

Final Project — II. Data Analysis Replication

For the data analysis replication, you can choose one papers attached below for replication tasks by the following rule. Please select a paper from group A if your student number is an even number (ends with 0, 2, 4, 6, 8), and select a paper from group B if your student number is an odd number (ends with 1, 3, 5, 7, 9). You can download the detailed instruction, questions, dataset for each paper from the corresponding folder from the following links:

https://drive.google.com/drive/folders/1Vr-FvxKUZg9wsU9qiejepJhGhnO9Nh?usp=share_link

or

https://pan.baidu.com/s/14P_szVsUMJqowkhGYvDzqQ?pwd=z3gj

提取码: z3gj

For each paper, I prepared five questions for you, each worth 20 points. You should answer all five and prepare the related R-script. For the third papers in both Groups, I prepared an extra question for each, which is related to the more advanced IV estimation. This question is optional, you can get an extra 20 points if you answer it correctly. Please prepare your answers and r-script for the paper you have chosen and submit it to moodle by 18:30pm 25 December, 2022.

Group A

1. Gerber, Alan S., Dean Karlan, and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics*, 1 (2): 35-52. See the pertinent instrument on pp.2-4. **OLS**
2. Abramitzky, Ran, Adeline Delavande, and Luis Vasconcelos. 2011. "Marrying Up: The Role of Sex Ratio in Assortative Matching." *American Economic Journal: Applied Economics*, 3 (3): 124-57. See the pertinent instrument on pp.5-6. **OLS** **DID**
3. Compulsory Licensing: Evidence from the Trading with the Enemy Act
See the pertinent instrument on pp.7-8. **OLS** **DID** **IV*(optional)**

Group B

1. Rose, Andrew, K. 2004. "Do We Really Know That the WTO Increases Trade?" *American Economic Review*, 94 (1): 98-114. See the pertinent instrument on pp.9-11. **OLS**
2. Zhang, Xiaoquan (Michael), and Feng Zhu. 2011. "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia." *American Economic Review*, 101 (4): 1601-15. See the pertinent instrument on pp.12-13. **OLS** **DID**

3. Imai, Masami, and Seitaro Takarabe. 2011. "Bank Integration and Transmission of Financial Shocks: Evidence from Japan." *American Economic Journal: Macroeconomics*, 3 (1): 155-83.
See the pertinent instrument on pp.14-15. OLS DID IV*(optional)

Group A

1. Gerber, Alan S., Dean Karlan, and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics*, 1 (2): 35-52.

In this paper, authors try to study whether the exposure to newspaper affect one's political knowledge and stated opinions, or increases voting turnover, and affect a voter's support to a Democrat (more liberal) or Republican (conservative) candidate. Using a randomized experiment, in which authors assigned individuals to a free subscription of either the liberal Washington Post or the more conservative Washington Times newspaper (two major newspaper in Washington, DC.), authors study these questions using a sample of residents in Prince William County VA, a county 25 miles outside of Washington, DC. In term of the econometric method, this paper is relatively easy. Since it is a randomized experiment, many of the omitted variable bias has already been taken care of by design. Nevertheless, authors still add a lot of controls variables so as to get a more robust estimation.

You can find the variable list and definition to each variable below:

| Variable Name | Variable Definition |
|------------------|--|
| treatment | TREATMENT |
| wave2 | Newspaper subscription processed in 2nd wave |
| completematch | Successfully matched to state voting data |
| Bgetsmag | Subscribes to any news or political magazine |
| post | Received Post treatment |
| times | Received Times treatment |
| surveyed | Surveyed in post-election survey |
| Bfemale | Female |
| reportedage | Age of respondent |
| Bvoted2004 | Self-report, voted in 2004 general election |
| Bvoted2002 | Self-report, voted in 2002 general election |
| Bvoted2001 | Self-report, voted in 2001 general election |
| Bconsumer | Respondent drawn from consumer database sample frame |
| Bpreferrepub | Prefers the Republican candidate for upcoming VA gubernatorial election |
| Bpreferdem | Prefers the Democratic candidate for upcoming VA gubernatorial election |
| Bprefernoone | No stated preference for upcoming VA gubernatorial election |
| paper | Received either Post or Times treatment |
| voted | Self-report voted in 2005 VA Gubernatorial election |
| voteddem | Voted for Democrat in 2005 Gub election, set to missing if did not vote |
| voteddem_all | Voted for Democrat in 2005 Gub election, set to 0 if did not vote |
| preferdem | Did not vote in 2005 Gub election, but preferred Democrat |
| voteddem2 | Voted for or preferred Democrat in 2005 Gub election |
| mostimp_scandals | Most important problem facing country (1=issue other than scandals, 0=scandal) |

