Reproduction of Civic Honesty around the Globe*

TBD

Bongju Yoo and Najma Osman

23 March 2021

Abstract

By reproducing the research paper, Civic honesty around the globe, we examined what sampling methods the authors chose and how the data was gathered, and verified the regression models used in the original paper. We discovered that some of the variables in the original dataset contained the data collector's estimation and judgment, and the rationale for selecting the countries was weak and unclear. The original research paper can be extended and/or improved with right choices of nations, more concrete criteria to measure abstract concepts, and considerations of socioeconomic status of the areas.

1   Introduction

This paper is the reproduction of the earlier research paper, Civic honesty around the globe, studied by Alain Cohn, Michel André Maréchal, David Tannenbaum, and Christian Lukas Zünd and proceeded between July 2013 and December 2016 (Cohn et al. 2019). The research team conducted field experiments in 40 countries around the world to examine the "trade-off between honesty and self-interest" (Cohn et al. 2019). The purpose of this reproduction work is to identify potential limitations and hidden biases from the original dataset and experimental design of the research paper, and suggest ways to extend and/or improve the original research paper.

Firstly, we looked into the original dataset and examined what sampling approach the research team used and how the data was collected. We found that some of the variables in the dataset were recorded highly based on the research assistant's estimation and judgment since they had to measure abstract concepts such as actions and behaviors. Also, we found that the rationale for selecting the countries was ambiguous. These inaccurate and unclear methods of collecting the data can lead to lower the reliability and validity of the results of the research. Additionally, we ran power analysis to verify whether or not the sample sizes used in the dataset were appropriate and confirmed that the sizes were large enough to detect a statistically significant difference between the control group and treatment group. Secondly, we identified which independent variables impact the most significant influence on dependent variables, using ordinary least squares (OLS) regression. The results of our regression models were slightly different from the original ones: the absence or presence of coworker/coworkers was statistically significant; otherwise, being ages 40+ was not statistically significant. However, in general, most of the results were the same as the original paper.

2   Data

2.1 Description of Dataset

This type of the dataset is a csv file, it is imported from the Harvard Dataverse Repository and uploaded June 15th, 2019 (Harvard College 2021). There were 57 variables and 17,303 observations, but we selected 12 variables using the R programming language (R Core Team 2020) and the 'tidyverse' package (Wickham et al. 2019). The selected variables are: id, country (unique numeric code for each country), Country (the name of the country), city, institution (drop-off locations), cond (no money, money, and big money conditions), response (reported or not), age, above40 (the estimated age of the recipient), male (the gender of the recipient), coworkers, computer, other_bystanders. All of those attributes are coded as numerical values except for Country, which is code as alphabetical values. The key variables are response, country, cond, and response: the response variable was used as a dependent variable, the no money condition was used as a control variable, and the other two conditions were used as treatments.

In the original dataset, there were two variables which may cause confusion due to their similarity: 'country' with lower case 'c' and 'Country' with upper case 'C'. We used the 'janitor' package's 'clean_names" function to change the names to 'country' and 'country_2' to avoid confusion (Firke 2021).

2.2 Methodology and Data Collection

The research team visited 355 cities in 40 countries and dropped a total of 17,303 wallets between July 2013 and December 2016 (Cohn et al. 2019). The target sample size was set at 400 per country except for the US, UK and Poland: the sample size for these three countries is greater than 400 since two additional treatment conditions (BigMoney and Money-NoKey conditions) are included (Cohn et al. 2019). The dataset contains not just where and when the wallets were dropped and reported but also situational factors that may affect the response of the experiment, such as whether there was a computer at the recipient's desk, how many witnesses or coworkers were there when the wallet was handed over to the recipient, or whether a security camera was visible at the recipient's desk (Cohn et al. 2019).

These situational factors were recorded by research assistants, and some of the factors highly rely on their individual judgment. For example, the 'Busyness' category, which is measured on a 7-point scale from "not at all" (0) to "very busy", and the 'Friendliness' category, which is measured on a 7-point scale from "very unfriendly" (0) to "very friendly" (6), were graded by the assistants according to their estimation and judgment since they had to measure actions and behaviors which are intangible unlike a computer or security camera (Cohn et al. 2019).

