The cost of illiteracy

A causal inference study on the impacts of illiteracy on physical and mental health using the Chinese Longitudinal Healthy Longevity Survey

Qiyang Zhang*

Graduate School of Education, University of Pennsylvania

Abstract

This paper uses the Chinese Longitudinal Healthy Longevity Survey (CLHLS) to investigate the causal relationship between illiteracy and health. Matching methodologies control for confounding variables including age, sex, living sites, access to tap water, and financial support. Through nearest neighbor matching, mahalanobis metric matching, and propensity score matching, illiteracy is found to have significant impacts on physical health, exercising habits, anxiousness, loneliness, and happiness. Matching suggests that, on average, illiteracy decreases physical health by 19.9%, decreases exercising habits by 7%, increases anxiousness by 11.56%, increases loneliness by 17.6%, and decreases happiness by 11.3%. The paper's results confirm external validity of similar findings in the U.S context. Additionally, this paper uniquely finds that illiteracy has a higher cost on mental health as compared to physical health for the elderly in China.

Keywords: causal inference – matching – education and health – China – the elderly

1 Introduction

Past literature in Econometrics has established a positive association between education and health using both theoretical and empirical evidence. In human capital theories, investment in education yields long-term returns in health (Grossman, 1972). Higher income levels, more comprehensive skill sets, and accumulated social capital are mediators between health and education (*Social epidemiology*, 2014). The illiterate, representing society's lowest educated population, are predicted to have lower health status in their senior years. Apart from the above mentioned reasons, the illiterate also lack health literacy, which is defined as the personal characteristics and social resources needed for people to access, understand and use information to make decisions about their health.

China's population census shows that there were around 55 million illiterate Chinese individuals in 2010 (National Bureau of Statistics of China, 2010). The number of Chinese illiterates is similar to the population of Thailand, Turkey, Egypt, or Iran. While there are some causal inference studies between illiteracy and health in the US and the UK, there is virtually no such literature studying China. This paper serves to bridge this gap in the literature and provide causal evidence between education and health.

2 Data

2.1 Data Source

This paper uses longitudinal observational data from the Chinese Longitudinal Healthy Longevity Survey

(CLHLS) Parent-Child Dyads, which is composed of two waves conducted in 2002 and 2005. CLHLS mainly investigated physical health, mental health, and quality of life of the elderly aged at least 65 years old in 22 provinces of China. CLHLS provides information on the aged population's socioeconomic characteristics, family, lifestyle, demographic profile, respondents' health conditions, daily functioning, self-perceptions of health status, quality of life, life satisfaction, mental attitude, and feelings about aging. This rich data set has 4240 rows and 2957 columns. The second wave followed up with the same sample group from the first wave and contained detailed answers about the elderly's socioeconomic, demographic, and other background variables. This paper uses cross-sectional data from 2002 to conduct most analyses, with the exception of the outcome variable of health deterioration.

Code	Answer	Frequency	%
1	illitearte and semiliterate	1041	24.6%
0	others	3199	75.4%
	total	**4240**	**100.0%**

Table 1: Literacy Indicator

Code	Answer	Frequency	%
2	very good	504	11.9%
1	good	1437	33.9%
0	so so	1561	36.8%
-1	bad	509	12.0%
-2	very bad	42	1.0%
-2	not able to answer	187	4.4%
	total	**4240**	**100.0%**

Table 2: Self-reported Health Indicator

2.2 Variables

In total, 4240 senior Chinese individuals participated in face-to-face interviews in this survey. The 2002 wave has almost no missing data whereas the 2005 wave lost contact with 12.5% and suffered death of 30.0% of the original participants. The independent x is a variable representing illiteracy. CLHLS participants were asked "what kind of educational degree did you gain finally?" — "illiterate or semiliterate," "elementary school," "junior high school," "senior high school," "technical secondary school," "junior college," "undergraduate," "graduate," or "Ph.D." Among the 5% illiterate Chinese, the distribution of illiterate groups tilted towards the older population. The Chinese population above 65 years old has survived the Great Leap Forward, Henan Famine, and the infamous Cultural Revolution. During those difficult times, obtaining higher education was almost a "luxury" and illiteracy is not uncommon among this age group. This is proven by CLHLS data, in which almost 25% of the sample are illiterate. As shown in table 1, the illiterate treatment group is coded as 1 while the literate control group is coded as 0. This division is appropriate because matching method requires a much larger control group than treatment group to ensure efficient matching. In order to have a more objective health indicator, physical health is represented by a combination of self reported health and interviewer-rated health. CLHLS participants were asked "how do you rate your own health?" — "very good," "good," "so so," "bad," "very bad," and "not able to answer." As shown in table 2, the self-reported health variable is coded from a range of -2 to 2. CLHLS interviewers were asked to rate the health status of interviewees as "surprisingly good," "relatively good," "moderately ill," "very ill," "missing." As shown in Table 3, the interviewer-rated health variable is also coded from a range of -2 to 2. The mean of interviewer-rated health is 1.076 while the mean of self-reported health is 0.4665. Another indicator called health deterioration depicts changes in health by comparing the difference between health status in 2005 and health status in 2002.