| | |
|----------------|--|
| iraq_post | Most important issues in Iraq (1=constitution or Hussein trial, 0=other) |
| bushapproval | Bush approval (4=strong approval, 1=strong disapproval) |
| repfavorable | Overall opinion of Republican Party favorable (0=v unfavorable, 3=v favorable) |
| demunfavorable | Overall opinion of Democratic Party unfavorable (1=v favorable, 4=v unfavorable) |
| iraq | Progress in Iraq (0=very badly, 3=very well) |
| leak | Leak case (3=no one did anything wrong, 1=something illegal) |
| alito | Alito confirmation (3=should confirm, 1=should not confirm) |
| iraqdead | Knew number of dead in Iraq |
| libby | Identified Libby as involved in leak |
| miers | Identified Miers as Supreme Court nominee |
| getapaper | Reports receiving a paper |
| gettimes | Reports receiving Times |
| getpost | Reports receiving Post |
| readfrqr | Frequency respondent reads paper (0=Never, 3=Every day) |
| readsome | Reads paper at least several times per week |
| conservative | Self-identified Conservatism (7=extreme conservative, 1=extreme liberal) |
| consindexgen | Broad policy index (higher scores conservative) |
| consindexpol | Specific Issue Index (mostimp_scandals, iraq_post, iraq, leak, alito) |
| factindex | Fact index (iraqdead, libby, meirs) |
| cells | Strata indicator |
| dateoperator | Surveyor/date indicator |
| voted2006g | Voted 2006 - Administrative state voting data |
| voted2005g | Voted 2005 - Administrative state voting data |
| voted2004g | Voted 2004 - Administrative state voting data |
| voted2003g | Voted 2003 - Administrative state voting data |
| voted2002g | Voted 2002 - Administrative state voting data |
| voted2001g | Voted 2001 - Administrative state voting data |
| voted2000g | Voted 2000 - Administrative state voting data |
| treatmentn | Experimental treatment (control,post or times) - numeric |
| MBfemale | Female, missing |
| Mreportedage | Age of respondent, missing |
| MBvoted2004 | Self-report voted in 2004 general election, missing |
| MBvoted2002 | Self-report voted in 2002 general election, missing |
| MBvoted2001 | Self-report voted in 2001 general election, missing |
| MBconsumer | Respondent drawn from consumer database sample frame, missing |
| MBgetsmag | Subscribes to any news or political magazine, missing |
| MBpreferdem | Prefers the Democratic candidate for gubernatorial election, missing |
| MBpreferrepub | Prefers the Republican candidate for gubernatorial election, missing |
| MBprefernoone | No stated preference for gubernatorial election, missing |
| Mwave2 | wave2, missing |
| MZBfemale | Female, missing coded as zero |
| MZreportedage | Age of respondent, missing coded as zero |
| MZBvoted2004 | Self-report voted in 2004 general election, missing coded as zero |
| MZBvoted2002 | Self-report voted in 2002 general election, missing coded as zero |
| MZBvoted2001 | Self-report voted in 2001 general election, missing coded as zero |
| MZBconsumer | Respondent drawn from consumer database sample frame, missing coded as zero |
| MZBgetsmag | Subscribes to any news or pol magazine, missing coded as zero |
| MZBpreferdem | Prefers the Democratic candidate for gubernatorial election, missing coded as zero |
| MZBpreferrepub | Prefers the Republican candidate for gubernatorial election, missing coded as zero |

MZBprefernoone
MZwave2

No stated preference for gubernatorial election, missing coded as zero
wave2, missing coded as zero

Question 1) To replicate the empirical result in Table 3 about the effect of the newspapers on political knowledge and attitudes and Table 4 on voting behavior.