In addition, the estimation of the assistants can be incorrect, and eventually they might record false information during the data collection process. For example, in terms of the 'Recipient age' category, the assistants had to guess the recipient's age and select one from six options: < 20, 20-30, 30-40, 40-50, 50-60, and > 60. What if the recipient looks 30? The assistants might choose either 20-30 or 30-40 based on their judgement since it is a huge challenge to correctly guess someone's age by looking at his or her face. A research team from the Department of Computer Science and Engineering at Michigan State University studied the ability of humans to estimate age using face images collected from crowdsourcing and found that the accuracy of guessing the age correctly with a margin of error of plus/minus 5 years is between 50% and 70% (Han, Otto, and Jain 2013). Namely, there are high chances that the assistants might have recorded inaccurate information for some of those variables that require the researcher's estimation. As a consequence, this can lead to distorted results of the analysis.

The selection of countries raises another issue of the dataset: Do those countries really represent the globe? 40 countries are selected and examined for the research, and almost half (19 countries) of them are European countries. However, in reality, the population of all European countries is approximately 10% of the world's population (Worldometer 2021a). In terms of the number of countries, there are merely 44 countries in Europe, out of 195 countries in the world, which accounts for about one-fourth of the total countries (Worldometer 2021b). Is it because those unselected countries could not meet the requirements of selection criteria? The selection criteria requirements for the country category are as follows:

- The country must have a sufficient number (at least 100,000) of large cities;

- The country must have a plenty of "feasible drop-off locations" where can be easily accessible;
- The country must be "easy to visit and safe enough" for research assistants to proceed with the wallet dropping experiment; and
- The country must have less strict customs, immigration, and banking regulation (Cohn et al. 2019).

It can be arguable that UAE, Malaysia, and Kazakhstan were selected instead of other Asian countries with larger population size and more large cities such as Japan, Philippines, Vietnam, and South Korea (World Population Review 2021). Japan was not included in the category even though the country has the most and tenth-most populous cities in the world: Tokyo and Osaka (World Population Review 2021). Furthermore, Japan was one of the "10 Safest Countries In The World To Visit" chosen by Forbes in 2020; otherwise, none of the selected Asian countries was not included in the list (Ledsom 2020).

The research data was collected by 13 research assistants: 11 male and 2 female born between 1985 and 1993. The assistants were instructed to say: "Hi, I found this [showing the wallet] on the street just around the corner" (Cohn et al. 2019). After handing the wallet over to the recipient, they were also guided to say: "Somebody must have lost it. I'm in a hurry and have to go. Can you please take care of it?" (Cohn et al. 2019). If the recipient contacts the email written on the business card inside the wallet to return it, the response variable is recorded as 1, meaning being responded. Even if the recipient makes a minor typo on the email address, the recipient email is designed to receive it unless misspelling the domain name (Cohn et al. 2019).

2.3   Cost

The wallets for the money condition contained the equivalent of US $13.45, and the ones for the big money condition contained the equivalent of US $94.15 (Cohn et al. 2019). So, the total amount of the cash spent for the experiment was $249,000, including exchange fees. Additionally, $3,600 was spent for the production of the wallets: $1,000 for printing business cards, $800 for producing plastic wallets, and $1,800 for copying keys. The research assistants were paid an hourly wage of $25 per hour, and the total wage expense was $300,000. The total expenditure on transportation such as flights and car rentals was $150,000.

