Università Commerciale "Luigi Bocconi"
PhD School

PhD program in Economics and Finance

Cycle: 33

Disciplinary Field (code): SECS-P/01

# Essays on Socioeconomic Inequality

Advisor: Jérôme ADDA

PhD Thesis by

**Manuela PUENTE BECCAR**

ID number: 3051970

**Year 2022**

# Contents

*I thank Gianmarco Ottaviano, Paola Profetta, Pamela Giustinelli and Jan Baker for always being willing to give me their time and advice.*

*I thank Diego Ubfal and Selim Gulesci, who trusted me and taught me so many things.*

*I thank Jaime, Alberto, Daria and Goonj for helping me whenever I needed.*

*I thank Jérôme Adda for everything.*

# Studying Socioeconomic Inequalities with a focus on health

In this thesis I study socioeconomic inequalities with a focus on health. In the first two chapters I study the causes of these inequalities: first looking at geographic inequalities in health behavior, and then looking at inequalities in health behavior by different socioeconomic groups throughout time. The last chapter instead studies a policy that can reduce inequalities in access to the labor market and finds unexpected beneficial effects on victims of domestic violence.

In Chapter 1, I explore the role of sorting in the observed differences in health between neighborhoods of the same city. I find that individuals who care about their health are more likely to live in neighborhoods close to health amenities and this sorting mechanism based on individuals' demographics and their health preferences explains 60% of the observed geographical health inequalities. I use the model for a counterfactual evaluation that shows that a well-intended information campaign on health can drive poor individuals further from health amenities; this results from a rise in demand of health amenities which capitalizes into house prices.

In Chapter 2, together with Jérôme Adda, we study the role of social sharing of information in the evolution of beliefs about the harmfulness of smoking and how these explain the differences in smoking prevalence by socioeconomic group. We find that social sharing is a key mechanism through which information related to health behaviors reaches people. We find that if there had been no formal dissemination of information regarding the harms of smoking in the 20th century, the learning rate of these harms would have decreased, but it would have eventually converged. This can be explained by two factors: (1) an increased life expectancy, allowing many smokers to experience smoking-related diseases,

and (2) individuals sharing this information within their social networks.

Finally, in Chapter 3, together with Selim Gulesci and Diego Ubfal, we find that a youth empowerment program that aimed at improving adolescents skills for the labor market, was able to reduce violence reporting among adolescent girls during the COVID-19 lockdown. The program offered training in soft skills and technical skills, sexual education, mentoring and job-finding assistance. To measure the effects of the program, the study conducted a randomized control trial with 600 vulnerable adolescents. Results indicate that 7 months after its completion, the program increased girls' earnings and decreased violence against girls. Violence was measured with both direct self-report questions and list experiments. The findings imply that multi-faceted empowerment programs can reduce the level of violence experienced by young women during high-risk periods.

Each of the three chapters of this thesis explains these findings in depth.

# Chapter 1

# Health preferences and sorting in the city

JEL Code: D01, I12, I14, R20

## 1.1 Introduction

In many cities of the developed world, individuals in one neighborhood can expect to live 10 years longer than those in other neighborhoods a few miles away (Gourevitch, 2019; O'Mahony, 2019), a life expectancy gap comparable to that between the United States and Laos or Senegal[1]. Understanding such big health inequalities in such small geographical units is very complex and requires a simultaneous analysis of individuals' health and the neighborhood they live in.

There are many variables and mechanisms that can play a role in this neighborhood-health relation. First, many characteristics of the individual, such as her race, income and education, might have a direct effect on her health but also an indirect effect through her choice[2] of neighborhood. Second, there are neighborhood characteristics, such as health amenities or disamenities that can be heterogeneously valued by individuals and which will also be

[1] World Health Organization 2019.

[2] "Choice" here and throughout the paper is understood as constrained choice. I explicitly consider income as an important constraint but I do not control for all other possible forms of constraints in neighborhood choice.

capitalized into house prices. Finally, an individuals' health might be directly affected by neighborhood characteristics or by those of her neighbors.

For the implementation of public policies attempting to reduce these health inequalities, it is fundamental to understand how much of the neighborhood-health relation can be explained by a causal effect of neighborhoods on health as opposed to similar people choosing to live in the same neighborhoods. In this paper, I examine this second channel by studying the location choice of individuals and their resulting selection into neighborhoods based on amenities.

In order to have a comprehensive view, I propose a structural model of neighborhood choice where individuals have heterogeneous preferences for health amenities and other neighborhood characteristics, and their choice is constrained by their income and house prices, which are an equilibrium outcome. In the model, individuals differ on their level of education, age, race, gender, family composition and health behavior, which I take as a proxy of individuals' health preferences. A higher valuation of health amenities by individuals with better intrinsic health will not only capitalize into house prices but it will lead to a sorting pattern of healthy individuals living close to health amenities and to each other. Notice that disregarding this sorting, better health in neighborhoods close to health amenities could be interpreted as a causal sign of the effect of amenities on health or as peer effects. Similarly, worse health next to polluted areas might be erroneously attributed to pollution.

To identify the sorting mechanism from the observed location of individuals, I exploit variation in health amenities of different neighborhoods given by parks, since their location was determined many years ago, especially in developed, densely populated cities. Parks provide space for physical activities, clean air (Jo et al., 2019) and can positively impact mental health (Wood et al., 2017). Furthermore, to avoid mechanic interactions between parks and certain health behaviors, such as exercising, I construct a health behavior index using health behavior variables that do not directly relate to parks, such as being an ever-smoker (a predetermined variable) or having unsafe sex. The health behavior index that I construct with these selected behaviors is highly correlated with an index that considers all the health behaviors present in my data, which also implies that the constructed indices are able to proxy for unobserved health preferences.

I find that independent of other individual characteristics such as income, education and racial homophily, individuals with good health behavior are more likely to choose neighborhoods closer to health amenities. The estimated model explains 62% of the variation in health behavior by neighborhood and it is an improvement of 35% with respect to a model that does not consider health amenities nor health preferences. The fact that I find more likely that individuals with good health behaviors–that are independent from parks–locate close to parks allows me to conclude that sorting is the main driver of my findings.

The location choice model I estimate is a discrete choice random utility model initially developed by McFadden (1973) and improved in the Industrial Organization literature (Berry et al., 1995; Nevo, 2000). Bayer et al. (2005) adapted this model for the urban setting by providing equilibrium proofs for endogenously generated neighborhood characteristics, present for example in case of racial homophily. I adapt their model to the choice of neighborhood instead of housing units, to preferences based on health instead of school quality and to the use of survey data with the incorporation of survey weights. This model of demand allows me to calculate a different marginal willingness to pay for health amenities depending on individuals' characteristics. I find that an individual who is non-poor, in working age and highly educated is willing to pay half as much to be one standard deviation closer to a park than an individual with the same characteristics who also has good health behavior.

In order to estimate the model, I got restricted-access to a pool of cross-sections for the period 2009-2013 at the individual-level from a health survey in New York City for almost 40.000 individuals located in 127 neighborhoods, allowing me to observe individuals' detailed health behavior, their income category and their geolocation, a very unique combination of data types. Using New York City as a case study has several other advantages. The first one is data availability, which allows me match to the health survey publicly available geographical data on park zones and administrative data on housing and neighborhood characteristics, including house prices. Another advantage is that due to geographical constraints the structure of the city remains constant, which allows me to interpret the results of a static model. Finally, most individuals in New York City choose their neighborhood; it is not a city where location is determined through inheritance[3].

I use the model to analyze a counterfactual evaluation that improves everyone's health behavior by 0.3 standard deviations (the observed improvement in the last ten years). This could be achieved, for example, with an information campaign addressing different health

[3] Homeownership rate in New York City was 32%, the second lowest in large cities after Miami, according to data from ACS 2019.

behaviors. I find that this counterfactual generates a 3% increase in prices of neighborhoods close to parks, and a reduction of 7% in prices of neighborhoods that are distant from parks. With respect to health, the model predicts that the relocation of individuals would generate a worsening of the average health behavior in some originally disadvantaged neighborhoods. More importantly, the increase in prices in neighborhoods close to health amenities pushes poor individuals towards more unhealthy neighborhoods.

This counterfactual evaluation shows that a well-intended policy, such as an information campaign that successfully improves everyone's health behavior, would have unintended negative consequences if individuals' sorting based on health preferences is not taken into account. Moreover, the model uncovers an important connection between individuals' health preferences and the housing market, with a health policy resulting in considerable changes in house prices.

This paper contributes to the growing literature in urban economics studying health amenities as determinants of location choice. So far the focus has been on temperature and air quality at coarse geographical levels (e.g. Chay and Greenstone, 2005; Bayer et al., 2009; Albouy et al., 2016; Mathes, 2021; Heblich et al., 2021). One exception is Darden (2021), who estimates a model of location choice to understand the urban/rural differences in smoking prevalence. This literature has not yet studied health amenities and sorting at the neighborhood level and so is not able to explain health inequalities between neighborhoods. I also contributes to the literature of location choice within cities based on amenities (e.g. Brueckner et al., 1999; Bayer et al., 2007; Walsh, 2007; Albouy and Stuart, 2020) by showing that health amenities can have an impact on location choice; and to the literature that finds capitalization of open space amenities (e.g. Bolitzer and Netusil, 2000; Morancho, 2003; Anderson and West, 2006) by showing that preferences for health can be one of the driving mechanisms.

On the other hand, I contribute to the literature on neighborhood (or place) effects on health (e.g. Ellen et al., 2001; Bilger and Carrieri, 2013; Alexander and Currie, 2017; Deryugina and Molitor, 2020;2021; Percoco, 2021) and in particular on health behavior (e.g. Kling et al., 2007; Eid et al., 2008; Zhao and Kaestner, 2010; Ludwig et al., 2013; Ou, 2019; Allcott et al., 2019), which has found clever ways to overcome the problem of individuals' sorting but has not studied it. The papers focusing on health behavior find inconsistent or small endogenous neighborhoods effects, which is in line with an important role of individuals' sorting.

Finally, this paper is an improvement over Boone-Heinonen et al. (2011) and Dang (2015) who look at neighborhood effects on physical activity and obesity respectively. These two papers try to address residential sorting, but they do not consider health amenities for location choice and they assume that health preferences can be accounted for by income and education. It is also an extension to Plantinga and Bernell (2007), who find evidence supporting the hypothesis that obese individuals sort into counties with a higher sprawl index.

The rest of the paper is organized as follows. Section 2 describes the data. A reduced form analysis is presented in Section 3. Section 4 sets out the structural model of location choice. Section 5 shows the estimation results while the counterfactual evaluation is presented in Section 6. Section 7 presents a discussion of the results and concludes.

## 1.2 Data

To understand individuals' location decisions as determinants of health in neighborhoods we need data on both neighborhood and individual characteristics. I use data at the ZIP code level for New York City from the following sources:

- *Community Health Survey*, 2009-2013 *(DOHMH)*
- All sales in the city, 2009-2013 (*NYC Open Data*)
- *Primary Land Use Tax Lot Output (PLUTO)* files, 2009 (*NYC Open Data*)
- Violent crime, 2009-2013 (*NYPD crime statistics*)
- ZIP code map (*NYC Open Data*)
- Park zones map (*NYC Open Data*)

Individual level data comes from the Community Health Survey (CHS) of the New York City Department of Health and Mental Hygiene (DOHMH, 2013). I got restricted access to the geolocation of individuals at the ZIP code level[4]. For confidentiality reasons, the data is a pooled cross-section for the period 2009-2013. This data gives information on health

[4] In the data made available by the DOHMH, respondents residing in a ZIP Code with a population less than 30.000 were combined to an adjacent ZIP code within their United Hospital Fund (UHF).

behavior and demographics for almost 40 thousand individuals in New York City located in 127 ZIP codes.