Hint: a) Control variables in each regression include: gender; reported age; three separate indicators for voting in the 2001, 2002, and 2004 general elections; an indicator for whether the respondent was drawn from a consumer list; self-reports of receiving any news or political magazines; baseline survey self-reports of preferring the Democratic candidate in the gubernatorial election and having no preference in the gubernatorial election; and an indicator for the wave of the study. In addition, authors also included the fixed effects of strata indicators and surveyor/date indicators. b) for Panel A of Tables 3 and 4, after running the regression, conduct an F test to test if the coefficient of “post” (Received Post treatment) is the same as that of the variable “time” (Received Times treatment).

Question 2) Now write two paragraphs to describe your replication results of table 3 and 4, please use your own language.

Question 3) suppose we are interested in finding out if the treatment effect of receiving Washington Post on Voted for Democrat (column (4) Panel A in Table 4) is different between male or female, design a new regression to test if it is true. Show your results in a new table (like the authors, you can conceal other results and only report the results on treatments, gender and their interactions).

Question 4) look at your replication result of column (1) in Table 4, whether if age has a significant effect on whether a respondent voted in 2005 election? Now suppose we are interested in finding out whether age has non-linear effect on one’s probability to voted in 2005, design a new regression to test if it is true. Show your results in a new table (like the authors, you can conceal other results and only report the results on age).

Question 5) In table 4, authors applied linear probability model for each regression with dummy variable as dependent variable. Now we want to use logit regression to re-run the regressions in Table 4 and report your results in a new table.

2. Abramitzky, Ran, Adeline Delavande, and Luis Vasconcelos. 2011. "Marrying Up: The Role of Sex Ratio in Assortative Matching." *American Economic Journal: Applied Economics*, 3 (3): 124-57.

Marital assortative matching refers to a phenomenon that males and females from the same social class (defined by family background, education and income level etc.) are more likely to marry each other (in Chinese we usually call it “門當戶對”). In this paper, authors are interested in finding out whether the relative population of one gender contributes to marital assortative matching. Authors answer this question by a very smart differences-in-differences design. In French, a lot of young males lost their life during WWI. Therefore, after the war, those regions with a higher wartime mortality rate in the army, authors hypothesize men were more likely to marry women with higher social classes (an indicator of marital assortative matching). The key differences-in-differences variable thus is the interaction term between the dummy variable of post-war period (“post”) and the military mortality rate for each region or department in French (“mortality”). The interaction term is titled “post_mortality” in the data.

You can find the variable list and definition to each variable below:

| Variable Name | Variable Definition |
|----------------|---|
| depc | departement code |
| year | year |
| post | post war |
| clgr | groom's class |
| clbr | bride's class |
| clfgr | groom's father's class |
| clfbr | bride's father's class |
| clmgr | groom's mother's class |
| clmbr | bride's mother's class |
| agegr | groom's age |
| agebr | bride's age |
| rural | rural |
| city | city size |
| secmargr | groom is remarrying |
| secmarbr | bride is remarrying |
| mortality | military mortality rate |
| post_mortality | post war * military mortality rate |
| sr | sex ratio (#males 18-59/#females 15-49) |
| sr_39 | sex ratio (#males/#females, ages 15-39) |
| sr_49 | sex ratio (#males/#females, ages 15-49) |
| fdgr | groom's father is dead |
| mdgr | groom's mother is dead |
| fdbr | bride's father is dead |
| mdbr | bride's mother is dead |
| classdiff | class difference |
| mardn | married down |
| lowbr | low class bride |
| agebrd | bride's age(/100) |
| agegrd | bride's age(/100) |

Question 1) To replicate the empirical result in Table 3 in the paper.

Hint: a) Control variables in each regression include: Rural, bride's age (/100), groom's age (/100), the fixed effects of groom class and department.