In addition to physical health, illiteracy is also negatively associated with mental health. With decreas-

Code	Answer	Frequency	%
2	surprisingly healthy	1390	32.8%
1	relatively healthy	2254	53.2%
-1	moderately ill	520	12.3%
-2	very ill	74	1.7%
NA	missing	2	0.0%
	total	**4240**	**100.0%**

Table 3: Interviewer-rated Health Indicator

Code	Answer	Frequency	%
4	always	35	0.8%
3	often	140	3.3%
2	sometimes	983	23.2%
1	seldom	1690	39.9%
0	never	1145	27.0%
NA	not able to answer	247	5.8%
	total	**4240**	**100.0%**

Table 4: Anxiousness Indicator

Code	Answer	Frequency	%
4	always	55	1.3%
3	often	213	5.0%
2	sometimes	1094	25.8%
1	seldom	1432	33.8%
0	never	1211	28.6%
NA	not able to answer	235	5.5%
	total	**4240**	**100.0%**

Table 5: Loneliness Indicator

ing illiteracy rates in modern China, literacy is an assumption taken for granted in most social activities while illiteracy is almost equivalent to a disability. For the illiterate elderly, the difficulties illiteracy causes are frequently observed in daily grocery shopping, signing mailed packages, ordering restaurant food, or even finding washrooms if no restroom graphics are displayed. Unpleasant encounters due to illiteracy cause huge mental pressure and frustrations. The CLHLS participants were surveyed on "the frequency of feeling anxious, lonely, useless with age, and as happy as younger" — "always," "often," "sometimes," "seldom," "never," and "not able to answer". As shown in Tables 3, 4, 5, and 6, the frequencies are coded into integer numbers while "not able to answer" is coded as NA.

The indicators of physical and mental health are listed in detail in Tables. Other outcome of interests include edentulism (toothlessness), death, cancer, number of serious illnesses in the past two years, smoking habits, alcohol drinking habits, and exercising habits. Through those indicators, the causal impacts of illiteracy on physical health, mental health, and lifestyle habits are investigated in the following sections.

Code	Answer	Frequency	%
4	always	248	5.8%
3	often	595	14.0%
2	sometimes	1604	37.8%
1	seldom	894	21.1%
0	never	650	15.3%
NA	not able to answer	249	5.9%
	total	**4240**	**100.0%**

Table 6: Feeling Useless with Age Indicator

Code	Answer	Frequency	%
4	always	992	23.4%
3	often	540	12.7%
2	sometimes	1098	25.9%
1	seldom	1021	24.1%
0	never	273	6.4%
NA	not able to answer	316	7.5%
	total	**4240**	**100.0%**

Table 7: Feeling as Happy as Younger Indicator

3 Method

Matching is an essential statistical tool to ensure similar distribution of covariates in the treatment and control groups. In research contexts where randomized experimental data are lacking, matching is a promising tool to reduce confounding biases in observational data. In this paper, three matching methods are used, including nearest matching, mahalanobis matching, and propensity score matching. Information with regard to age, sex, residential area (rural, city), and ethnicity are included as potential matching covariates. Income covariates are represented using the survey question "whether you have access to tap water" and "whether you have enough financial support at the moment." Through controlling these identified covariates, the treatment and control groups have reduced background differences. In Table 8, t-test is performed to check the hypothesis of whether the covariate's means are significantly different in literate and illiterate groups. All except ethnicity are found to be statistically significant. Thus, in the following matching analysis, ethnicity is deleted as a matching covariate.