I use all sales in the city to calculate average house prices for every neighborhood. I cleaned the house price data by excluding sales of non residential units, sales of empty (or almost) lots, and sales registered with a symbolic price (usually from family transfers). To be able to compare housing expenditures with yearly income, I calculate annual rents with the price to rent ratio for each borough in NYC. I considered sales in 2009, 2011 and 2013[5] to get representative prices faced by the individuals in the health survey. The cleaned file contains approximately 135 thousand sales.

PLUTO files are used to calculate housing characteristics at the neighborhood level, namely, average number of floors (a proxy for population density), percentage of detached houses, average age of the building, and area destined to commercial or industrial use. The cleaned file for 2009[6] has approximately 850 thousand observations.

To calculate crime rates, I estimated the yearly frequency of violent crimes[7] per 1000 residents by neighborhood, with a mapping between police precincts and ZIP codes. There are approximately 350 thousand occurrences of these crimes in the period 2009-2013.

The other relevant neighborhood characteristics needed to understand individuals' location decisions are distance to the city center and distance to the closest park, as an exogenous health amenity. To calculate distance to the city center I used the Google maps API from the centroids of each ZIP code to Times Square[8] in Manhattan, using transit mode. Distance to the closest park was calculated by the minimum aerial distance from the centroids of each ZIP code to the centroids of park zones using QGIS. This can be seen in the left map of Figure 1.1, where green dots represent the centroids of park zones while brown dots represent the centroids of ZIP codes. The map at the right of Figure 1.1 shows the resulting distance by neighborhood.

A couple of clarifications are useful on the definition of parks and distance. I consider "parks" all the green areas that the NYC Department of Parks and Recreation considers park zones,

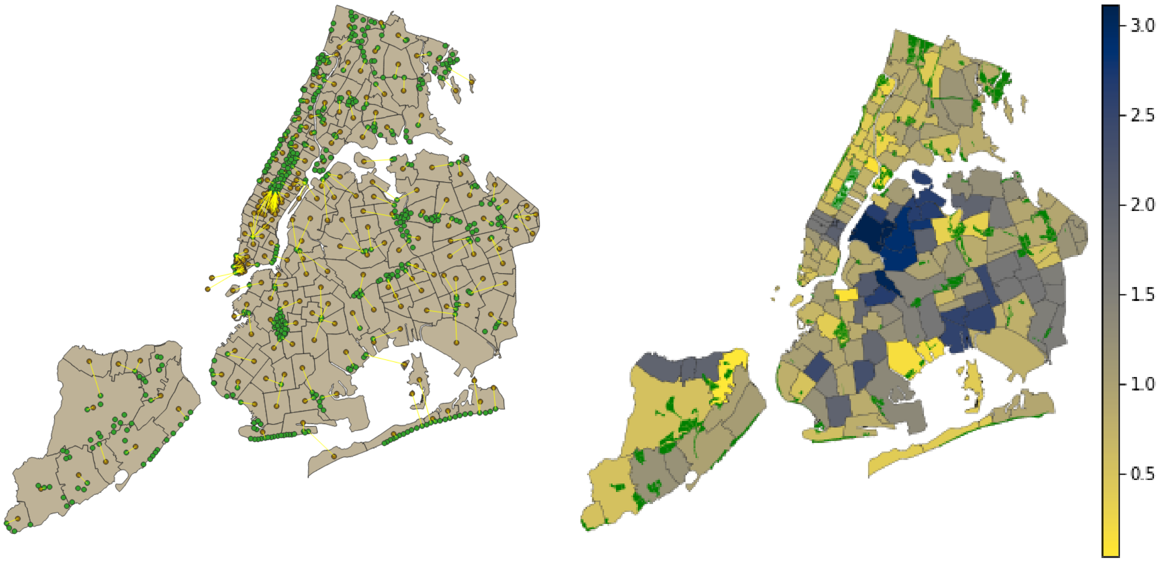
[5] House prices in NYC reached a peak in 2007 and touched bottom in the first quarter of 2009. In the second quarter of 2013 prices were back to the 2007 level.

[6] The average characteristics of 2009 are virtually equal to those of 2013. This is a confirmation that these characteristics are fixed in the short term.

[7] Violent crime includes murder, rape, felony assault, dangerous drugs and dangerous weapons.

[8] Using Wall Street gives practically the same results.

Figure 1.1: Distance to the closest park



The left panel shows the calculation of distance to the closest park by taking the linear distance from the centroids of ZIP codes to the centroids of park zones. The right panel shows the resulting attribution of distance (in Km) to each neighborhood (ZIP code).
Source: Own elaboration from geographic data on ZIP codes and park zones from NYC Open Data.

except for park areas smaller than 2 hectares (5 acres - to exclude playgrounds and to avoid endogeneity stemming from conversion of some areas into green areas). Park zones, therefore, include beaches and exclude green areas such as cemeteries or natural reserves. The analysis includes more than 250 parks and the majority of these were established between 1700 and 1950[9], which implies that their location is not the result of neighbors' decisions. Regarding distance, I choose to calculate linear distance to parks and transit distance to the CBD because it is more likely that New Yorkers go to the CBD by car or public transportation but that they go walking, running or cycling to the closest park in their neighborhood.

Table 1.1 shows descriptive statistics of individuals' demographics, health behavior variables and neighborhood characteristics. Most of the latter can be considered exogenous or fixed in the short term. The exceptions are crime rate –which I include because it is a very important determinant of neighborhood choice– and percentage white, percentage black/hispanic, which are determined in equilibrium with the estimation of the model to allow for racial

[9] The website https://www.nycgovparks.org/about/history/timeline shows the history of NYC parks. Only 7 large parks were created after 1950.

homophily[10].

Table 1.1: Descriptive statistics

| | Individual | | | | Neighborhood | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | mean | min | max | N | | mean | sd | min | max |
| Income (1000 USD) | 49.3 | 9.4 | 136 | 43955 | Annual rent (1000 USD) | 11.54 | 6.98 | 4.02 | 45.76 |
| Poor ($<$ 200% FPL) | 0.39 | 0.0 | 1.0 | 43955 | Distance CBD (min) | 48.12 | 20.26 | 12.00 | 108.68 |
| College | 0.33 | 0.0 | 1.0 | 43955 | Distance park ($Km$) | 1.10 | 0.69 | 0.04 | 3.11 |
| Female | 0.54 | 0.0 | 1.0 | 43955 | Detached houses | 0.26 | 0.24 | 0.01 | 0.85 |
| Parent (approx) | 0.21 | 0.0 | 1.0 | 43955 | Number of floors | 3.03 | 1.78 | 1.66 | 10.85 |
| Married | 0.49 | 0.0 | 1.0 | 43955 | Building age | 78.58 | 13.66 | 35.48 | 102.59 |
| Age 25-44 | 0.40 | 0.0 | 1.0 | 43955 | House size ($m^2$) | 112.25 | 17.82 | 70.19 | 159.50 |
| Age 45-64 | 0.32 | 0.0 | 1.0 | 43955 | Commercial area ($Km^2$) | 1.35 | 2.04 | 0.13 | 16.42 |
| Age 65+ | 0.15 | 0.0 | 1.0 | 43955 | Industrial area ($Km^2$) | 0.07 | 0.15 | 0.00 | 1.21 |
| White | 0.36 | 0.0 | 1.0 | 43955 | Crime rate (1000 res) | 19.20 | 16.15 | 0.83 | 83.56 |
| Black/Hispanic | 0.49 | 0.0 | 1.0 | 43955 | Percentage white | 0.37 | 0.28 | 0.0 | 0.89 |
| Insured | 0.81 | 0.0 | 1.0 | 43448 | Percentage black/hisp | 0.48 | 0.32 | 0.05 | 0.98 |
| Ever smoker | 0.36 | 0.0 | 1.0 | 43955 | | | | | |
| Current smoker | 0.15 | 0.0 | 1.0 | 43955 | | | | | |
| Unhealthy drinker | 0.09 | 0.0 | 1.0 | 43249 | | | | | |
| Unsafe sex | 0.11 | 0.0 | 1.0 | 43955 | | | | | |
| Flu vaccine | 0.37 | 0.0 | 1.0 | 43702 | | | | | |
| Colonoscopy | 0.28 | 0.0 | 1.0 | 43955 | | | | | |
| >= 1 sweetened bev. | 0.29 | 0.0 | 1.0 | 43558 | | | | | |
| <= 1 fruit/veg | 0.35 | 0.0 | 1.0 | 42997 | | | | | |
| >= 4 fruit/veg | 0.20 | 0.0 | 1.0 | 42997 | | | | | |
| Exercise | 0.75 | 0.0 | 1.0 | 43917 | | | | | |

Individual characteristics come from the CHS, as well as percentages of whites and blacks/hispanics by neighborhood. Income is approximated as is explained in the Appendix. Poor=1 if household income is below 200% of the Federal Poverty Line and “parent” is an individual in the age group 25-44 that is also married (or cohabiting). Definitions of health behavior variables can be found in Appendix Table A1.

The right panel shows descriptive statistics of the estimated neighborhood variables that are used in the analysis that follows. Annual rent is estimated from administrative data on all sales in the city; distance to the CBD is calculated with the Google maps API; distance to the closest park is estimated with GIS software; crime rates are estimated from NYPD crime statistics while all other neighborhood characteristics come from PLUTO files. There are 127 observations for all neighborhood variables.

As can be seen in Table 1.1, all individual level variables are dummies except for income. Two of the demographics are created using survey variables: an individual is poor if her household income is below 200% of the Federal Poverty Line and “parent” is an individual in the age group 25-44 that is also married (or cohabiting). Regarding neighborhood characteristics, the numbers presented in the table are averages of averages, e.g. all neighborhood have on average 3 floor buildings; the “lowest” neighborhood has an average of 1.7 floors, while the

[10] I could also allow for homophily in education and income. I do not do so due to endogeneity concerns through reverse causality.

“tallest” neighborhood has an average of 11 floors[11].

Unfortunately, some of the demographics in the individual-level data are coded into categories. The data gives information on four age groups (omitted category 18-24) and five income groups: from less than a 100% of the Federal Poverty Line (FPL) to more than 600% of the FPL. Income is an important variable for the location choice of an individual, so I include an approximation of it on the analysis. To take into account that not all individuals pay the same rent in each neighborhood, I assign rents (divided in 5 quantiles) to each individual corresponding to their income category. Details on this assignment and the approximation of income can be found in the appendix.

### 1.2.1 Characterizing Health Behavior

I take observable health behaviors as a proxy for unobservable health preferences[12]. If different health behaviors would be uncorrelated with each other, we would not know how to characterize an individual’s general health behavior and they would not be a consistent proxy for preferences. Figure 1.2 shows a correlation analysis between the different health behavior variables present in the survey[13]. Looking at correlation coefficients that are greater or equal than 0.05, we can see that in general negative (positive) health behaviors are positively correlated with each other and negatively correlated with positive (negative) health behaviors. There is only one exception (in any case close to the 0.05 threshold) which is that having unsafe sex is positively correlated with doing exercise.

The correlations presented show consistency in health behavior which allows us to create an index summarizing the information of these variables to quantify how good or bad is an individual’s health behavior[14].

