf sf \<-
max(ts$hom w) / max(ts$migr w)

2. Multiply hom w by sf in the graph to
make it fall in the range of 0 - 0.055: p1 \<- ggplot(ts, aes(x = year)) +
geom line(aes(y = migr w), color = "grey40", linewidth = 1) +
geom point(aes(y = migr w), color = "grey40", shape = 21) +
geom line(aes(y = hom w / sf), color = "blue", linewidth = 1) +
geom point(aes(y = hom w / sf), color = "blue", shape = 24) +
scale y continuous( name = "Stock of migrants over population", \# Use
\^ on the right axis to "reverse" the true murder rate sec.axis =
sec axis(\~ . \* sf, name = "Homicide rate per 100,000 inhabitants") ) +
scale x continuous(breaks = seq(1990, 2020, by = 5)) + theme minimal() +
theme(axis.title.x = element blank()) print(p1)
```

(OpenAI, 2025)

Here I follow the instruction, use y = hom_w / sf when drawing the blue line (equivalent to "compressing" the murder rate to the small interval of the proportion of immigrants), Then use sec_axis(~ . * sf) to "expand" the label on the right back to the true murder rate value. After following the steps of chatgpt, the two lines are in the correct position. Lastly, I corrected the scales on both sides to 0.035-0.055 on the left axis and 6-8.5 on the right axis.

After following the steps of ChatGPT, the two lines are now in the correct position. Lastly, I corrected the scales on both sides to 0.035-0.055 on the left axis and 6-8.5 on the right axis to make it more alike to original one.

I also noticed that the gray line ended up closer to 8.5, while the original graph looked between 8 and 8.5. I asked ChatGPT and checked three things according to it's instructions. 1. Did I draw 2020 ? 2. Confirm that migr_pop is a proportion, not a number of people 3. Re-ensure that the weighted average uses the pop1990 column I checked all of them, but the graph did not change. ChatGPT said that it might be that the weighting method in Stata is slightly different from the implementation details in R, or that the original graph was slightly adjusted manually when it was published.(OpenAI, 2025) I remember learning in the course last semester that the key to doing replication is to be able to reproduce the results and trends that the chart wants to present after following the code provided by the author and the same method. In this graph, the direction of the lines is consistent with the changes (immigration increased, murders decreased), and the values are roughly in the same area, with no obvious misalignment or data errors. I think the points where the two lines fall and where they intersect are not significantly different from the original graph, so I believe this is a successful replication.

Discussion on Figure 2 top

Figure 2 top graph uses United Nations immigration data. I think the author could be more precise on that. Did they use International migrant stock from United Nations data or Total population, or both? At first, I compared the numbers on the data and the website, and found that there was a slight difference between the total population on the website and the file provided by the author. For example, the total population of Australia in 1990 was 16960.600 in the author's data and 17126.298 on the website. The total population of Armenia in 1990 was 3538.164, and 3552.128 on the website. I thought that this might be because the data on the website had been updated. If the difference was not big, I thought the data provided by the author was credible, so I directly used the author's data for replication. Later, I further wanted to confirm where the number of migr_pop in the author's file came from, but I couldn't calculate a similar number. Finally, I realized the problem is that I downloaded the latest version of the data, and the author used the 2020 version. The latest version is Armenia 1990 Total population at mid-year 3552.128, migrant stock 433541. The version used by the author, Armenia 1990 Total population at mid-year 3538.164, migrant stock has been updated a lot, which led to my calculation of 433541/3552.128 ≈0.1225, not ≈0.186. Another issue was the numbers for migrant stock are missing from the file provided by the authors; it has only the author's own calculation of migr_pop 0.186195165. I think the author's readme file can include the file year used and how migr_pop is calculated. This can improve the overall transparency and credibility of the research and reduce disputes over errors caused by different versions.

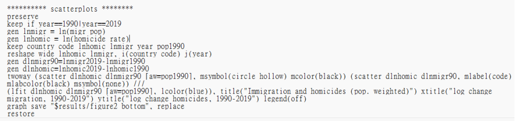
Figure 2 Bottom Replication Process

The bottom plot is the logarithmic change scatter plot + weighted regression line for 1990 vs. 2019. X axis: log change migration, Y axis: log change homicides

Below is the original Stata code for Figure 2

Figure 4.