Question 2) Now write a paragraph to describe your replication results of table 3, do the results confirm author's hypothesis? Please use your own language to describe.

Question 3) For typical differences-in-differences design using the interaction term, authors should also control for the main effects of both "post" and "mortality" variables. Why did the authors exclude the "mortality" variable in each regression in Table 3? Please answer.

Question 4) Now let us confirm if the parallel time trend assumption of differences-in-differences can be satisfied in authors' setting. Now divide all the observations by high or low mortality rate (>15.037 [the mean of mortality] or ≤ 15.037), and then calculate the proportion of individuals that were married down by year for the high and low mortality groups. The last step is to draw a line plot on the pertinent proportions by year with two lines corresponding to high and low mortality groups. Are these two lines parallel to each other pre-war (before 1918)?

Question 4) Now suppose you want to once for all take care of the parallel time trend assumption, how can you modify the regression to solve it? (Hint: utilize the department dummies and year variables)

Question 5) In table 3, authors applied linear probability model for each regression with dummy variable as dependent variable. Now we want to use probit regression to re-run the regressions in Table 3 and report your result in a new table.

3. Compulsory Licensing: Evidence from the Trading with the Enemy Act

Compulsory licensing allows firms in developing countries to produce foreign-owned inventions without the consent of foreign patent owners. Have you watched the movie “dying to survive” (我不是藥神) two years ago? The movie is somewhat related to this topic. Every year countries such as Brazil, Thailand, and India have used this policy to procure life-saving drugs for millions of patients, who otherwise cannot pay for the original drugs produced by the patent-holders. However, in the policy realm, the compulsory licensing is a very controversial policy, some opponents to the policy raise concerns that compulsory licensing may discourage invention by the patent holder firms as well since most of the products produced under compulsory licensing are sold under market price. In this paper, authors analyze the welfare effect of this policy from a new perspective: whether compulsory licensing can also increase or discourage domestic invention. After World War I, the US Congress passed the “Trading with the Enemy Act”, allowing US firms to violate enemy-owned patents if they contributed to the war effort. After some amendment, by February 1919, all German-owned patents were systematically licensed to US firms, a de facto compulsory licensing to all those patents owned by German in other words. Authors try to test if the effect of this policy treatment will increase or decrease the invention in the same subclass of those German-owned patents, a typically differences-in-differences research design. The treatment variable “treat” in the data is the interaction term of the dummy variable indicating the pertinent subclass has at least one license (the “licensed_class” variable) and the dummy variable of post-1919 (the “post-1919 dummy”).

You can find the variable list and definition to each variable below:

| Variable Name | Variable Definition |
|-------------------|--|
| uspto_class | subclass of patent (text) |
| grntyr | year of patent granted |
| count_usa | number of patents by the US inventors |
| count_france | number of patents by the France inventors |
| count_germany | number of patents by the Germany inventors |
| count | number of patents |
| count_for | number of patents by foreign inventors |
| count_for_2 | number of patents by foreign inventors squared |
| count_noger | number of patents by the non-Germany inventors |
| count_for_noger | number of patents by the non-Germany foreign inventors |
| main | first half of subclass id (text) |
| subcl | second half of subclass id (text) |
| year_conf | remaining lifetime of licensed patents |
| year_conf_2 | remaining lifetime of licensed patents squared (*100) |
| count_cl | number of licenses |
| count_cl_2 | number of licenses squared |
| licensed_class | subclass has at least one Germany license |
| confiscated_class | number of confiscated patents in class |
| class_id | subclass of patent |
| treat | (subclass has at least one Germany license)*post-1919 |
| year_conf_itt | remaining lifetime of enemy patents |
| count_cl_itt | number of enemy patents |
| post | post-1919 dummy |

Question 1) To replicate the empirical result in Table 2 in the paper.

Hint: a) Control variables in each regression include: the fixed effects of the year (of the patent granted) and subclass.

Question 2) Now write a paragraph to describe your replication results of table 2, do the results confirm the author's hypothesis? Please use your own language to describe.