One might be worried about the possible reverse causality between neighborhood characteristics and health behavior. We could think that if we randomly allocate an individual close to a park, she starts exercising which then causes her to eat healthier and to smoke and

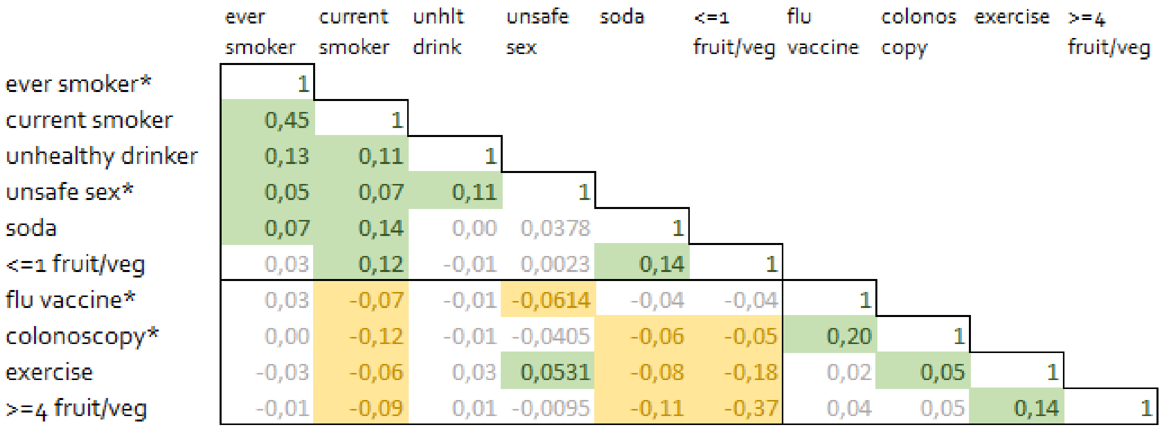
[11] Taking the median values instead of the mean does not change the results of the analysis.

[12] Behaviors are the result of preferences, information, beliefs and other constraints. Due to data availability I cannot separate these factors, but this is not fundamental for the purposes of my analysis.

[13] Appendix Table A1 presents the definition of each variable.

[14] To generate the index I add (subtract) 1 for each positive (negative) health behavior and then standardize, so the index has mean 0 and standard deviation 1. Using factor analysis gives a highly correlated index (0.92).

Figure 1.2: Correlations among health behaviors



| | ever smoker | current smoker | unhlt drink | unsafe sex | soda | <=1 fruit/veg | flu vaccine | colonos copy | exercise | >=4 fruit/veg |
|---|---|---|---|---|---|---|---|---|---|---|
| ever smoker* | 1 | | | | | | | | | |
| current smoker | 0,45 | 1 | | | | | | | | |
| unhealthy drinker | 0,13 | 0,11 | 1 | | | | | | | |
| unsafe sex* | 0,05 | 0,07 | 0,11 | 1 | | | | | | |
| soda | 0,07 | 0,14 | 0,00 | 0,0378 | 1 | | | | | |
| <=1 fruit/veg | 0,03 | 0,12 | -0,01 | 0,0023 | 0,14 | 1 | | | | |
| flu vaccine* | 0,03 | -0,07 | -0,01 | -0,0614 | -0,04 | -0,04 | 1 | | | |
| colonoscopy* | 0,00 | -0,12 | -0,01 | -0,0405 | -0,06 | -0,05 | 0,20 | 1 | | |
| exercise | -0,03 | -0,06 | 0,03 | 0,0531 | -0,08 | -0,18 | 0,02 | 0,05 | 1 | |
| >=4 fruit/veg | -0,01 | -0,09 | 0,01 | -0,0095 | -0,11 | -0,37 | 0,04 | 0,05 | 0,14 | 1 |

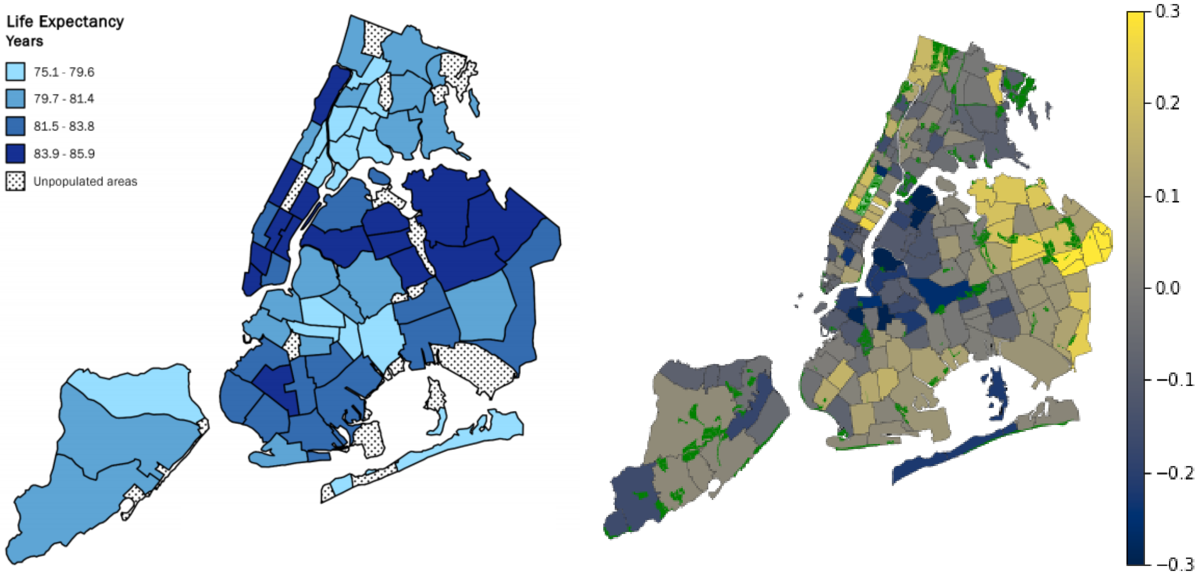
The figure shows the correlations between all the health behavior variables present in the survey data. The first six are negative health behaviors while the last 4 are positive health behaviors. Behaviors with an asterisk (*) can be considered exogenous from parks. Green numbers show a positive correlation between behaviors while yellow numbers show negative correlations. Correlations below 0.05 in absolute value, in gray, are omitted from this analysis.
Source: own elaboration using data from CHS 2009-2013.

drink less, while if the random allocation would have been distant from parks, this change in behavior would have not happened. To minimize this concern, I generate two health indices. The first index uses all the variables in Figure 1.2 while the second index uses the more exogenous health behavior variables marked with an asterisk. This second health index is less affected by the endogeneity problem, since moving next to a park should not change whether an individual is an ever smoker (that is a predetermined variable) or if she has unsafe sex. Even if the second index is generated with less than half of the variables used for the first index, the correlation between the two is very high (0.73) so in the analysis I use the second, more exogenous health index. Notice that this high correlation of indices generated with different variables is also an indication of the consistency in health behaviors and therefore the ability to proxy unobserved health preferences with this index.

It is an established fact that health behavior is an important determinant of life expectancy. Figure 1.3 shows a map of life expectancy, provided by the Bureau of Vital Statistics of NYC DOHMH on the first panel, and a map of the second health index on the right panel. This figure shows that the geographic inequalities of the second health index are very similar to the geographic inequalities in life expectancy.

Another important point to consider is the correlation between income, education and health

Figure 1.3: Life expectancy and health behavior



Life Expectancy, NYC DOHMH, Bureau of Vital Statistics, 2006-2015

Average of Health index 2 by ZIP code, using data from CHS, 2009-2013. Higher values indicate better health behavior.

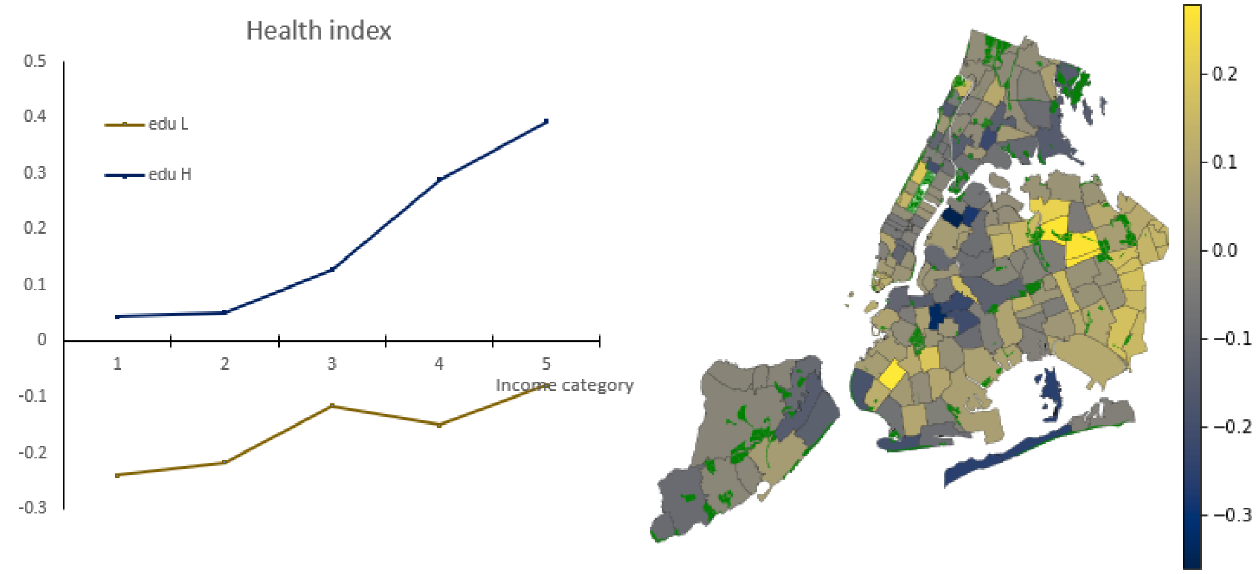
behavior, since we could expect them to be highly correlated[15]. The data I use gives information on income category (from 1 to 5) and education category (from 1 to 6). The relation between education and health behavior is not linear, but highly educated individuals –those with 4 years of college or more– have distinctly better health behavior. The first panel of Figure 1.4 shows that the health behavior index increases with income category and it is always higher for individuals with college or more.

The second panel of Figure 1.4 shows the residuals from a regression at the neighborhood level of the health behavior index on income categories, a dummy for college and age categories. We observe that after controlling for these characteristics, the geographic inequalities of this health index are still very similar to the ones observed in Figure 1.3, with healthy areas in the center of Manhattan, in the Northeast of Brooklyn and in the south of Queens[16]. Even if income and education are important drivers of health behavior, the health index variable has information on the health behavior of individuals that goes beyond the information that

[15] Cutler and Lleras-Muney (2010) show that income, health insurance, family background, knowledge and measures of cognitive ability can explain a big part of the relationship between education and health behavior.

[16] A map of New York City boroughs is presented in Appendix 1.8.1.

Figure 1.4: Relation between income, education and health behavior



Average values of the health index for different levels of income and education.

Residuals from a regression of health behavior on income, education and age at the neighborhood level.

In the left panel eduH is a dummy for individuals with college education. The income categories are as follows: 1=<100% Federal Poverty Line (FPL), 2=100-200% FPL, 3=200-400% FPL, 4=400-600% FPL, 5=>600% FPL.
Source: own elaboration using data from CHS 2009-2013.

we can obtain from income and education. Indeed, a regression at the individual level of the health index on all the income and education dummies, controlling also for age, has an $R^2$ of 0.13, even if most of these dummies are significant.

## 1.3 Reduced form analysis

If healthy individuals would choose to live closer to health amenities such as parks, we should observe better health behavior in neighborhoods closer to parks. However, as we saw in the previous section, there are other (observable) variables that influence the health behavior of individuals. Table 1.2 shows linear regressions at both neighborhood and individual level. Considering the relatively low number of neighborhoods, the first two columns show low precision in the estimates, nevertheless, we see that indeed health behavior worsens with distance to parks. This negative correlation is maintained if we control for individual's demographic characteristics and distance to the city center[17].