original Stata code Figure 2 bottom



```
********** scatterplots ********
preserve
keep if year==1990|year==2019
gen lnmigr = ln(migr_pop)
gen lnhomic = ln(homicide_rate)
keep country_code lnhomic lnmigr year pop1990
reshape wide lnhomic lnmigr, i(country_code) j(year)
gen dlnmigr90=lnmigr2019-lnmigr1990
gen dlnhomic=lnhomic2019-lnhomic1990
twoway (scatter dlnhomic dlnmigr90 [aw=pop1990], msymbol(circle_hollow) mcolor(black)) (scatter dlnhomic dlnmigr90, mlabel(code)
mlabcolor(black) msymbol(none)) ///
(lfit dlnhomic dlnmigr90 [aw=pop1990], lcolor(blue)), title("Immigration and homicides (pop. weighted)") xtitle("log change
migration, 1990-2019") ytitle("log change homicides, 1990-2019") legend(off)
graph save "$results/figure2_bottom", replace
restore
```

Same I followed the instructions of ChatGPT to do the steps, first select 1990 and 2019 from the original data

It looks very similar to the original image, and then I adjusted the scale.

Lastly, I use the code theme_minimal(base_size = 38) for the font size adjustment. I think this graph shares the same idea that the author wants to present in the paper. There is a nearly horizontal regression line, and most countries are concentrated between -1 and 1 on the X axis. This means that the proportion of immigrants and the murder rate in most countries have not experienced a huge change. If more immigrants lead to higher crime, there will be a cluster in the upper right corner of the graph. However, the graph shows that there is no consistent trend or causal relationship between changes in immigration and changes in murder rates.

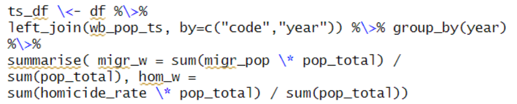
Using World Bank population data as weight

Because the group before us had some data generation problems, they didn't seem to use the data used by the author, but the graph was produced. The professor said that their data might be generated by ChatGPT itself, and then said that they could go to the World Bank to find the data they needed. At that time, I thought I had to use the World Bank data, so I used the World Bank data to make the second graph. But I didn't give ChatGPT instructions clear enough; I just said that I wanted to use the World Bank population data to make a new graph.So ChatGPT gave me the code:

The step after merging had mistake

Figure 5.

Mistake using incorrect weights



```
ts_df \<- df %\>%
left_join(wb_pop_ts, by=c("code","year")) %\>% group_by(year)
%\>%
summarise( migr_w = sum(migr_pop \* pop_total) /
sum(pop_total), hom_w =
sum(homicide_rate \* pop_total) / sum(pop_total))
```

Because I used the total population of all years from 1990 to 2019, what I got was the "weighted average of the population distribution of that year", which means that the weights change every year. Actually, I don't need the total population of all years from 1990 to 2019. I only need to replace the 1990 population (pop1990) with World Bank's data, because the author used only the 1990 population (pop1990) as the weight for all periods.

Second time, I only used the 1990 population as the weight for the whole period, and I still use the migr_pop calculated by the author as a share.

Discussion of using the data from the World Bank incorrectly

This graph looks very different from the last graph with mistakes, and doesn't look significantly different from the first time I used the data provided by the author.

This time, I used the data from the World Bank incorrectly and didn't notice it at all. I intuitively thought that the graph looked different only because I used the data from the World Bank. This caused me to be embarrassed during my oral presentation.

Problems with making a scatter plot at the bottom

I also tried to make the bottom plot using the data from the World Bank, but the problem I had was scatter plots did not appear.

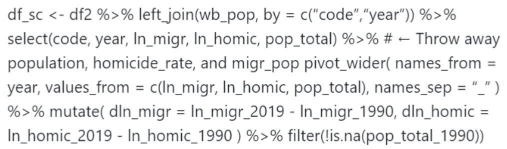
I asked ChatGPT why this code can't make a scatter plot; the plot is blank, there is no dots. At first, it couldn't find the reason, and ChatGPT kept going around in circles. After hours, I changed the conversation to a new chat room and reposted the code. It finally works and I just checked it step by step according to its instructions. For example, check whether there are rows in the final data frame before drawing, and checked which countries are not in wb_pop.

Finally, we found the problem, and ChatGPT explained it this way: The problem is in the plot data frame, all rows corresponding to x = dln_migr or y = dln_homic are treated as NA, so ggplot automatically discards them.

The solution provided from ChatGPT is to keep only the columns I need before pivoting

Figure 4.