The package `MatchIt` in R is used to perform sta-

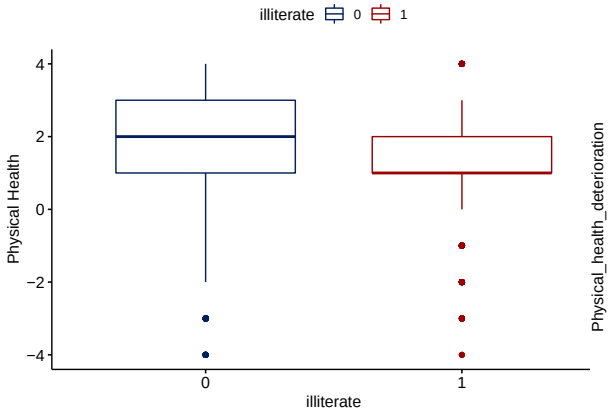

Figure 1: Box Plot for Physical Health

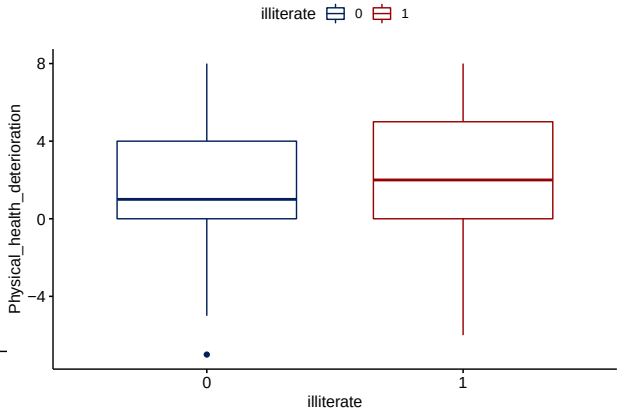

Figure 2: Box Plot for Physical Health Deteriaration

tistical analysis. However, `MatchIt` does not allow missing values and there are 880 missing values for certain variables in the CLHLS data. After removal of all NA values, the sample number decreases from 4240 to 3361.

4 Physical Health

In this section, the effects of illiteracy on physical health are examined using outcome variables including physical health (physical health = self-reported health + interviewer-rated health), physical health deterioration from 2002 to 2005, exercising habits, and smoking habits. Before matching, box plots assist visual detection of differences in literate and illiterate groups. Figures 1 and 2 show the comparison of literate and illiterate groups.

5 Mental Health

In this section, the effects of illiteracy on mental health are examined using outcome variables including feeling anxious, lonely, useless with age, and an evaluation of happiness to happiness at a younger age. Before matching, box plots assist in the visual detection of differences in literate and illiterate groups. Figures 3, 4, 5, and 6 show the comparison of literate and illiterate groups.