Question 3) Now let us confirm if the parallel time trend assumption of differences-in-differences can be satisfied in the authors' setting. Now draw a line plot on the average number of patents by the US inventors by year with two lines corresponding to treatment and control groups ("licensed_class"=1 for the treatment group, and "licensed_class"=0 for the control group). Are these two lines parallel to each other pre-war (before 1918)?

Question 4) Now suppose you want to once for all take care of the parallel time trend assumption, how can you modify the regression to solve it? (Hint: utilize the subclass dummies and year variables)

Question 5) Given the "Trading with the Enemy Act" gives licenses for all German-owned patents to the US firms, it may discourage patent application for Germany firms in the US patent office. Check if this is true to replace the dependent variable in Table 2 using the numbers of patents by Germany inventors (the "count_germany" variable). Report your result in a new table and describe it.

*Question 6) (optional) Now replicate the regressions using an instrumental variable approach in Table 4.

For this optional question, you can get an extra 20% of scores on top of the full mark for the first five questions.

Group B

1. Rose, Andrew, K. 2004. "Do We Really Know That the WTO Increases Trade?" *American Economic Review*, 94 (1): 98-114.

The World Trade Organization (WTO) may be one of the most important international organization aiming to promote free trade. However, does the membership to WTO really promote trade? The author tries to answer this question using a large panel data of bilateral merchandise trade between 175 countries over 50 years. In addition to the WTO, the author also examines whether its predecessor the General Agreement on Tariffs and Trade (GATT) has a positive effect on promoting trade between two countries.

You can find the variable list and definition to each variable below:

| Variable Name | Variable Definition |
|---------------|--|
| cty1 | IFS Country Code 1 |
| cty2 | IFS Country Code 2 |
| year | Year |
| pairid | Unique Country-Pair Identifier |
| landl | # Landlocked 0/1/2 |
| island | # Islands 0/1/2 |
| border | Land Border Dummy |
| comlang | 1 for Common Language |
| comcol | Dummy for Common Colonizer post 1945 |
| comctry | Dummy for Same Nation/Perennial Colonies |
| colony | Dummy for pairs ever in Colonial Relationship |
| curcol | Dummy for pairs currently in Colonial Relationship |
| custrict | Strict Currency Union |
| ltrade | Log Value of Bilateral Trade in Real \$ |
| avertltrade | Mean of Log value of bilateral trade in each year |
| above_average | A dummy of trade value above yearly average |
| regional | RTA Dummy |
| lareap | Log of Product of Land Areas |
| ldist | Log of Distance |
| lrgdp | Log of Product of Real GDPs |
| lrgdppc | Log of Product of Real GDPs per capita |
| rta | RTA: EU(1) usis(2) nfta(3) crcm(4) pter(5) anzd(6) ccm(7) mrc(8) |
| sasia1 | South Asian Dummy |
| ssafr1 | Sub-Saharan Africa Dummy |
| asia1 | East Asian Dummy |
| high1 | High Income Dummy |
| latca1 | Latin-Caribbean Dummy |
| least1 | Least Developed Dummy |
| lowin1 | Low Income Dummy |
| menaf1 | Middle East and North African Dummy |
| midin1 | Middle Income Dummy |
| sasia2 | South Asian Dummy |
| ssafr2 | Sub-Saharan Africa Dummy |
| asia2 | East Asian Dummy |
| high2 | High Income Dummy |