2.4   Power Analysis

We ran power analysis to check whether the research team used an appropriate sample size for the original paper. Firstly, we calculated the proportions of the returned wallets in the control group (the 'no money' condition) and the treatment group ('the 'money' condition), using the raw data using the dplyr's select, group_by, summarise, and mutate functions (Wickham et al. 2020). Then, we calculated the effect size, using the formula written by Stephane Champely (Champely 2020a). Lastly, we ran the two-sample test for proportions to estimate the required sample size which is required to show a significant difference between the two groups, using the 'pwr' package (Champely 2020b). The result of the test shows that the most appropriate size of a sample to compare the two condition groups is 312 for each. Since the sample size of the control group is 7,888 and that of the treatment group is 8,211, we can say that the sample size is large enough to find a statistically significant difference between the two groups. The result of the power analysis is displayed below.
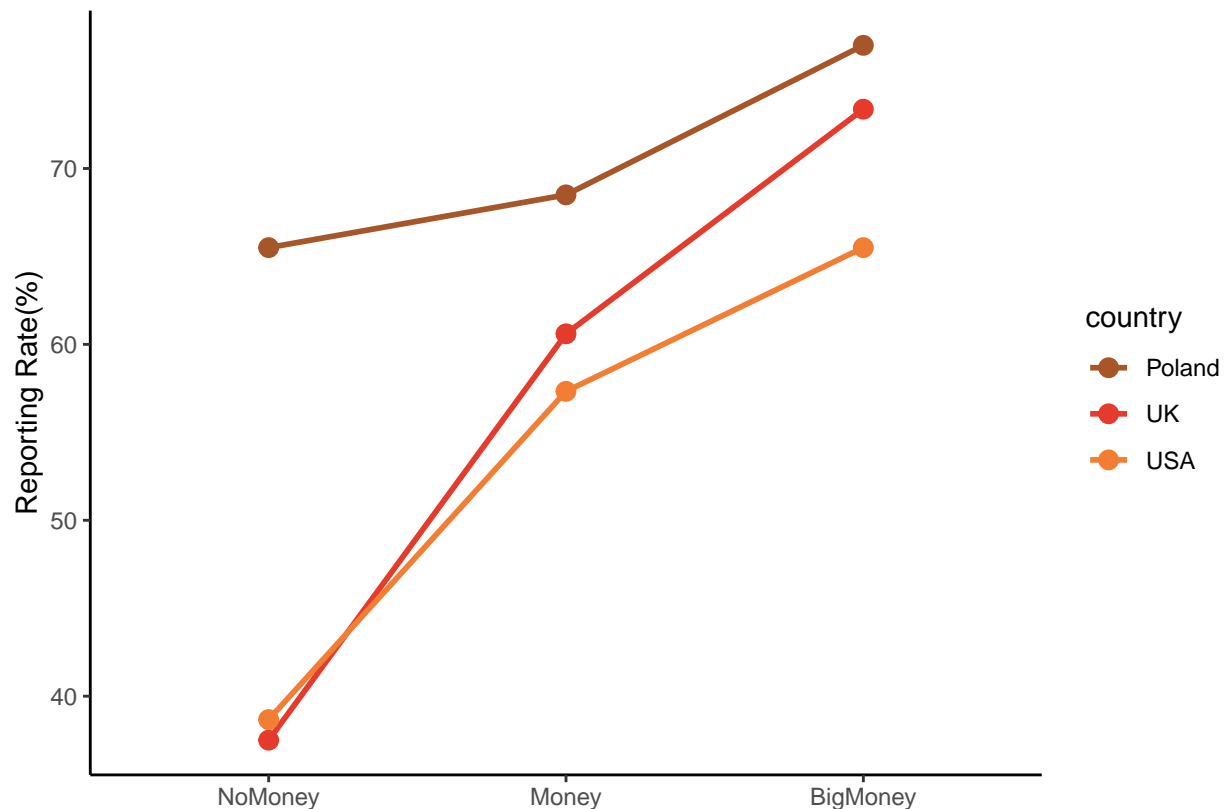
```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##              h = 0.2242923
##              n = 312.0382
##      sig.level = 0.05
##          power = 0.8
##    alternative = two.sided
```

```
## 
## NOTE: same sample sizes
```