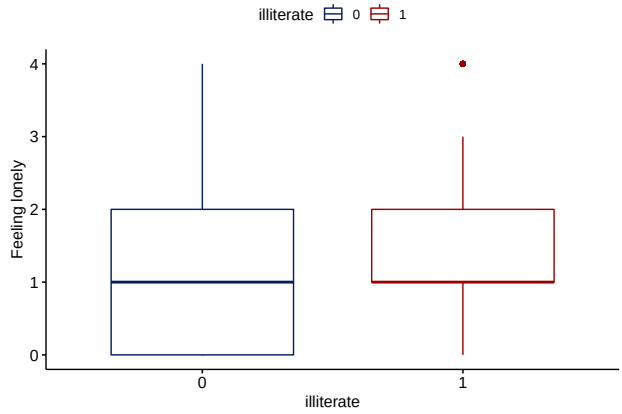

Figure 3: Box Plot for Mental Health–Loneliness

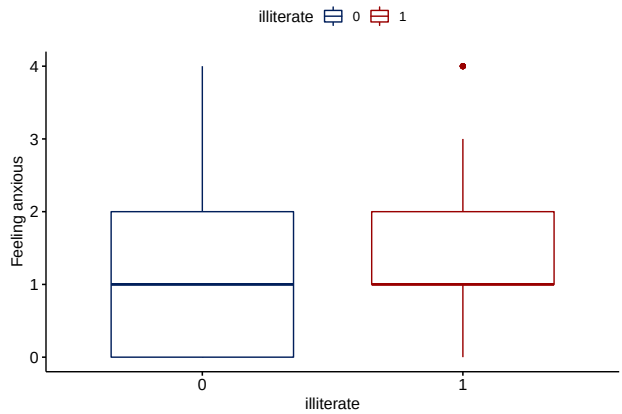

Figure 4: Box Plot for Mental Health–Anxiousness

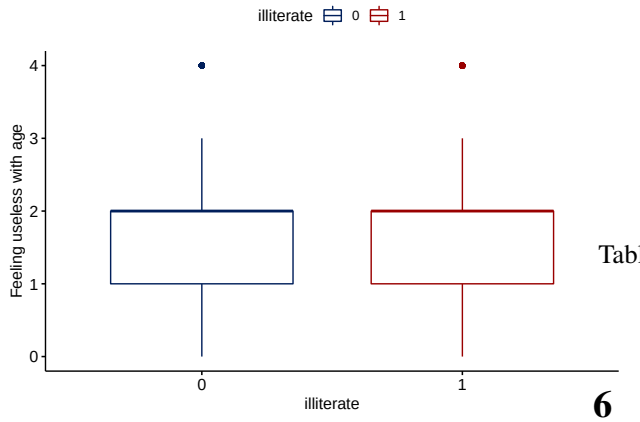

Figure 5: Box Plot for Mental Health–Uselessness

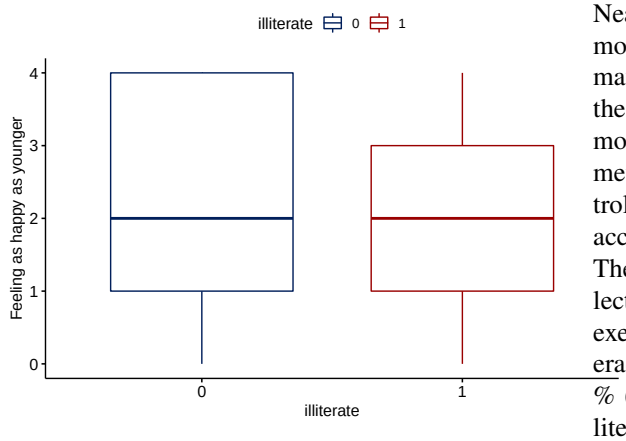

Figure 6: Box Plot for Mental Health–Happiness

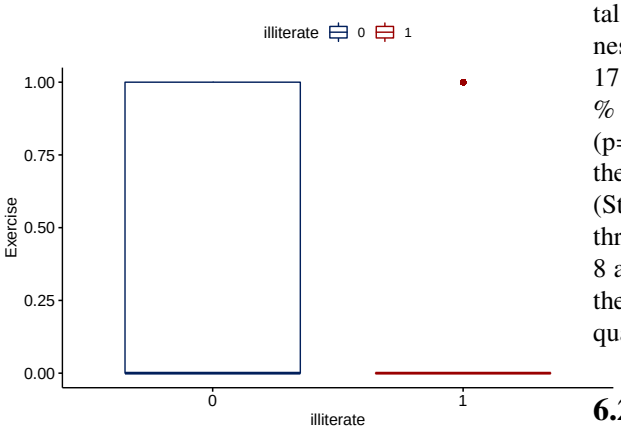

Figure 7: Box Plot for Exercising Habits

Covariate	P-value
age	2.2e-16*
living in rural areas	2.2e-16*
access to tap water	1.41e-11*
ethnicity	0.3548
sex	0.002188*
enough financial support	2.981e-12*

Table 8: T-test on whether the covariate's means are different in literate and illiterate groups
p-value smaller than α (0.05)

6 Statistical Analysis

6.1 Nearest Neighbor Matching

Nearest neighbor matching is one of the most common and easiest methods (Rubin, 1973). In nearest matching, the treatment group stays constant while the control group is partially discarded to ensure more accurate k:1 matching. The nearest matching method matched 829 treatment groups with 829 control groups on variables including age, sex, rural, access to tap water, and enough financial support. The regression results of nearest matching are collected in Table 9. Illiteracy significantly decreases exercising habits by 6 % (p=0.004). Besides, illiteracy significantly decreases physical health by 20 % (p=0.012). The counter-intuitive result is that illiteracy decreases cancer rate by 4.3 % (p=0.09). Compared to impact on physical health, illiteracy has more significant and large scale impacts on mental health. Specifically, illiteracy increases anxiousness by 11.3 % (p=0.009), increases loneliness by 17.9 % (p=0.0001), increases useless feelings by 10.9 % (p=0.037), and decreases happiness by 1.34 % (p=0.03). One limitation of nearest matching is that the matching order can change the matching quality (Stuart, 2010). This limitation does not pose much threats to this paper as the Quantile-quantile plots in 8 and 9 show the high nearest matching accuracy. In the supporting documents, detailed reports prove the quality of the matched samples.