Table 1.2 shows that the signs of the correlations with all the variables included in the regression are those that we would expect: health behavior improves with high education and age (omitted category 18-24), and it is higher for females, married and insured individuals, while the value of the health index is lower if an individual is poor or is white, black or Hispanic, as compared to the omitted races, mainly composed by Asian.

If we look at the standardized coefficient in column 5 we can see that the correlation with distance to parks is even higher than that with being poor.

On the other hand, the urban economics theory tells us that house prices must be higher closer to amenities. If parks are considered amenities, we should observe higher prices in neighborhoods closer to parks[18]. Table 1.3 shows a negative correlation between average house prices in a neighborhood and the distance to the closest park, which is maintained after controlling for other neighborhood and housing characteristics.

In Table 1.3 we also see that house prices are lower if we move away from the CBD and if there is a higher crime rate, while house prices are higher in neighborhoods with taller buildings and bigger houses. The low number of observations does not allow precision in the estimation of some of these coefficients. The third column shows that the effect of distance to parks is close to 30% of the effect of distance to the CBD, which is thought of one the main determinants of house prices.

This reduced form analysis has two important findings. The first one is a positive correlation between having a good health behavior and living close to parks. Considering that this correlation is positive when we take into account health behavior variables that are unlikely to be affected by proximity to parks and that it is maintained if we control for other relevant variables for health behavior, such as age, race, income and education, this evidence supports the hypothesis that individuals that care about their health choose to locate close to parks. The second important finding is that, as Table 1.3 shows, this preference for parks has an effect on the housing market, by increasing housing prices close to parks.

[17] The correlation between health behavior and distance to parks is virtually the same using the index including all the health behavior variables: the coefficients are -0.02, -0.01, -0.03 and -0.03 for the four columns respectively.

[18] Indeed other papers in other contexts find a positive marginal willingness to pay to live close to parks, e.g. Poudyal et al. (2009); Wen et al. (2015); Du and Zhang (2020).

Table 1.2: Health behavior and distance to parks

| Health Index 2 | Neighborhood level | | Individual level | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| | Coef./std.err | Coef./std.err | Coef./std.err | Coef./std.err | beta coef. |
| Distance park (Km) | -0.02 | -0.03*** | -0.03** | -0.03*** | -0.02 |
| | (0.02) | (0.00) | (0.02) | (0.01) | |
| Distance CBD (min) | | -0.00*** | | -0.00 | -0.01 |
| | | (0.00) | | (0.00) | |
| Poor ($<$ 200% FPL) | | 0.04*** | | -0.03* | -0.01 |
| | | (0.01) | | (0.01) | |
| College | | 0.37*** | | 0.13*** | 0.06 |
| | | (0.01) | | (0.02) | |
| Female | | 0.35*** | | 0.17*** | 0.09 |
| | | (0.02) | | (0.02) | |
| Parent | | -0.70*** | | 0.10*** | 0.04 |
| | | (0.02) | | (0.03) | |
| Married | | 0.92*** | | 0.30*** | 0.15 |
| | | (0.02) | | (0.02) | |
| Age 25-44 | | -0.21*** | | -0.31*** | -0.15 |
| | | (0.02) | | (0.03) | |
| Age 45-64 | | 0.69*** | | 0.37*** | 0.17 |
| | | (0.02) | | (0.03) | |
| Age 65+ | | 1.07*** | | 0.85*** | 0.31 |
| | | (0.02) | | (0.03) | |
| White | | -0.33*** | | -0.28*** | -0.14 |
| | | (0.00) | | (0.03) | |
| Black/Hispanic | | 0.03*** | | -0.06*** | -0.03 |
| | | (0.01) | | (0.02) | |
| Insured | | -0.07*** | | 0.23*** | 0.09 |
| | | (0.01) | | (0.02) | |
| Constant | 0.02 | -0.66*** | 0.04* | -0.40*** | |
| | (0.02) | (0.02) | (0.02) | (0.04) | |
| Obs. | 127 | 43955 | 43955 | 43448 | |
| $R^2$ | 0.01 | 0.59 | 0.00 | 0.22 | |

The first two columns show the coefficients from a linear regression where the dependant variable is the mean of the health behavior index by neighborhood. Columns 3, 4 and 5 use instead the individual health behavior index. Column 5 shows the coefficients of column 4 standardized.

Standard errors in parentheses, clustered by neighborhood, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: own elaboration using data from CHS 2009-2013 and distances calculated as explained in section 1.2.

Table 1.3: House prices and distance to parks

| | Annual rent (thou USD) | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| | Coef./std.err | Coef./std.err | beta coef. |
| Distance park ($Km$) | -1.292 | -0.963 | -0.10 |
| | (0.897) | (0.698) | |
| Distance CBD (min) | | -0.118$^{*}$ | -0.34 |
| | | (0.063) | |
| Detached houses | | 2.473 | 0.08 |
| | | (4.144) | |
| Number of floors | | 1.719$^{***}$ | 0.44 |
| | | (0.617) | |
| Building age | | 0.011 | 0.21 |
| | | (0.054) | |
| House size ($m^2$) | | 0.148$^{***}$ | 0.38 |
| | | (0.036) | |
| Commercial area ($Km^2$) | | -0.033 | -0.01 |
| | | (0.313) | |
| Industrial area ($Km^2$) | | 8.177$^{**}$ | 0.17 |
| | | (3.415) | |
| Crime rate (per 1000 res) | | -0.049 | -0.11 |
| | | (0.031) | |
| Constant | 12.966$^{***}$ | -7.511 | |
| | (1.165) | (7.613) | |
| Obs. | 127 | 127 | |
| $R^2$ | 0.02 | 0.62 | |

Coefficients from a linear regression at the neighborhood level. Column 2 includes borough dummies, Column 3 shows the coefficients of column 2 standardized.
Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Source: own elaboration using data from different sources as explained in section 1.2.

The connection with the housing market implies that for any policy that we wish to implement, we must consider the effects on relocation, housing prices and the feedback between them. More importantly, this reduced form analysis provides some evidence for the fact that we cannot take neighborhood configuration as given if we want to study neighborhood effects. For example, if we see that certain neighborhoods have high obesity prevalence after controlling for income and education, we cannot deduce that this is explained by neighborhood effects since it could well be that each individual is unhealthy by itself, and would be so in any neighborhood. This does not mean that there is no role for neighborhood effects, it just means that we should not attribute all the observed differences to these effects. As evidenced by Figure 1.4 and by Table 1.2, income and education are not enough to account for health preferences and how these influence location choice.

The reduced form analysis presented so far is useful to understand the basic relations between the variables of interest, but it also suffers from many drawbacks. The $R^2$ in the second column of Table 1.3 is 0.46 which is quite high but this is a regression of averages, which greatly reduces the true variability. Additionally, the number of observations is small. A discrete choice model that allows us to exploit the individual level data is a better alternative. Moreover, the interactions between health, location choice and equilibrium prices cannot be studied with reduced form regressions. A structural analysis allows to answer the following questions:

- What is the role of health preferences in individual's location choice?
- How much of the geographical health inequalities can be explained by sorting?
- How do people relocate if there are changes in people's health behavior?
- How does this affect equilibrium prices and the geographical health inequalities?

An important characteristic of New York City is that location is not determined by inheritance of a house, implying that most of its inhabitants actually choose where to live[19]. First of all, the home ownership rate is around 30%. From the NYC-HANES survey 2014, we know that 50% of individuals moved in the past 5 years while 90% moved after age 18. From another survey[20] we know that 80% of the individuals who moved, did so to a different

[19] This does not imply that they can live wherever they want, but they have a degree of choice among the available alternatives.

[20] https://www.brickunderground.com/live/why-new-yorkers-move-so-often

neighborhood. This evidence allows me to interpret the observed location of individuals as their choice.

In the next section I present a structural model of location choice, that allows to study the choice of neighborhood considering individual preferences, including preferences for health as measured by the health behavior index, and neighborhood characteristics, including distance to the closest park as a relevant health amenity.

## 1.4 Location choice model

To model the choice of neighborhood, I use a discrete choice model based on McFadden (1973) and I follow Bayer et al. (2005) –which I will refer to as BMR– for the estimation procedure[21].

The indirect utility of agent $i$ in neighborhood $j$ is given by:

$$V_{ij} = \beta^{DI} log(I_i - P_j) + \mathbf{Z_j'}\beta_\mathbf{i}^\mathbf{Z} + \mathbf{\overline{R}_j'}\beta_\mathbf{i}^\mathbf{R} + \xi_j + \epsilon_{ij} \tag{1.1}$$

where $I_i$ is the individual's income, $P_j$ is the price of housing at neighborhood $j$, $\mathbf{Z_j}$ is a vector of relevant characteristics of neighborhood $j$ and $\overline{R}_j$ is the average of individuals' race in neighborhood $j$, a variable determined in equilibrium. The coefficient $\beta^{DI}$ captures the effect of Disposable Income, i.e. the income individuals can dispose of after paying for housing.

The error term is divided into two components, $\xi_j$, an unobserved (by the econometrician) element of neighborhood characteristics relevant for location choice, and $\epsilon_{ij}$, an individual specific taste shock. The vector of observed neighborhood characteristics, $\mathbf{Z_j}$ contains the distance to the Central Business District (CBD), the distance to the closest park, as an important and exogenous health amenity, and other housing and neighborhood characteristics, such as the average number of floors of houses in the neighborhood or the commercial area. The CBD is a proxy for occupation location[22] and for other amenities such as theaters. I include the logarithm of income minus price to allow for wealth effects and by including only

[21] BMR, in turn, base their model and estimation procedure on Berry et al. (1995).

[22] Appendix table A2 shows the relation between residents and employment in each county in New York City.

distance to the park as a relevant health amenity, I am able to avoid endogeneity problems with the unobservable characteristics in $\xi_j$ since the location of parks can be considered exogenous.

The model allows each individual to respond differently to $\mathbf{Z_j}$ and $\overline{R}_j$:

$$\beta^r_{i,k} = \beta^r_{0,k} + \beta^{\mathbf{r}}_{\mathbf{1},\mathbf{k}}\mathbf{y_i}, \quad r = Z, R \tag{1.2}$$

where $k$ indicates each component of $\mathbf{Z_j}$ and $\overline{R}_j$, and $\mathbf{y_i}$ is a vector of individual characteristics, such as education, race, and health behavior. This specification relaxes the Independence of Irrelevant Alternatives assumption since different individuals will have a different ordering of preferred alternatives (Nevo, 2000).

Individuals will choose the neighborhood that gives them the highest utility. If we denote $W_{ij}$ all the deterministic components of equation (1.1), individual $i$ will choose neighborhood $j$ only if

$$V_{ij} \geq V_{ik} \iff W_{ij} - W_{ik} \geq \epsilon_{ik} - \epsilon_{ij} \quad \forall k \neq j$$

Assuming that the individual specific shocks follow the extreme value distribution, the probability that individual $i$ chooses neighborhood $j$ is given by:

$$p_{ij} = \frac{exp(\beta^{DI}log(I_i - P_j) + \mathbf{Z_j'}\beta^{\mathbf{Z}}_{\mathbf{i}} + \overline{\mathbf{R}}_{\mathbf{j}}'\beta^{\mathbf{R}}_{\mathbf{i}} + \xi_j)}{\sum_l exp(\beta^{DI}log(I_i - P_l) + \mathbf{Z_l'}\beta^{\mathbf{Z}}_{\mathbf{i}} + \overline{\mathbf{R}}_{\mathbf{l}}'\beta^{\mathbf{R}}_{\mathbf{i}} + \xi_l)} \tag{1.3}$$

Notice that if utility depended linearly on income minus price (instead of log-linearly), income would drop from this equation since it does not vary with neighborhood choice. Considering that individuals' expenditure on housing is large relative to their income, it is more appropriate to allow for income effects.