| | |
|----------|--|
| latca2 | Latin-Caribbean Dummy |
| least2 | Least Developed Dummy |
| lowin2 | Low Income Dummy |
| mena2 | Middle East and North African Dummy |
| midin2 | Middle Income Dummy |
| carib1 | Caribbean dummy |
| carib2 | Caribbean dummy |
| join1 | Date of GATT/WTO accession (-999 2002 observer; 10000 never) |
| join2 | Date of GATT/WTO accession (-999 2002 observer; 10000 never) |
| onein | One country in GATT/WTO |
| bothin | Both countries in GATT/WTO |
| nonein | Neither country in GATT/WTO |
| found1 | One founding GATT member |
| found2 | Both founding GATT members |
| years1 | Years inside GATT/WTO |
| years2 | Years inside GATT/WTO |
| minyrs | Minimum years in GATT/WTO |
| maxyrs | Maximum years in GATT/WTO |
| gsp | GSP Dummy |
| ecd | Dummy for EC/EU |
| usi | Dummy for USIS |
| naf | Dummy for NAFTA |
| car | Dummy for CARICOM |
| pat | Dummy for PATCRA |
| anz | Dummy for ANZD |
| cac | Dummy for CACM |
| mer | Dummy for MERCOSUR |
| ase | Dummy for ASEAN |
| spr | Dummy for SPARTECA |
| cty1name | Name for Country 1 in the pair |
| cty2name | Name for Country 2 in the pair |

Question 1) To replicate the empirical result in Table 1 about the effect of the GATT/WTO membership on bilateral trade between two countries.

Hint: a) Apart from the control variables reports in each column, all four regressions in Table 1 also add the year fixed effects. b) all regressions use robust standard errors and clustering by county pairs (the “pairid” variable). c) for column (2) to exclude all the industrial countries, we need to exclude those observations with IFS country code less than 200 (cty1<200 & cty2<200), see footnote 16 on p. 104 for detail. D) for column (4) to add the fixed effects for country 1 and country 2 (using the variables “cty1” and “cty2”).

Question 2) Now write a paragraph to describe your replication results of table 1. Does joining the WTO help a country’s trade with others? Please use your own language to illustrate.

Question 3) suppose we are interested in a test that if the treatment effect of GATT/WTO membership is significant for country pairs speaking a common language (the “comlang” variable), but not for country pairs speaking different languages. Design a new regression to test if it is true. Show your results in a new table.

Question 4) The data used in this paper is a panel data of country pairs over time. In other words, the cross-sectional variation is at country pair level while time variation at the year level. So we can use a two-way fixed effects model to estimate the effect of GATT/WTO membership, controlling for both the fixed effects of country pair (using the “pairid” variable) and the fixed effects of the year (now the model only use one-way fixed effects. Now apply the two-way fixed effects to the model. Show your results in a new table and compare it with those in original Table 1. Explain what kind of change it brings by adding the country pair fixed effects. Also, do you find some variables automatically drop in the estimation? If yes, which variables and why?

Question 5) Now suppose we concern more about the relative size of trade instead of the absolute size. We generate a new dummy variable “above average” which equals to 1 if the log value of bilateral trade between country i and j is larger than the yearly average in year t (if $\ln \text{trade}_{ijt} \geq \text{averltrade}_t$) and otherwise 0 (if $\ln \text{trade}_{ijt} < \text{averltrade}_t$). Now use the same sample, control variables, and specification as column (1) in Table 1, replace the dependent variable “above average” as a dependent variable to run the regression. Use both linear probability model, Probit and Logit model for the new regressions. Show your results in a new table.

2. Zhang, Xiaoquan (Michael), and Feng Zhu. 2011. "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia." *American Economic Review*, 101 (4): 1601-15.

Today many public goods on the Internet are provided to users for free, many of which rely entirely on free user contributions. Taking Wikipedia as an example, as a free online encyclopedia, Wikipedia relies on volunteer contributors around the world to create and edit content. It is an interesting question, then whether the size of the user of a platform could change the incentive for users to contribute. Authors utilize the shock event of the block of Chinese Wikipedia in mainland China in October 2005 as a natural experiment to test if the content contribution decrease as a result of the block event. During the block, mainland Chinese could not use or contribute to Chinese Wikipedia, while contributors outside mainland China can still use and contribute; naturally, this causes a dramatic decrease of users of the platform. Authors then test if the contribution levels of the nonblocked contributors also decrease within several weeks of the event.