6.2 Mahalanobis Metric Matching

The mahalanobis distance offers another type of distance criterion based on a caliper to judge matches.

Outcome Variable	Coefficient	Standard Error	P-value	Adjusted R-square	F-statistics
Physical Health	-0.200	0.079	0.012*	0.0032	6.3
Illness	-0.062	0.423	0.884322	-0.00059	0.02117
Edentulism	-0.013	0.023	0.567	-0.0004	0.3
Cancer	-0.043	0.026	0.090.	0.0011	2.9
Exercise	-0.060	0.207	0.004**	0.0183	63.8
Smoke	-0.029	0.019	0.124	0.0008	2.4
Health Deterioration	-0.154	0.128	0.230	0.0002	1.4
Anxious	0.113	0.043	0.009**	0.0035	6.9
Lonely	0.179	0.047	0.0001***	0.0082	14.7
Useless	0.109	0.052	0.037*	0.0020	4.4
Happy	-0.0134	0.062	0.03*	0.0022	4.7

Table 9: Nearest Matching Results
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Outcome Variable	Coefficient	Standard Error	P-value	Adjusted R-square	F-statistics
Physical Health	-0.211	0.079	0.007**	0.0037	7.195
Illness	-0.041	0.423	0.9227	-0.00059	0.009411
Edentulism	-0.014	0.023	0.532	-0.0004	0.391
Cancer	-0.027	0.025	0.286	$8.3e^{-}5$	1.137
Exercise	-0.103	0.021	$1.51e^{-}6 * **$	0.01328	23.3
Smoke	-0.024	0.019	0.197	0.0004	1.662
Health Deterioration	-0.1308	0.129	0.311	$1.51e^{-}5$	1.028
Anxious	0.124	0.043	0.0039**	0.0044	8.336
Lonely	0.186	0.047	$18.19e^{-}5 * **$	0.0087	15.59
Useless	0.084	0.052	0.104	0.0016	2.64
Happy	-0.192	0.061	0.00178**	0.0053	9.791

Table 10: Mahalanobis Matching Results
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