Share of wallets reported by country
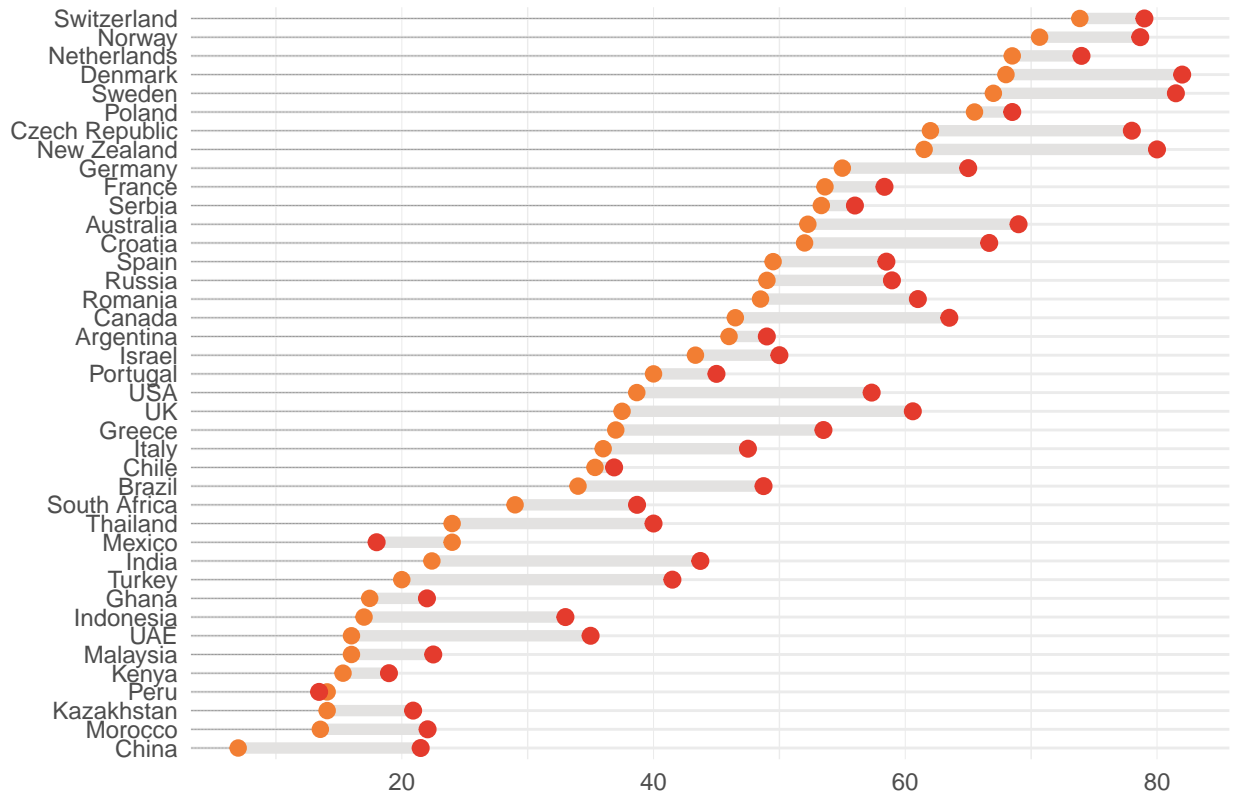
3 Model

We used a linear regression model, specifically ordinary least squares (OLS), which are used to examine relationships between variables, specifically looking to determine which independent variables hold the most and/or significant influence over dependent variables - in this case, whether the wallet is reported or not. In addition to this, we can see how changes in the independent variables are related to changes in the dependent variable through use of dummy variables. It allows us to understand the mean change in a dependent variable, given a 1 unit change in each independent variable. Three regressions were run. Table 1's second column and Table 2 consists of multiple linear regressions, Table 1's first column and Table 3 are simple linear regressions.

3.1 Formulae

Table 1 (column 1) + Table 3:
$$Y \sim \beta_0 + \beta_1 X_1 + \epsilon$$

Table 1 (column 2):

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

Table 2:

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

3.2 Model Features/Aspects

Like stated prior, 3 OLS regressions are run within this paper to replicate tables S8, S9, and S10 in the original paper, respectively Tables 1, 2, and 3 within this paper.

- Treatment Condition (cond, as found in the cleaned model dataset):
  - Originally, the treatment condition is a categorical variable from 0-3, where 0 = NoMoney, 1 = Money, 2 = Big Money, 3 = Money-NoKey.
  - To replicate S8, the treatment condition had to be re-coded into a dichotomous variable where 0 is the NoMoney condition, and 1 comprises the Money, BigMoney, and Money-NoKey conditions.
  - For S9, dummy variables had to be created in order to focus on the BigMoney and Money conditions,
  - And for S9, the treatment condition was again re-coded so 1 referred to the Money-NoKey condition, and 0 comprised all other conditions.

In doing so, we could isolate for the correct conditions when running the model.

The remaining variables within the model, such as whether the respondent is male, above 40, coworkers, a computer, or other bystanders are present, are all binary variables. Using dummy variables, and/or avoiding categorical ones was necessary to ensure that those demarcated with a value of 0 will cause that particular variable to have no role in influencing the dependent variable, which in this instance is response/reporting of the wallet.

3.3 Justification

For the purposes of this study, a regression analysis makes sense due to the use of a categorical/binary dependent variable, whether the person reports the lost wallet to the owner, and continuous or binary independent variables. Linear regression, as opposed to nonlinear regression models, are a better fit for this study due to their simplicity, assuming adequate fit in the residual plots.