The model can be estimated with maximum likelihood, by selecting the parameters that increase the probability that each individual makes the choice observed in the data. The likelihood function is:

$$L = \prod_i \prod_j \mathbb{1}_{ij} p_{ij} \tag{1.4}$$

where $\mathbb{1}_{ij}$ is an indicator function equal to 1 if individual $i$ does live in neighborhood $j$.

With this setup of the model, however, we cannot identify all the parameters of equation (1.3). This is because the individual-specific coefficients have a fixed component which is not interacted with individual characteristics. The model can only identify one constant per alternative, denoted $\delta_j$, that comprises the following elements:

$$\delta_j = \mathbf{Z_j'}\beta_{\mathbf{0}}^{\mathbf{Z}} + \overline{\mathbf{R}}_{\mathbf{j}}'\beta_{\mathbf{0}}^{\mathbf{R}} + \xi_j \tag{1.5}$$

If the individual-level variables are constructed in such a way that they have zero-mean, then $\delta_j$ can be interpreted as the mean indirect utility of each neighborhood. In other words, $\delta_j$ measures the level of utility that each individual enjoys by living in neighborhood $j$, regardless of individual characteristics and preferences.

Grouping all the neighborhood specific variables in one constant per neighborhood, the estimated model is a Conditional Logit with alternative varying regressors and an alternative specific constant (McFadden, 1973). If we further group all the variables that vary by both individuals and neighborhoods in $\lambda_{ij}$, the equation to estimate can be conveniently expressed as:

$$p_{ij} = \frac{exp(\delta_j + \lambda_{ij})}{\sum_l exp(\delta_l + \lambda_{il})} \tag{1.6}$$

with $\lambda_{ij} = \beta^{DI} log(I_i - P_j) + \mathbf{Z_j'}\beta_{\mathbf{1}}^{\mathbf{Z}}\mathbf{y_i} + \overline{\mathbf{R}}_{\mathbf{j}}'\beta_{\mathbf{1}}^{\mathbf{R}}\mathbf{y_i}$

When the number of alternatives is large, estimating all the $\delta_j$ coefficients with an optimization algorithm can be very time consuming. One alternative is to estimate these coefficients with a contraction mapping for each combination of parameters in $\lambda_{ij}$, where $t$ denotes the iteration:

$$\delta_j^{t+1} = \delta_j^t + log\Big(\frac{S_j}{\sum_i \hat{p}_{ij}}\Big) \tag{1.7}$$

In this way, the $\delta_j$ coefficients will be those that equate the observed shares of each alternative, $S_j$ to the shares estimated by the model[23].

The components of $\delta_j$ can be estimated in a second stage through a linear regression of equation 1.5, by plugging in the left side the alternative specific constants previously estimated. The residual of this regression will be the unobserved neighborhood characteristics.

[23] BMR show that this equivalence can be derived from the first order condition of the optimization of the likelihood.

Before estimating these equations we must address the endogeneity stemming from $\xi_j$, since these unobserved characteristics are probably correlated with prices. For example, school quality, that would be part of $\xi_j$, could affect both the utility of living in a given neighborhood –and therefore the probability of choosing it– and the price to live in that neighborhood.

To address this endogeneity, I use an adaptation of the instrument proposed by BMR. They instrument the price of a specific house using exogenous housing characteristics of houses at a reasonable distance. In my case, I need to instrument for the average price of living in a given neighborhood, so the instrument I use are the exogenous characteristics of adjacent neighborhoods. To understand the instrument, notice that when an individual is considering a neighborhood in which to live in, the price of living in this neighborhood will not only depend on the characteristics of the neighborhood itself, but also on the characteristics of similar neighborhoods that the individual considers as alternative options. However, once the individual has made his choice, the characteristics of neighborhoods other than the one actually chosen, will not have any effect on his utility. Therefore, exogenous characteristics of neighborhoods other than $j$ are related to the price of living in $j$ without affecting the utility of living in $j$.

Finally, I need to take into account that the data used is survey data, which implies that each observation has a different weight. Survey weights change equation 1.4 into equation 1.8[24] and equation 1.7 into equation 1.9:

$$L = \prod_i \prod_j \mathbb{1}_{ij} w_i * p_{ij} \tag{1.8}$$

$$\delta_j^{t+1} = \delta_j^t + log\Big(\frac{S_j}{\sum_i w_i * \hat{p}_{ij}}\Big) \tag{1.9}$$

This completes the estimation of the location choice model. All the variables included in $\mathbf{Z_j}$ are arguably exogenous but important for the location decisions of individuals. The coefficient of interest is the one of the interaction between distance to parks and an individual's health behavior[25]. If it is true that individuals choose where to live taking into account their

[24] See Bruch and Mare (2012).

[25] My interest in this interaction coefficient is the main reason why I use a Discrete choice model instead of a hedonic approach, which would only provide a weighted average of the marginal utilities of individuals (Wong, 2018).

health preferences, then this coefficient should be negative and significant, implying that healthier individuals are less likely to choose neighborhoods distant from health amenities such as parks.

## 1.5 Results

This section shows the parameter estimates of the location choice model, calculations of marginal willingness to pay and an analysis of model performance. Definitions and sources of all the variables used were presented in Section 1.2.

### 1.5.1 Parameter estimates

Table 1.4 shows most interaction coefficients between neighborhood and individual characteristics, i.e. the $\beta_1$ (Table A3 shows the full set of interaction coefficients), the signs of the coefficients can be directly interpretable. Conditional on prices and other characteristics, poor individuals are less likely to live closer to the CBD, they are less likely to live in neighborhoods with detached houses, neighborhoods with skyscrapers, older buildings and larger houses. Ceteris paribus, poor individuals are less likely to live in neighborhoods with large percentages of whites or large percentages of blacks/Hispanics. Notice, however, that the racial homophily coefficient is quite large[26], so that a poor, white individual is indeed more likely to choose a white neighborhood. Almost all the coefficients in Table 1.4 are those that we could expect.

The coefficient of interest is that of the interaction between distance to parks and being healthy. This coefficient is negative and significant, which implies that healthy people are less likely to choose neighborhoods that are more distant from parks.

Table 1.4 also shows a negative relation (even if non-significant) between being healthy and living in neighborhoods with large factory areas. Industrial areas could be related with health preferences. Industry in New York City includes the manufacturing of processed foods, chemicals, fabricated metals, plastics and others. It is possible that people that care about their health prefer to be in neighborhoods distant from these industrial areas. Notice that by looking at health behavior instead of health, we reduce concerns of reverse causality, since it could well be that individuals develop health problems because they are living close

[26] The racial homophily coefficient for blacks/hispanics is 3.5, presented in Appendix Table A3

Table 1.4: Selected Interaction coefficients

| | poor | educated | parent | age65m | white | healthy |
|---|---|---|---|---|---|---|
| Distance CBD | -0.0042*** | -0.0082*** | -0.0028 | -0.001 | 0.0019 | -0.0016 |
| Distance park | -0.018 | -0.0311** | -0.0164 | 0.0262 | -0.0381* | -0.043*** |
| Detached house | -0.2617*** | 0.2713*** | -0.0726 | 0.0576 | -0.2377** | 0.1002 |
| Nr floors | -0.0824*** | 0.0849*** | -0.0086 | 0.0113 | -0.0077 | 0.033*** |
| Building age | -0.0004*** | 0.0036*** | -0.0019 | -0.0008 | -0.0026* | -0.0015 |
| House size | -0.0021*** | 0.0021*** | 0.0004 | 0.0009 | -0.0012 | 0.0017** |
| Comm area | -0.001 | -0.0128** | 0.0076 | 0.0004 | -0.0104 | -0.0083 |
| Indus area | -0.1122 | -0.0903 | -0.1381 | -0.2376** | 0.1911* | -0.0007 |
| Crime rate | 0.0007 | -0.0018** | -0.0025** | -0.0003 | -0.0009 | -0.0007 |

Coefficients from the interactions between neighborhood and individual characteristics, with clustering at the individual level. Omitted columns show interaction coefficients with *female, married, age25-44, age45-64, insured* and *black/hisp*, the full set of coefficients and standard errors is in Appendix Table A3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

to industrial areas, but it is much less likely that for this reason they develop bad health behaviors.

Table 1.4 shows that after controlling for age, gender, marital status, income, education, and racial homophily, whether an individual is intrinsically healthy has predictive power in the choice of neighborhood: healthy individuals will choose to live closer to health amenities such as parks.

As explained in section 1.4, the prices used for this estimation are the predicted prices using instrumental variables. These regressions are presented in Appendix Table A4, where we see that exogenous characteristics of adjacent neighborhoods are able to explain a lot of the variability in equilibrium prices and the F-test is above the rule of thumb acceptance level (F-test from 16 to 36).

Table 1.5 shows the second stage of the estimation: the decomposition of the neighborhood specific constants into its components. Given that the individual level characteristics have been constructed to have zero-mean, these $\delta_j$ measure the mean indirect utility of living in neighborhood $j$. In Table 1.5 we see that neighborhoods with larger commercial areas and newer buildings provide higher utility while neighborhoods with high crime rates, and many tall buildings provide in general lower utility to its residents. The number of observations is dictated by the number of ZIP codes which does not allow for a lot of precision in the estimation.

The total effect on utility from a given characteristic is the sum between the $\beta_0$ presented

Table 1.5: Delta decomposition

| | Coef. | Std. error |
|---|---|---|
| Distance CBD (min) | -0.001 | (0.00) |
| Distance park ($Km$) | 0.008 | (0.05) |
| Detached houses | 0.049 | (0.26) |
| Number of floors | -0.175*** | (0.04) |
| Building age | 0.006** | (0.00) |
| House size ($m^2$) | -0.004 | (0.00) |
| Commercial area ($Km^2$) | 0.097*** | (0.02) |
| Industrial area ($Km^2$) | -0.167 | (0.26) |
| Crime rate (per 1000 res) | -0.011*** | (0.00) |
| Percentage white | -0.335 | (0.31) |
| Percentage black/hisp | -0.092 | (0.28) |
| Obs. | 127 | |
| $R^2$ | 0.85 | |

Coefficients from a linear regression on the alternative specific constants on neighborhood characteristics. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

in Table 1.5 and the $\beta_1$ from the interactions presented in Table 1.4 interacted with the individual's characteristics. These effects, therefore, are individual specific. The Marginal Willingness to Pay (MWTP) for characteristic $x$ would be the marginal rate of substitution between $x$ and income. From the indirect utility function in equation 1.1 I obtain:

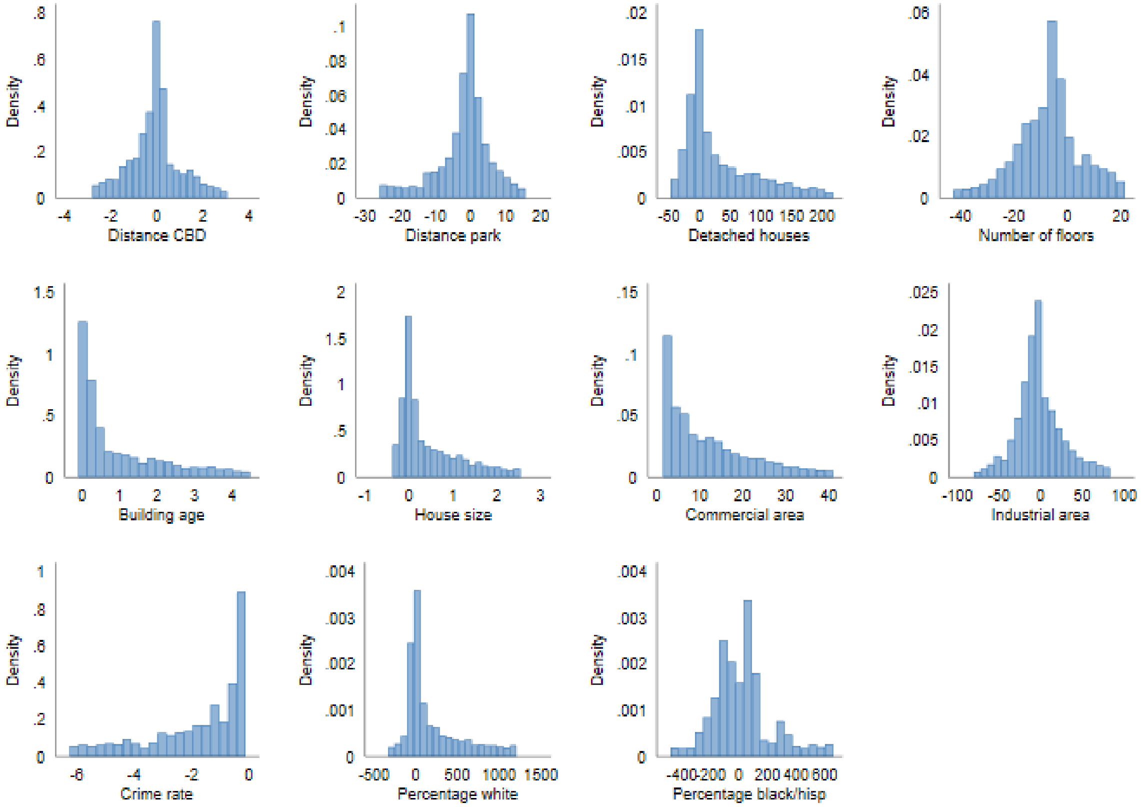
$$MWTP_i(x) = \frac{\partial V/\partial x}{\partial V/\partial I} = \frac{\beta_i^x}{\beta^{DI}}(I_i - P_j)$$

Figure 1.5 shows the distribution of the MWTP for all the neighborhood characteristics. The figure shows that everybody appreciates commercial areas, newer buildings and low crime, while preferences for other characteristics, including distance to parks, are more heterogeneous.

In Figure 1.6, I separate individuals in two groups, the upper half of individuals according to the health index is classified as healthy while the bottom half is classified as unhealthy. We clearly see that the mass of healthy individuals is located at the left of the zero, implying a *positive* MTWP to live *closer* to parks, while the mass of unhealthy individuals is located at the right of zero.

To understand how this MWTP to live close to parks compares to the MWTP for other neighborhood characteristics, I calculate the MWTP for a change in one standard deviation

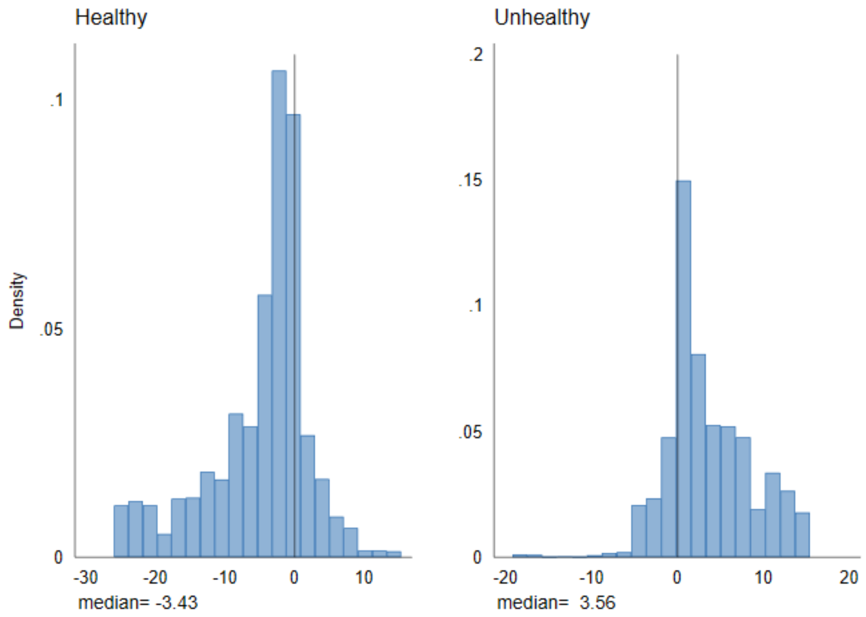
Figure 1.5: Distribution of MWTP for neighborhood characteristics



Each plot shows the distribution of the marginal willingness to pay estimated for all individuals in the sample for each neighborhood characteristic. The value in the x-axis is in thousands of USD.

of each characteristic for a typical New York individual: not poor, working (age below 65) and with high education, and for these individuals who are also healthy (health behavior index above average). Results are presented in table A5 in the Appendix. These calculations show that a typical individual would be willing to increase her annual rent by 36 thousand dollars (37% of their income) to live one standard deviation (20 minutes) closer to the CBD keeping all else equal, and she would be willing to increase her rent by 7 thousand dollars (7% of their income) to live one standard deviation (0.7 Km) closer to a park. An individual with the same characteristics who is also healthy, would be willing to pay a 43 thousand dollars (43% of their income) to live closer to the CBD, but she would be willing to increase her annual rent by more than 15 thousand dollars (15% of their income) to live one standard

Figure 1.6: Distribution of MWTP to live distant from parks



Distribution of the marginal willingness to pay for being distant from parks for healthy (unhealthy) individuals, defined as those with a health index above (below) the median. The value in the x-axis is in thousands of USD.

deviation closer to a park[27], more than twice as much the MWTP of a typical individual.

### 1.5.2 Model performance

I use the model to assign each individual to a neighborhood according to the estimated probabilities. To do this assignment, I first calculate the cumulative probability that each individual lives in each (ordered) neighborhood. This cumulative probability goes from 0 to 1. Then, I draw random variables from a Uniform distribution in [0 , 1] and assign the individual to the corresponding neighborhood in the cumulative probability. For example, suppose an individual can choose between 4 neighborhoods and the following are the probabilities given by the model:

$$[0.1 \quad 0.5 \quad 0.2 \quad 0.2]$$

[27] Considering that both income and the actual rents paid by individuals are approximations, these values are useful mainly to compare the relative effect of each characteristic and should not be taken as precise quantitative estimates. See section 1.7 for a more detailed discussion.

Then the cumulative probability vector would be

$$[0.1 \quad 0.6 \quad 0.8 \quad 1]$$

A variable drawn from a Uniform [0 , 1] would have exactly a 10% probability of falling in the interval [0 ; 0,1] implying that the chosen neighborhood is the first one, 50% probability of falling in the interval [0,1 ; 0,6] so that the chosen neighborhood is the second one and 20% probability of falling in the intervals [0,6 ; 0,8] or [0,8 ; 1], implying the choice of the third or the fourth neighborhood respectively. To avoid results to be driven by the randomness of the draw, I repeat this procedure many times and take the average.

With individuals assigned to neighborhoods, I can calculate the inequality in health behavior predicted by the model and compare it to the observed health inequality in the data. To better understand the fit of the model and the importance of the variables considered, I estimate the model omitting health behavior and distance to parks. I refer to this as the "standard" model of location choice, since it predicts the sorting of individuals in the city taking into account all the demographics and neighborhood characteristics except for health preferences and health amenities.

As a first measure of fit, I look at the correlation between the observed and the predicted average health behavior by neighborhood. The standard model has a correlation of 0.27, while my model has a correlation of 0.39, improving performance by 44%. As a second measure of fit, I calculate the percentage of the true variance the model is able to predict, similar to an $R^2$ measure. My model predicts a distribution of health behavior by neighborhood with a variance equal to 62% of the observed one. The standard model is only able to capture 46% of the variation, an improvement of 35%. Having a good approximation of the dispersion in health is important if we want to understand geographic inequalities.

These numbers show two things. First, that the model is able to replicate a lot of the geographical variation in health observed in the data. Second and most important, that this model of residential sorting explains a large part of the inequality in health behavior within a city.

### 1.5.3 Robustness checks

I perform several robustness checks that confirm the findings presented so far. First, to check robustness to the sample, I estimate the model excluding each borough at the time. The resulting coefficients are numerically similar and the qualitative results are maintained. Appendix Table A6 presents these coefficients for the most relevant variables.

Appendix table A8 shows robustness to the definition of income. The first 5 columns show the main coefficients for estimations excluding each income category at the time. Column 6 shows these coefficient for a model that allows for heterogeneous preferences for disposable income, i.e. estimating $\beta_i^{DI}$ instead of $\beta^{DI}$.

To check robustness with respect to the main variables of the analysis, I estimate the model using the first health index (the one calculated with all the health behavior variables) instead of the more exogenous health index 2. The interaction coefficient between distance to the closest park and the health index gives a negative and significant coefficient (-0.035). This result is presented in the first column of Table A7. I also consider an alternative index formed only with the variables *ever smoker* and *unsafe sex*, to take into account the possibility that both *flu vaccine* and *colonoscopy* are capturing individuals' frailty (and therefore bad health) or insurance coverage.

The other columns of this table show robustness to the variable measuring distance to parks. In column 3 the model is estimated with the logarithm of distance to parks, to allow for nonlinear effects. Column 4, instead, shows that using park area in each neighborhood instead of distance to the closest park, gives a positive and significant coefficient (0.012), implying that healthy individuals are more likely to choose neighborhoods with (larger) parks in them.

The last two column of Appendix Table A7 show the results of the estimation with a weighted version of distance to parks to control for park quality. It is possible that some parks are less *amoenus* than others, either because they are neglected by the city government or because they are dangerous. For the coefficient presented in column 4, I use data on the cost of park maintenance[28] to calculate a weight that is inversely proportional to this investment. In this way, parks where the city government invests a lot can be considered closer while parks with lower investment can be considered further. In column 5, instead, I weight distance with

[28] Data on park maintenance from the NYC Department of Parks and Recreation.

reports of crime in each park[29] taking park area into account. In this way, parks with more occurrences of crime per acre are considered further. As the coefficients in Table A7 show, the qualitative results are maintained.

## 1.6 Counterfactual evaluation

A widely used policy to improve health behaviors are information campaigns. For example, from 2002 to 2013 the New York City government used information campaigns to address tobacco use, obesity and HIV infection (Fairchild et al., 2015). Such campaigns have the potential of improving health behavior in many dimensions. In this section, I use the estimated model to look at the consequences of an information campaign that would succeed in improving everyone's health behavior[30].

To run this counterfactual analysis, I increase everyone's health behavior index by 0.3 standard deviations, which is the observed improvement in the period 2009-2019[31]. This change generates a new ranking of preferences over neighborhoods for each individual.

To find the new probabilities that each agent $i$ chooses neighborhood $j$, we must also take into account that with a change in demand, house prices must change to find a new equilibrium. Moreover, since people have preferences over the racial composition of neighborhoods, we must make sure that the model is in equilibrium regarding both prices and racial composition. The steps to calculate the new equilibrium are:

1. Take the sociodemographic composition as given
2. Calculate the prices that clear the market with the following contraction mapping:

$$\hat{P}_j^{t+1} = \hat{P}_j^t + ln\Big(\hat{S}_j(\hat{P}_j^t)/S_j\Big)$$

   where $\hat{S}_j = \sum_i \hat{p}_{ij}(\hat{P}_j)$ and $t$ is the iteration

[29] Parks crime statistics from the NYPD.