You can find the variable list and definition to each variable below:

| Variable Name | Variable Definition |
|----------------------|---|
| date | calendar date |
| id | Registered contributor ID |
| Addition | Total number of characters added |
| Deletion | Total number of characters deleted |
| Total | Total number of characters added and deleted |
| joindate | Date of joining Wiki for the contributor |
| lastdate | Date of the last editing of the contributor |
| nonblocked | Dummy of nonblocked contributor |
| overseas | Dummy of oversea IP |
| week | week before/since the block event |
| id_week | week before/since the block event (text) |
| weekly_Addition | Weekly total number of characters added |
| weekly_Deletion | Weekly total number of characters deleted |
| age | Age |
| agesqr | age squared |
| logAddition | Log of (weekly total number of characters added + 1) |
| logDeletion | Log of (weekly total number of characters deleted + 1) |
| logTotal | Log of (weekly total number of characters added and deleted +1) |
| after | AfterBlock |
| social_participation | Log of (weekly average of total addition and deletion in user pages or user-talk pages before the block +1) |
| ifTotal | Dummy to indicate if the weekly total number of character add and delete is larger than zero |

Question 1) To replicate the empirical result in Table 2 in the paper.

Hint: a) columns (1) to (3) use OLS estimate with only AfterBlock, Age and Age Squared as a regressor, while columns (4) to (6) add contributor's fixed effects. b) all regressions use robust standard errors. c) restrict the observations to 4 weeks prior and after the block event and exclude the week zero (week ≥ -4 & week ≤ 4 & week $\neq 0$).

Question 2) To further examine the impact of social effects, authors use a differences-in-differences design, in which they further distinguish contributors' participation in social communication. They use the contributors' weekly average contribution to the user pages and user-talk pages—both of which are designed to facilitate communication between contributor in Wikipedia—prior to the block event to construct a measure of average social participation of each contributor. Then authors use the interaction term between the AfterBlock dummy and the measure of social participation pre-block as the differences-in-differences variable in Table 3. Now replicate the results in Table 3 and write a paragraph to describe your replication results of table 3. Do the results suppose the authors' hypothesis that group size matter to the user's free contributions? Please use your own language to illustrate.

Question 3) For an alternative differences-in-differences design using the interaction term, authors can also use the fixed effects of the contributor and the fixed effects of the week to replace the AfterBlock dummy and SocialParticipation variable. Now try to add the fixed effects of the week to re-run the results in columns (4) to (6). Show your results in a new table. How do the results change? Do you prefer to keep the AfterBlock dummy? Why?

Question 4) Now suppose we have a hypothesis that the treatment effect of the block event is driven by the contributor from oversea IP. Now use the triple interaction terms between the “oversea” dummy variable and our differences-in-differences variable using the specification in columns (4) to (6) in Table 3. Show your results in a new table. What do the results tell us about this hypothesis?

Question 5) Now suppose we concern more about whether the non-block contributors contribute after the event instead of how much they contribute. We generate a new dummy variable “ifTotal” which equals to 1 if the weekly total number of character add and delete is larger than zero for contributor i in week t (if $\text{weekly_Addition} + \text{weekly_Deletion} > 0$) and otherwise 0 (if $\text{weekly_Addition} + \text{weekly_Deletion} = 0$). Now use the same sample, control variables, and specification as column (4) in Table 3, replace the dependent variable “ifTotal” as a dependent variable to run the regression. Use both linear probability model, Probit and Logit model for the new regressions. Show your results in a new table.

3. Imai, Masami, and Seitaro Takarabe. 2011. "Bank Integration and Transmission of Financial Shocks: Evidence from Japan." *American Economic Journal: Macroeconomics*, 3 (1): 155-83.

In a major financial crisis, the large nationwide banks were often the troublemakers, take the Lehman Brothers in 2008 Financial Crisis as an example. On the one hand, some believe the geographically diversified, big banks are more resilient to crisis or in other words, are "too big to fall" compared to the regional, small-size banks. On the other hand, some argue these nationwide operated big banks often play the role of an active transmitter of the bad debt during the crisis. To investigate whether the banking integration (a jargon to describe to what extent the banks operate in a region are more geographically diversified big banks while in contrast, the banks operate in a region are predominantly small-sized regional banks) has a significant effect in transmitting the financial shocks across regions during the financial crisis, the authors use the collapse of the real estate bubble in 80s Japan as a case study. Specifically, authors examine whether the effect of land price drop (an indicator of bubble bust) has a larger effect on debt and GDP in the prefectures with a more integrated banking system.