Assumptions underpinning OLS models include:

- No heteroscedasticity: the error term has a constant variance,
- Linearity in Parameters: the regression model is linear in coefficients and error term,
- Zero conditional mean: error term has a population mean of zero
- Variation in X: there is variation in the explanatory variable,
- Normality - No IV is a perfect linear function of other explanatory variables,
- And Random Sampling: the observed data represents a random sample from the population

3.3.1 Alternatives

According to Cohn et al's (2019) supplementary materials, OLS was chosen over non-linear models (e.g., logistic regression) as they return virtually identical results. In addition to this, OLS was used for simplicity and clarity as "coefficients can be directly interpreted as percent point changes." OLS is easier to perform, and interpret. Nonlinear regressions use iterative algorithms as opposed to the linear, solving with matrix equations. This introduces worry about algorithm choice, starting values, and convergence possibilities.

3.4 Model Convergence/Checks

RMSE is the square root of the variance of residuals, and indicates the absolute fit of the model to the data. Specifically, it tells us how close the observed data points are to the model's predicted values, and as such,

can be interpreted as the standard deviation of the unexplained variance. Lower values of RMSE indicate better fit, with 0 meaning a perfect fit. Once again, the "response" in this particular model is whether the recipient of the wallet does or does not report it stolen - 0, or 100. The RMSE for each regression are as follows:

- Table 1
    - Col 1: RMSE = 49.76789
    - Col 2: RMSE = 49.01223

- Table 2
    - Col 1: RMSE = 48.53371
    - Col 2: RMSE = 48.30933
    - Col 3: RMSE = 46.21193
    - Col 4: RMSE = 49.22719

- Table 3
    - Col 1: RMSE = 49.13414
    - Col 2: RMSE = 49.43042
    - Col 3: RMSE = 46.35512
    - Col 4: RMSE = 49.9143

The high RMSE here may partially be as a result of failure to account for fixed effects in the models, such as institution and city. That withstanding, since the response data ranges from 0 to 100, the RMSE's result uses the same unit, the numbers here are incredibly large and do not suggest a great fit. However, considering doesn't truly range from 0 to 100, and is rather a binary 0 or 100 for non-response or response, this may have affected the result.

3.5   Software Used

To run the OLS model, we used the R Core Team (2020) inbuilt function, lm(), which is used to fit linear models. It has two key arguments: formula, and data. Formula takes on the form y ~ x1 + x2, where y is the dependent variable and x1, x2, and onward are the independent variable, while data is the data frame containing the columns specified in the formula. The stargazer package by Hlavac (2018) was used to create the LaTeX regression tables.

4   Results

Table 1's results are aggregated across the 40 countries visited to conduct the experiment. Results show the following: * Lost wallet reporting rates increase by 12 percentage points in the Money relative to the NoMoney condition * Men are less likely to report a wallet than woman the presence of a computer increased likelihood of reporting the lost wallet, * but the presence of other bystanders decreased reporting rates, * However, unlike in the original paper, age groups (above and under 40), and the presence of coworkers had no statistical significance regarding reporting or failing to report lost wallets.

Table 2 examines reporting rates for the Money and BigMoney conditions in the UK, Poland, and the USA. * Column 1 shows that across the three countries, they were more likely to report a lost wallet containing greater amounts of money. * Columns 2-4, where the countries are isolated, shows a trend that wallets with larger amounts of money are more likely to be reported. * With smaller amounts of money, only the UK and US showed statistically significant likelihood of reporting the lost wallets.

Table 3 looks at reporting rates for the Money-NoKey conditions in the UK, Poland, and the USA. * Column 1 shows that fewer wallets were reported when they did not contain a key in the 3 countries. * Columns

Table 1: Reporting rates in the Money and No Money Condition

	Dependent variable:	
	Response	
	(1)	(2)
Money	12.157***	10.527***
	(1.540)	(1.527)
Male		−6.910***
		(1.041)
Above 40		−1.449
		(1.035)
Computer		17.210***
		(1.275)
Coworkers		1.011
		(1.075)
Other Bystanders		−6.121***
		(1.085)
Constant	51.053***	45.185***
	(0.549)	(1.544)
Observations	9,407	9,407
R^2	0.007	0.037
Adjusted R^2	0.006	0.036
Residual Std. Error	49.773 (df = 9405)	49.030 (df = 9400)
F Statistic	62.281*** (df = 1; 9405)	59.378*** (df = 6; 9400)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Reporting rates in NoMoney, Money, and Big Money condition