[30] I do not assume differential effects in subgroups since Stead et al. (2019), in a review of reviews, find neutral or inconsistent evidence of differential effects of information campaigns by subpopulations such as ethnicity or socioeconomic group.

[31] For this calculation I used the publicly available yearly datasets from the Community Health Survey of DOHMH. The improvement in health behavior is virtually the same in both high- and low-educated individuals.

3. With the new equilibrium prices and the corresponding choice probabilities, calculate the new sociodemographic composition:

$$\hat{Y}_j = \sum_i y_i \cdot \hat{p}_{ij}$$

4. Repeat steps 1 to 3 until $\hat{Y}_j^{t+1} = \hat{Y}_j^t$

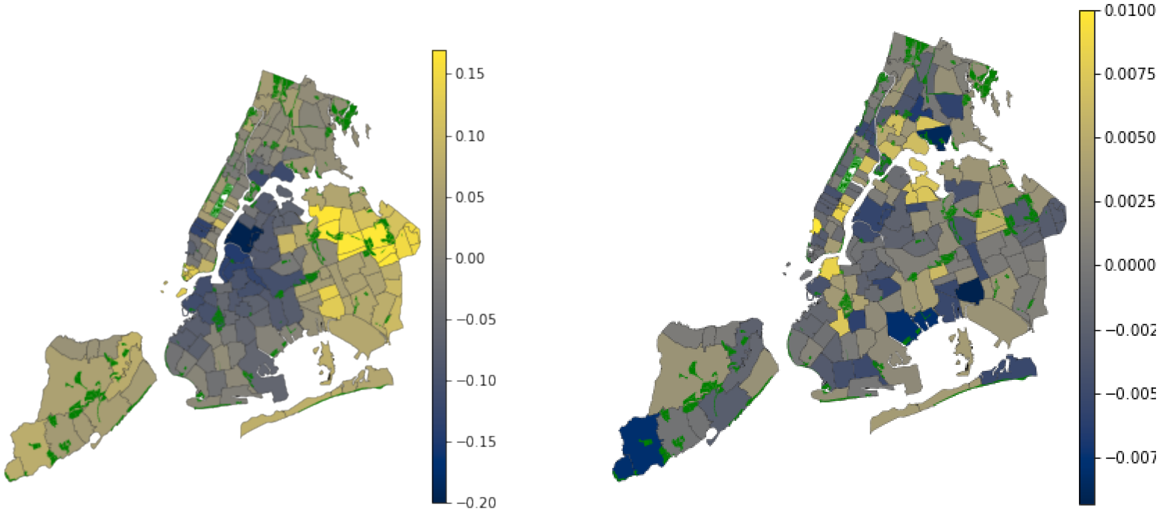
Understanding the second step of this iterative procedure is important to understand the mechanics of the model. Notice first that the observed shares must be maintained since the supply of housing is fixed in the short term. Therefore, if the observed share (supply) is higher than the share predicted by the model (demand), prices must go down. Lower prices will attract individuals until an equilibrium is reached. Analogously, if the supply of housing is lower than the demand predicted by the model, prices must increase to reach an equilibrium.

Figure 1.7 shows the geographical health inequality predicted by the model and the changes in the geographic distribution of health with the counterfactual evaluation[32]. Considering that the counterfactual increases everyone's health behavior, we could expect no changes in the distribution of health by neighborhood, instead, the figure shows that some neighborhoods become healthier while others become less healthy. In particular, some originally unhealthy neighborhoods become even more unhealthy after the application of this information campaign. The observed changes in the average health behavior of neighborhoods are, however, small, with a maximum effect of +/- 0.01 standard deviations.

Where we do see stronger effects is in equilibrium prices. Figure 1.8 shows the original prices (annual rents) by neighborhood and the changes in prices with the counterfactual evaluation. We can see that the Southwest part of Queens suffers reductions in prices of up to USD 2000 in annual rent, since this is an area that had relatively large prices but is the most distant from parks (see Figure 1.1). The North of Queens and a big part of Manhattan and the Bronx experience price increases, being areas closer to parks. On average, prices of neighborhoods within 1 Km of a park experience an increase on the annual rent of USD 402, an increase of 3% with respect to original prices. Instead, neighborhoods that are 1 Km distant from parks see a reduction in annual rents of USD 524, a 7.3% reduction; while

[32] In Figure 1.7b, I plot the differences in the averages of the original health index, without the improvement of 0.3 standard deviations.

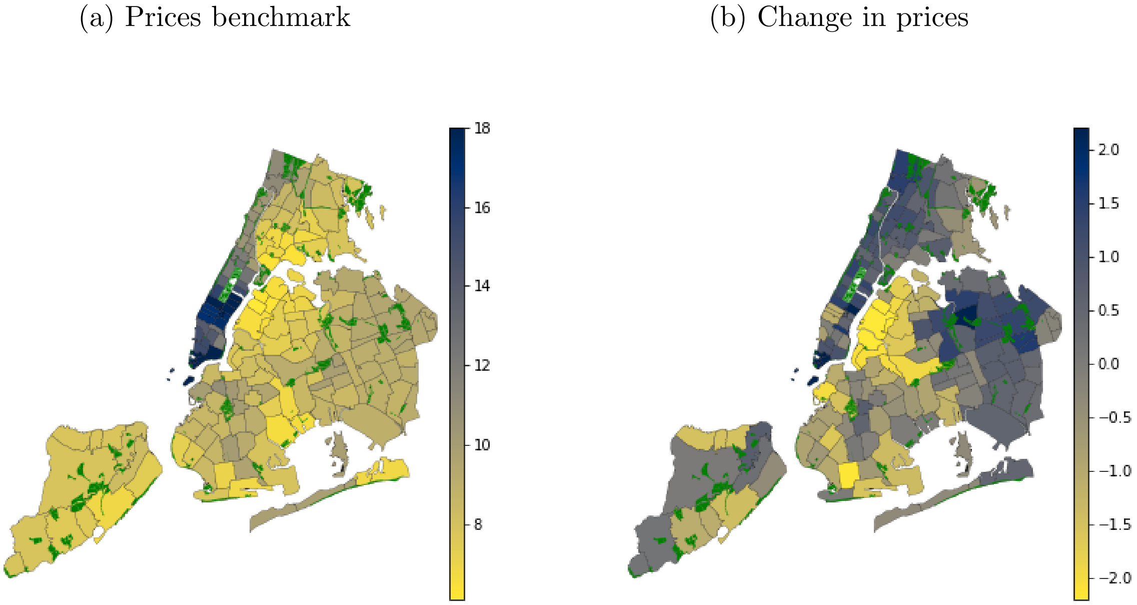
Figure 1.7: Counterfactual - Health behavior


(a) Health index benchmark

(b) Change in health index

The left plot shows the distribution of health behavior by neighborhood predicted by the model (benchmark). This is calculated by averaging the health index of all individuals assigned to each neighborhood. The right plot shows the difference between the average (original) health behavior by neighborhood after relocation due to the counterfactual evaluation and the benchmark health distribution.

Figure 1.8: Counterfactual - Prices


(a) Prices benchmark

(b) Change in prices

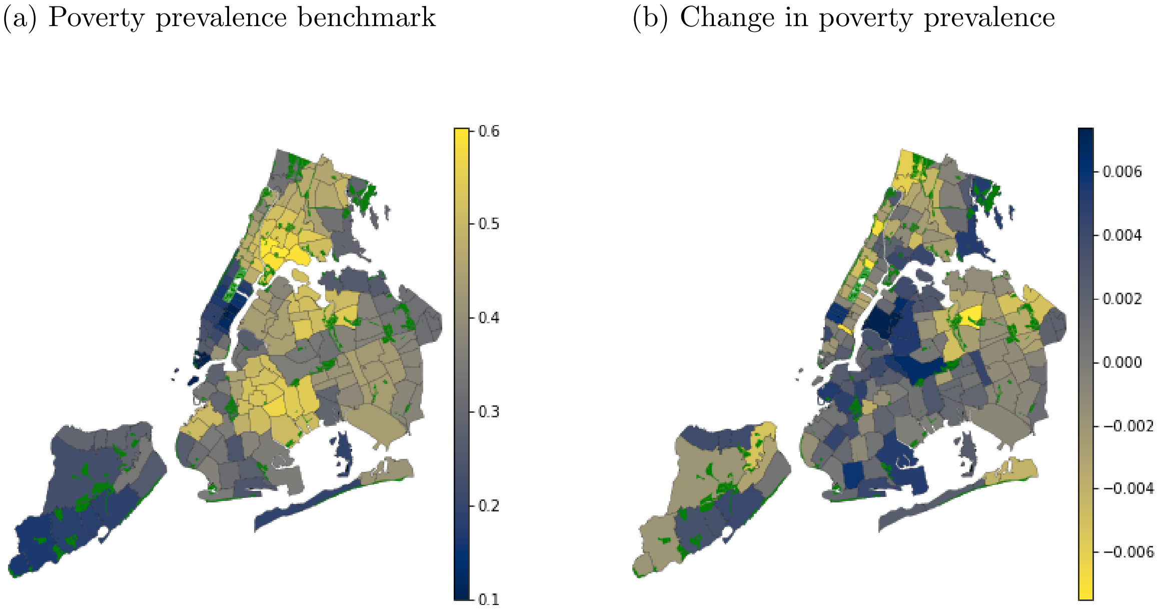
The left plot shows the observed average annual rent by neighborhood (benchmark). The right plot shows the difference between equilibrium prices from the counterfactual evaluation and benchmark prices. Both axis are in thousands of USD.

those that are more than 2 Km distant from parks experiencing a decrease of USD 1065, a 14.8% reduction with respect to original prices.

These changes in equilibrium prices are followed by changes in poverty prevalence in each neighborhood. As can be see in Figure 1.9, the neighborhoods that become cheaper attract individuals under the poverty line that are pushed away from neighborhoods that become more expensive. This displacement of low-income residents could in turn negatively affect their mental health, as was shown by Lim et al. (2017).

To summarize, a policy that can be considered successful from a public health point of view, can generate unexpected negative consequences through the interactions with the housing market. The rise in demand for neighborhoods closer to health amenities due to the health campaign, causes an increase in prices in these neighborhoods, while neighborhoods distant from health amenities suffer a reduction in prices. These cheaper neighborhoods experience an increase in poverty prevalence and a worsening of health behaviors, aggravating the initial geographical inequalities.

Figure 1.9: Counterfactual - Poverty


(a) Poverty prevalence benchmark

(b) Change in poverty prevalence

The left panel shows the distribution of poverty prevalence predicted by the model (benchmark). The right panel shows the difference between the poverty prevalence predicted in the counterfactual evaluation and benchmark poverty.

These results highlights the importance of considering the location choice of individuals to understand the general equilibrium effects of policies addressing health behavior.

## 1.7 Discussion and Conclusion

### 1.7.1 Discussion

The interpretation of my results deserves some discussion on a few points specified below:

*a) Interpretation of results*
From the estimation of this model of location choice, we can conclude that 60% or more of the observed geographical health inequality can be explained by sorting of healthy individuals in certain neighborhoods. This result must be taken into account when designing public policies with the aim of reducing these inequalities. Clearly, such policies could still be implemented after an appropriate assessment of costs and benefits.

It is important to notice that this sorting is determined by individuals' characteristics, but concluding that 60% of the observed geographical health inequality is the "natural" or "optimal" inequality since it is the result of individuals' choices is not correct. This is the observed level of inequality generated by individuals' constrained choices: constrained by their level of education, race, income and other unobservable characteristics.

Another important point to consider is that my model does not capture all the relation between neighborhood choice and health preferences, since it only uses parks as a relevant health amenity.