Solution for blank scatter plot



```
df_sc <- df2 %>% left_join(wb_pop, by = c("code","year")) %>%
select(code, year, ln_migr, ln_homic, pop_total) %>% # ← Throw away
population, homicide_rate, and migr_pop pivot_wider( names_from =
year, values_from = c(ln_migr, ln_homic, pop_total), names_sep = "_" )
%>% mutate( dln_migr = ln_migr_2019 - ln_migr_1990, dln_homic =
ln_homic_2019 - ln_homic_1990 ) %>% filter(!is.na(pop_total_1990))
```

(OpenAI, 2025)

After correcting according to the instructions, the scattered points appeared.

Conclusion Summarizing

Through this replication process and the professor's explanation in class, I understand the importance of replication for research. It not only confirms the transparency and credibility of the research, can also checks whether there are errors or artificial manipulation of data in the analysis process. The first time I used all the data provided by the author to make the chart, and only compared the numbers on the website with the file. The second time I used the data from the World Bank, although it was under a misunderstanding of the discussion in class, which resulted in many errors, but on the other hand taught me a lot.

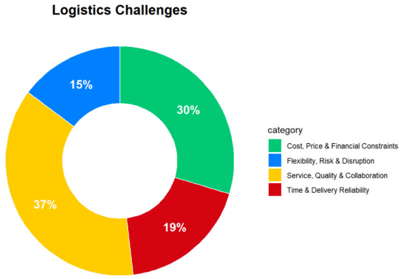
Honestly, after the presentation, I still don't know if it makes sense for me to use the data from the World Bank, cause I still use the original calculation of migr_pop by the author. I asked ChatGPT, and its answer was: "If the purpose of your research is just want to see what happens if you use the WDI population as weight, then it makes sense to do so: it can answer a new question: If I use the World Bank's 1990 population as weight instead, how will the trend of the global immigrant share be fine-tuned? This may also be an interesting robustness check in policy or methodological discussions." Seeing the answer, I was glad that it was not a waste of time. Although using the World Bank data may not help replicate the graph itself, it's meaningful to the learning process, and it allowed me to learn how to add the World Bank data in R and use it correctly.

If I hadn't taken this class, I would have never thought about learning how to use R or how to write code. I often couldn't understand what the professor was doing in class, I felt lost in the lecture or I forgott what to do after Professor finished demonstrating. It started changing until I had to make this Replication report. At the beginning I did everything slowly, and when R shows that I had error, I didn't know what happened. I had to take screenshot and sent it to ChatGPT.Under the guidance of ChatGPT, I was able to create the graph step by step. Although I still made a lot of mistakes like the first time I tried to replicate the graph, I put all the code into the console, and it worked,the graph showed, but I was not able to keep the code for the report. When modifying some part of the code, I also made mistakes such as missing a) or missing a comma, adding an extra comma, and made an Incorrect data path etc.Some of the mistakes took me a lot of time to figure the problems out and fixed them.Lack of familiarity with R often slows down our progress and was very frustrating.

After I learnt how to create graphs, I even used R to create more graphs in another group project, which was something I had never expected at the beginning of the semester, and I am proud of my learning process.

Figure 6.

Graph for DHL Project created with R



Limitations and Future Directions

The Limitations of this report, we only found data that the author had cleaned up, which contained the migr_pop calculated by the authors. We only compared the total population and murder rate provided by the authors with the numbers on the websites, and did not download the original data of these two catagories and recalculate migr_pop to verify the numbers provided by the authors.

I think the further direction of this replication study could be to first use the latest version of immigration data on the UN website for Figure 2. As I mentioned earlier, the numbers of migration stock are quite different from the previous version. However, this would involve recalculating migr_pop ourselves. second, other violent crimes can also be included to re-examine the relationship between immigration and crime rates. Third, use the same analysis method to explore different cities in a certain country.

References

Bouter, L. M., & Riet, G. ter. (2021). Replication Research Series-Paper 2 : Empirical research must be replicated before its findings can be trusted. *Journal of Clinical Epidemiology*, *129*, 188–190. https://doi.org/10.1016/j.jclinepi.2020.09.032

Marie, O., & Pinotti, P. (2024). Immigration and Crime: An International Perspective. *Journal of Economic Perspectives*, *38*(1), 181–200. https://doi.org/10.1257/jep.38.1.181

OpenAI. (2025). ChatGPT [Large language model]. https://chat.openai.com/chat

Appendix A

Appendix B