	Dependent variable:			
	Response			
	UK, Poland, and US	United Kingdom	Poland	United States
	(1)	(2)	(3)	(4)
Money	11.317***	17.600***	5.779	12.333***
	(2.025)	(3.200)	(4.020)	(3.600)
Big Money	22.225***	30.367***	14.600***	20.500***
	(2.403)	(4.196)	(4.033)	(4.125)
Constant	49.923***	43.000***	63.065***	45.000***
	(1.348)	(2.419)	(2.321)	(2.205)
Observations	2,926	1,132	794	1,000
R^2	0.030	0.050	0.016	0.028
Adjusted R^2	0.029	0.048	0.014	0.026
Residual Std. Error	48.559 (df = 2923)	48.373 (df = 1129)	46.299 (df = 791)	49.301 (df = 997)
F Statistic	45.429*** (df = 2; 2923)	29.585*** (df = 2; 1129)	6.582*** (df = 2; 791)	14.162*** (df = 2; 997)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Reporting rates in Money-No Key condition

	Dependent variable:			
	Response			
	UK, Poland, and US	United Kingdom	Poland	United States
	(1)	(2)	(3)	(4)
Money-NoKey	3.605***	3.691**	−10.875***	2.125
	(0.858)	(1.521)	(3.794)	(3.950)
Constant	53.484***	51.633***	70.875***	52.375***
	(1.489)	(2.528)	(1.904)	(1.767)
Observations	2,926	1,132	794	1,000
R^2	0.006	0.005	0.010	0.0003
Adjusted R^2	0.006	0.004	0.009	-0.001
Residual Std. Error	49.151 (df = 2924)	49.474 (df = 1130)	46.414 (df = 792)	49.964 (df = 998)
F Statistic	17.655*** (df = 1; 2924)	5.891** (df = 1; 1130)	8.215*** (df = 1; 792)	0.289 (df = 1; 998)

Note: *p<0.1; **p<0.05; ***p<0.01

2-4 show that this pattern holds for the UK, Poland, and the USA, but the difference was only statistically significant for the UK and Poland.

Here, we see that it is less likely for Money-NoKey wallets to be reported compared to Money condition wallets in Table 2. This may suggest that having a key present indicates greater loss than the small (USD$13.45) sum of money, as losing the key may cause the owner greater inconvenience.

5 Discussion

This paper seeks to replicate the results found in the study on civic honesty around the globe. Using their datasets, and limiting to results pertaining to reporting rates across the 4 treatment conditions (NoMoney, Money, BigMoney, and Money-NoKey), we attempt to validate their findings on civic honesty. Like Cohn et al. (2019), from the results, we found that citizens appear to display greater civic honesty when wallets contain money, the likelihood of contacting the wallet owner increasing when the wallet has larger amounts of money (Table 1, Table 2. This result holds when controlling for recipient and situational characteristics, such as gender, presence of coworkers, a computer, or bystanders, and age, as seen in Table 1. However, the effects of a coworker/coworkers being present, as well as being ages 40+ did not produce statistically significant results as found in the original paper. In addition, when the wallet contains money but no key, citizens are less likely to contact the owner compared to instances where there was money and a key (Table 3).

5.1 Limitations

There are several limitations associated with the study that must be addressed. The first relates to the survey data that was gathered. Cross-country surveys come with a number of limitations, e.g., comparisons of survey data may be biased due to cultural differences in interpretations of questions, how participants make use of response scales, and to what degree responses are influenced by social desirability concerns (Tannenbaum et al., (2020). It is challenging to draw clean comparisons through survey data alone with these limitations, not including the question of whether survey responses actually translate to meaningful behaviours despite cognitive biases and social desirability effects (Bertrand and Mullainathan, (2001). The limitations listed here raise the likelihood that surveys regarding social capital have little similarity with objective measures of social capital (Tannenbaum et al., (2020). In addition to survey limitations, it is possible that failure to return the wallet or contact the owner is not a dishonest act, but caused by the policies in place at the institutions the wallets were lost in, e.g., use of a lost and found as mentioned by Sulitzeanu-Kenan et al., (2020), or even fatigue.