In Japan, there are roughly two types of bank, the big, city banks which have bank branches in many different prefectures, and the small, regional banks that often operate only in one prefecture. Thus, the authors use the proportion of city bank branches to total bank branches in a prefecture as a measure of bank integration in a given prefecture.

You can find the variable list and definition to each variable below:

| Variable Name | Variable Definition |
|----------------------|---|
| name | Prefecture Name |
| prefid | Prefecture ID |
| year | Year |
| GDP | GDP Growth |
| averGDP | Average GDP growth rate at each year |
| above_average | A dummy of GDP growth rate above yearly average |
| LOAN | Loan Growth |
| gloan_city2 | LOAN (City Bank) |
| gland | Local Land Price |
| cityshare1 | City Bank Share |
| share_gland_city | City Land Price*City Bank Share |
| shareo_gland_city | City Land Price*City Bank Share (1979) |
| dist_closestcity | Distance to Nearest Major City |
| dist_gland_city | Distance*City Land Price |
| diff_gpc_gland_city | Income Diff*City Land Price |
| citybank_diff_city | Income Diff*City Bank Share |
| citybank_dist_city | Distance*City Bank Share |
| City_industry_shock | Industry Mix Control (Output) |
| gloan_noncity2 | LOAN (Non-City Bank) |
| citysharedep | City Bank Share |
| sharedep_gland_city | City Land Price*City Bank Share |
| sharedepo_gland_city | City Land Price*City Bank Share |
| cit~p_dist_city | Distance*City Bank Share |
| cit~p_diff_city | Income Diff*City Bank Share |
| mean_diff_gpc_city | Income Difference |

Question 1) To replicate the OLS results (columns (1) to (4) in Panel A) in Table 1 in the paper.

Hint: a) each regression controls for the two-way fixed effects of the prefecture (denoting the “prefid” variable) and year. b) use the robust standard errors for all regression.

Question 2) Now write a paragraph to describe your replication results of table 1, do the results confirm the author’s hypothesis? Please use your own language to describe.

Question 3) Suppose we now concern that the variable of city bank share may be endogenous to GDP growth rate, in other words, after the real estate bubble bust, some banks had gone bankrupt which may change the city bank share in a prefecture, a reverse causality problem. Therefore, instead of using city bank share in each prefecture i at each year t , we use the city bank share in each prefecture i at 1979 prior to the crisis to construct the interaction term. Now re-run the columns (2) and (3) in Table 1, Panel A and report your results in a new table. Now compare the coefficients in original columns (2) and (3) with those in the new regressions, what do the results tell you about the reverse causality? Do you think the new measures give you more robust results?

Question 4) Now use the specification in column (3) of Panel A, Table 1 to generate three new regressions results: first one to remove the fixed effects of year, the second one to remove the prefecture fixed effects, and then third to remove both fixed effects. Show your results in a new table. Now compare the coefficients of the interaction term in each of the new regression with the result of the original column (3). What do the differences tell you? How does it bias our results to remove each or both of the fixed effects?

Question 5) Now suppose we concern more about the relative growth rate of GDP instead of the actual growth rate. We generate a new dummy variable “above average” which equals to 1 if the GDP growth i is larger than the yearly average of GDP growth rate among prefectures in year t (if $GDP \geq \text{averGDP}$) and otherwise 0 (if $GDP < \text{averGDP}$). Now use the same sample, control variables, and specification as column (3) in Table 1, replace the dependent variable “above average” as a dependent variable to run the regression. Use both linear probability model, Probit and Logit model for the new regressions. Show your results in a new table.

*Question 6) (optional) Now replicate the regressions using an instrumental variable approach in columns (5) and (6) in Panel A and Panel B of Table 1.

For this optional question, you can get an extra 20% of scores on top of the full mark for the first five questions.