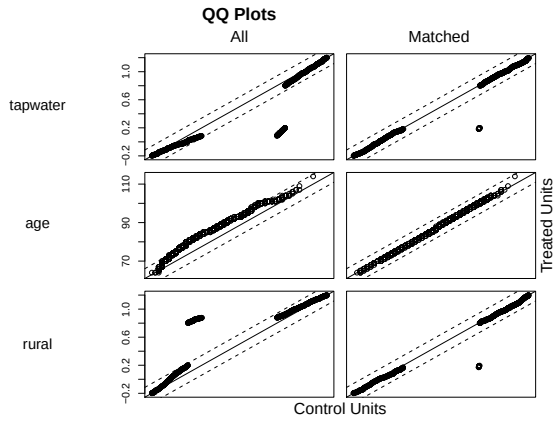

Figure 8: Matching

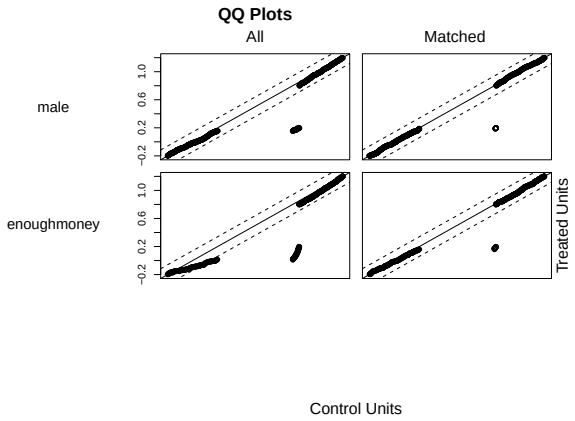

Figure 9: Matching

Outcome Variable	Coefficient	Standard Error	P-value	Adjusted R-square	F-statistics
Physical Health	-0.186	0.079	0.0193*	0.0027	5.481
Illness	-0.192	0.4359	0.66	-0.00049	0.1936
Edentulism	-0.021	0.023	0.377	-0.0001	0.781
Cancer	-0.029	0.025	0.238	0.0002	1.395
Exercise	-0.048	0.021	0.019*	0.0027	5.515
Smoke	-0.034	0.019	0.074.	0.0013	3.193
Health Deterioration	-0.228	0.130	0.0787.	0.0011	3.095
Anxious	0.1098	0.043	0.0108*	0.0033	6.518
Lonely	0.164	0.047	0.00044**	0.0068	12.4
Useless	0.0796	0.051	0.121	0.00085	2.412
Happy	-0.135	0.061	0.0276*	0.0023	4.81

Table 11: Propensity Score Matching Results
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Using the default caliper of 0.25, 829 treatment groups are successfully matched with 829 control groups. Details are included in the Appendix. The regression results of mahalanobis matching are collected in Table 10. Illiteracy significantly decreases exercising habits by 10.3 % ($p = 1.51e^{-}6$). Besides, illiteracy significantly decreases physical health by 21.1 % (p=0.007). The effects on cancer were shown to be not significant. Compared to impacts on physical health, illiteracy has more significant and large scale impacts on mental health. Specifically, illiteracy increases anxiousness by 12.4 % (p=0.0039), increases loneliness by 18.6 % ($p = 18.19e^{-}5$), and decreases happiness by 19.2 % (p=0.00178).

6.3 Propensity Score Matching

The regression results of propensity score matching are collected in Table 11. Illiteracy significantly decreases exercising habits by 4.8 % (p=0.019). Besides, illiteracy significantly decreases physical health by 18.6 % (p=0.0193). The effects on cancer were shown to be not significant. Compared to impacts on physical health, illiteracy has more significant and large scale impacts on mental health. Specifically, illiteracy increases anxiousness by 10.98 % (p=0.0108), increases loneliness by 16.4 % (p=0.00044), and decreases happiness by 13.5 % (p=0.0276).

7 Results

This study found two significant results.

1. All three matching methods conclude with significant impacts on physical health, exercise, anxiousness, loneliness, and happiness at $\alpha = 0.05$. Matching suggests that, on average, illiteracy decreases physical health by 19.9 %, decreases exercising habits by 7 %, increases anxiousness by 11.56 %, increases loneliness by 17.6 %, and decreases happiness by 11.3 %. The negative impacts of illiteracy are established by many rigorous studies. A 2006 paper studying the American ageing population found significant association between individuals with inadequate health literacy and a sedentary lifestyle (Wolf, Gazmararian, & Baker, 2007). A 2005 paper using American cross-sectional surveys also found significant worse physical and mental health conditions for individuals with inadequate health literacy (Wolf, Gazmararian, & Baker, 2005). This paper verified the above findings' external validity in the social context of China.

2. Illiteracy has a higher mental health cost as compared to its physical health cost. With rapidly decreasing illiterate groups, the illiterate elderly in China are increasingly feeling lonely and anxious. This result calls for more mental health care for the illiterate ageing groups

living in China. Ruling out the possibility of re-education, graphic assistance on medication, and audio devices to assist more convenient daily necessary activities are possible methods to improve illiterate people's life quality.

Acknowledgements

I wish to thank Dr. Wendy Chan, from Penn GSE, for taking time to kindly and patiently guide me in the process of defining my research topic, choosing appropriate analytical methods, and troubleshooting my code.

References

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political economy*, *80*(2), 223–255.

National Bureau of Statistics of China. (2010). China population census. China Statistics Press Beijing. Retrieved from %5Curl%7Bhttp://data.stats.gov.cn/easyquery.htm?cn=C01&zb=A0305&sj=2018%7D

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159–183.

Social epidemiology. (2014). Oxford University Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Wolf, M. S., Gazmararian, J. A., & Baker, D. W. (2005). Health literacy and functional health status among older adults. *Archives of internal medicine*, *165*(17), 1946–1952.

Wolf, M. S., Gazmararian, J. A., & Baker, D. W. (2007). Health literacy and health risk behaviors among older adults. *American journal of preventive medicine*, *32*(1), 19–24.