5.2 Next Steps

Based on the above limitations, we have some suggestions for how the experiment can be extended and/or improved upon. To start, the researchers need to ensure that all institutions used in the study do not have a policy against contacting owners of lost items would allow for a firmer conclusion in whether non-returns are a result of dishonesty or altruistic behaviour. The current model leaves room for non-responses being codified as dishonest where they may not be. Second, in the initial study, additional representative surveys were given in the UK, Poland, and the US, where respondents were given detailed descriptions of wallet drop-off procedure, and asked how likely it was they'd contact the owner. Across the 3 countries, the average was 100%, substantially higher than the wallet return rates seen in the regression results and our replication. Social desirability in the face of them knowing their answers will be seen by someone differs greatly from receiving and/or finding a wallet and the task being left up to you with no indication of who the owner may be.

5.3 Ethics/Bias

40 countries, and 355 cities were used to represent the globe, with 400 observations per country. However, is it fair to say that the 40 countries chosen, or even those 5-8 largest cities per country that were selected, are representative of the world? If we look to the selection criteria for the countries/cities, as seen in the methodology section, countries without 5-8 cities with a minimum population of 100,000, countries that have strict customs, immigration, and banking regulations, and those deemed difficult to visit and/or unsane are omitted. As such, entire regions of the world are omitted, such as north and central Africa (not including

Middle Eastern nations), and the Pacific Islands, while nearly half of all nations included are within Europe, introducing a heavy Western culture bias while attributing results to the world. There is also failure to mention the average socioeconomic status of the areas in which the wallets are dropped off, which can be a confounding variable.

In addition to this, drop-off procedure for the wallets involved research assistants waking into a building, and approaching an employee at the counter to give them the counter and immediately leave the building without leaving contact details or asking for a receipt. Two factors come into mind: is it reasonable to treat the employee as the average person, and did their telling the employee that the wallet was found outside the building and around the corner, as opposed to inside the building, have an effect on the responses? In the event they were told that the wallet was found on the premises, it is possible that reporting rates could have increased due to added likelihood that someone would come searching for the wallet. Furthermore, all research assistants were pooled from two German speaking universities. Assuming they were German or Swiss nationals, it is incredibly likely that their English is accented, and that they are white, which may have had an effect on response rates compared with seeking out assistants within the target countries.

Appendix

References

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91 (2): 5. https://doi.org/10.1257/aer.91.2.67.

Champely, Stephane. 2020a. *Package "Pwr".* https://cran.r-project.org/web/packages/pwr/pwr.pdf.

———. 2020b. *Pwr: Basic Functions for Power Analysis.* https://CRAN.R-project.org/package=pwr.

Cohn, Alain, Michael Andre Marechal, David Tannenbaum, and Christian Lukas Zund. 2019. "Civic Honesty Around the Globe." *Science Mag* 365 (6448): 70. https://doi.org/10.1126/science.aau8712.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Han, Hu, Charles Otto, and Anil Jain. 2013. "Age Estimation from Face Images: Human Vs. Machine Performance." *Proceedings - 2013 International Conference on Biometrics, ICB 2013*, June. https://doi.org/10.1109/ICB.2013.6613022.

Harvard College. 2021. *Havard Dataverse Support.* https://support.dataverse.harvard.edu/.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Ledsom, Alex. 2020. *The 10 Safest Countries in the World to Visit.* https://www.forbes.com/sites/alexledsom/2020/11/21/the-10-safest-countries-in-the-world-to-visit/?sh=3cd42d911f74.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sulitzeanu-Kenan, Raanan, Omer Yair, and Markus Tepe. 2020. "Are Public Sector Employees More Honest? Field Evidence from Forty Countries." *SSRN*, 37.

Tannenbaum, David, Alain Cohn, Christian Lukas Zund, and Michel Andre Marechal. 2020. "What Do Cross-Country Surveys Tell Us About Social Capital?" https://doi.org/10.5167/uzh-188351.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Worldometer. 2021a. *7 Continents.* https://www.worldometers.info/geography/7-continents/.

———. 2021b. *Countries in Europe.* https://www.worldometers.info/geography/how-many-countries-in-europe/.

World Population Review. 2021. *World City Populations 2021.* https://worldpopulationreview.com/world-cities.