*b) Importance of geographic inequalities*
The results of the counterfactual evaluation show that a well-intended policy that successfully improves everyone's health behavior can have negative consequences on the geographic distribution of health and poverty prevalence. This increase in inequalities is caused by the relocation of individuals, with the increase in prices of healthy neighborhoods displacing vulnerable individuals. In the absence of neighborhood effects of any type, this clustering of vulnerable individuals would not be a problem, but this does not seem to be the case. Even if some recent research show the small role of neighborhood effects on adults, especially in health (Kling et al., 2007; Ou, 2019; Allcott et al., 2019), other research shows the importance of neighborhoods effects on children, their future outcomes and therefore on

intergenerational mobility (Damm and Dustmann, 2014; Chetty and Hendren, 2018; Chyn and Katz, 2021).

*c) Identification of preferences*

With my empirical strategy I am not able to separately identify preferences for parks as health amenities from preferences for other amenities that might endogenously generate close to parks, nor from endogenous neighborhood effects. For example, it is possible that a group of healthy people locate close to parks because they have a preference to stay near green areas. Given the demand of healthy products or services by these people, healthy stores or gyms might open in the vicinity of parks, attracting other healthy individuals that did not have particular preferences for parks. It is also possible that individuals living in these neighborhoods improve their health behavior due to the direct influence of healthy neighbors or the presence of healthy stores and services. As previously mentioned, neighborhood effects on health are likely small, but endogenous amenities play an important role in the final offer of goods and services (Diamond, 2016; Almagro and Domınguez-Iino, 2021).

Despite these limitations, I am still able to identify preferences for healthy neighborhoods, regardless if they are healthy due to closeness to parks or due to a combination of closeness to parks and other endogenous health amenities. Considering that I use a static model of location choice, the counterfactual evaluations can be interpretable after a new equilibrium is reached, which also implies an equilibrium in endogenous amenities[33].

*d) Data Limitations*

Prices and income are extremely important for the choice of neighborhood. Unfortunately, it is very difficult to have access to detailed data on health that also provides detailed information on income and geographical location. The data I use provides information on five income categories and I constructed representative house prices for these five income groups. These approximations are useful to obtain an estimate of the effect of income and house prices on location choice and can give us an approximation of the marginal willingness to pay (MWTP), which is the most standard way to present results from demand models. The calculations on MWTP presented in section 1.5 are useful to compare the relative preference for one characteristic over others, but should not be taken at face value.

[33] The implicit assumption would be that endogenous amenities are generated in the same way before than after the counterfactual.

### 1.7.2 Conclusion

This paper addresses one acknowledged but untested channel in the neighborhood-health relation. I propose and estimate a model of location choice that takes into account health amenities, proxied by parks, and health preferences, proxied by health behavior. I find that individuals that care about their health are more likely to choose neighborhoods close to health amenities, confirming a role of sorting based on health. This model is able to explain 62% of the variability in health between neighborhoods, with an improvement of more than 35% compared to a sorting model based on standard individual characteristics (e.g. income, education and race) and standard amenities. I use the model for a counterfactual evaluation that shows unintended consequences in housing prices and the distribution of poor individuals after a well-intended information campaign on health.

The fact that I find more likely that individuals with good health behaviors–that are independent from parks–locate close to parks allows me to conclude that sorting is the main driver of my findings. I do not claim, however, that the total effect I find is due exclusively to sorting. It is possible that the presence of individuals who chose to locate close to parks because they care about their health generates endogenous health amenities, such as private clinics, which might have an effect on health behaviors. Likewise, I cannot rule out the existence of peer effects but other research finding a negligible or null peer effects in health make this a less relevant concern.

First, my results contribute to the literature on neighborhood effects in health by showing that the largest part of neighborhood differences in health can be explained by individuals' sorting. Second, my model shows an important connection between individuals' health preferences and house prices. Consequently, any policy that tries to affect health in neighborhoods must take into account potential equilibrium effects in the housing market.

Improvements in data availability by providing repeated observations on health, income and geographical location will allow future work to generate further insights into the mechanics and consequences of location choice based on health preferences and health amenities.

## 1.8 Appendix

### 1.8.1 New York City Boroughs

Bronx
Manhattan
Queens
Staten Island
Brooklyn

### 1.8.2 Calculations for income and house prices

The individual-level data gives information on five income groups: from less than a 100% of the Federal Poverty Line (FPL) to more than 600% of the FPL. The 2010 poverty line guidelines of the Department of Health and Human Services gives an FPL of USD 14.500 for two-person households (the mean household size in NYC is 2.4). I consider the threshold for two-person households because the assumption that housing is paid by two household members is less strong that the assumption that it is paid by one of them. With household income data from ACS 5-year estimates for 2009-2013 (downloaded from IPUMS), I calculate the median income for each income category in the survey. The table below shows the approximations I used to estimate a yearly income in USD from the income categories.

In the health survey data, 27% of individuals declare a household income below 100% FPL,

| Category | Definition | Range | Approximate income |
|---|---|---|---|
| 1 | < 100% FPL | min – 14,499 | 9,430 |
| 2 | 100%– < 200% FPL | 14,500 – 28,999 | 21,380 |
| 3 | 200%– < 400% FPL | 29,000 – 57,999 | 42,161 |
| 4 | 400%– < 600% FPL | 58,000 – 86,999 | 71,037 |
| 5 | > 600% FPL | 87,000 – max | 136,000 |

Table A1: Definition of Health behavior variables

| Variable | Definition |
|---|---|
| Ever smoker | Smoked at least 100 cigarettes |
| Current smoker | Smokes cigarettes now if ever smoker |
| Unhealthy drinker | More than one drink on average (per day) more than 5 times a month or more than 3 drinks on average (per day) if more than twice a month |
| Unsafe sex | No condom if not with stable partner (among sexually active in last 12 months) |
| Flu vaccine | Flu vaccine (shot or spray) in the last twelve months |
| Colonoscopy | Colonoscopy performed in the last ten years for individuals of 45+ years |
| >= 1 sweetened bevrg | One or more soda or other sugar sweetened beverages consumed per day |
| <= 1 fruit/veg | One or less servings of fruit and/or vegetables per day |
| >= 4 fruit/veg | Four or more servings of fruit and/or vegetables per day |
| Exercise | Physical activity (including walking as exercise) in the last month |

Source: Codebook of the CHS and own definition of variables.

while using ACS data only 12% should belong to this category. This is probably due to underestimation of income in the health survey but I attain to it in the analysis.

To be able to compare housing expenditures with yearly income, I calculate annual rents with the price to rent ratio for each borough in NYC (from smartasset.com – data from 2015 using the Wayback Machine from the Internet Archive): Bronx 28, Brooklyn 43, Manhattan 50, Queens 30 and Staten Island 36. Considering that income is capped at USD 136.000, I also cap rents to ensure a reasonable disposable income after rent at USD 100.000.

To assign the price that each individual paid for housing, I calculated five quantiles of rents and assigned quantile 10 to individuals in the first income category, quantile 30 to individuals in the second income category, quantile 50 for the third income category, quantile 70 for the fourth and quantile 90 for the last income category.

### 1.8.3 Appendix Tables

## Table A2: Employment NYC

| | Bronx | Brooklyn | Manhattan | Queens | Staten Island |
|---|---|---|---|---|---|
| Employment (thousands) | 272 | 656 | 2277 | 585 | 106 |
| Working age pop (thousands) | 881 | 1610 | 1119 | 1436 | 293 |
| Ratio | 0.31 | 0.41 | 2.03 | 0.41 | 0.36 |

The table shows that employment in Manhattan is much larger than its working age residents, attracting a large part of the labor force from other boroughs.
Source: own elaboration using Quick Facts New York City, Census Bureau.

## Table A3: Interaction coefficients, full set

| | poor | educated | female | parent | married | age2544 | age4564 | age65m | white | black/hisp | insured | healthy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Distance CBD | -0.0042 | -0.0082 | 0.0008 | -0.0028 | 0.0029 | 0.0018 | 0.0014 | -0.001 | 0.0019 | 0.000 | 0.0052 | -0.0016 |
| | (0.0021) | (0.0023) | (0.0004) | (0.0019) | (0.0014) | (0.0022) | (0.0017) | (0.0018) | (0.0019) | (0.0021) | (0.0021) | (0.0013) |
| Distance park | -0.018 | -0.0311 | 0.0011 | -0.0164 | 0.042 | 0.019 | 0.0125 | 0.0262 | -0.0381 | -0.0554 | -0.0156 | -0.043 |
| | (0.0277) | (0.0314) | (0.0057) | (0.0235) | (0.0193) | (0.029) | (0.026) | (0.0248) | (0.0229) | (0.026) | (0.0279) | (0.0197) |
| Detached house | -0.2617 | 0.2713 | -0.0615 | -0.0726 | 0.1543 | -0.047 | -0.0217 | 0.0576 | -0.2377 | -0.0133 | -0.041 | 0.1002 |
| | (0.1267) | (0.1548) | (0.0296) | (0.1104) | (0.0804) | (0.1375) | (0.1177) | (0.1178) | (0.1399) | (0.1315) | (0.1052) | (0.0878) |
| Nr floors | -0.0824 | 0.0849 | 0.0099 | -0.0086 | -0.0321 | 0.0133 | 0.0069 | 0.0113 | -0.0077 | 0.0331 | 0.0613 | 0.033 |
| | (0.0274) | (0.0254) | (0.0047) | (0.0219) | (0.0192) | (0.023) | (0.0202) | (0.0203) | (0.0149) | (0.0205) | (0.0228) | (0.0131) |
| Building age | -0.0004 | 0.0036 | -0.0002 | -0.0019 | 0.0003 | 0.0024 | 0.0001 | -0.0008 | -0.0026 | -0.0024 | 0.0004 | -0.0015 |
| | (0.0016) | (0.002) | (0.0004) | (0.0017) | (0.0012) | (0.0019) | (0.0017) | (0.0015) | (0.0017) | (0.002) | (0.0017) | (0.0011) |
| House size | -0.0021 | 0.0021 | 0.001 | 0.0004 | -0.0008 | 0.000 | 0.001 | 0.0009 | -0.0012 | -0.0004 | 0.0013 | 0.0017 |
| | (0.0013) | (0.0015) | (0.0003) | (0.0012) | (0.001) | (0.0015) | (0.0013) | (0.0014) | (0.0012) | (0.0012) | (0.0012) | (0.0008) |
| Comm area | -0.001 | -0.0128 | -0.0081 | 0.0076 | -0.015 | 0.0178 | 0.012 | 0.0004 | -0.0104 | -0.0145 | 0.0028 | -0.0083 |
| | (0.0107) | (0.0103) | (0.0032) | (0.0113) | (0.0097) | (0.0125) | (0.0083) | (0.007) | (0.0094) | (0.0114) | (0.0083) | (0.0059) |
| Indus area | -0.1122 | -0.0903 | 0.0213 | -0.1381 | 0.2732 | 0.0458 | -0.2047 | -0.2376 | 0.1911 | 0.1871 | 0.1314 | -0.0007 |
| | (0.1217) | (0.1876) | (0.0257) | (0.1431) | (0.1539) | (0.1701) | (0.1575) | (0.131) | (0.0897) | (0.0868) | (0.0827) | (0.0719) |
| Crime rate | 0.0007 | -0.0018 | -0.0001 | -0.0025 | -0.0007 | -0.0005 | -0.0008 | -0.0003 | -0.0009 | -0.0004 | 0.0003 | -0.0007 |
| | (0.0011) | (0.0014) | (0.0003) | (0.0011) | (0.0009) | (0.0012) | (0.0012) | (0.0011) | (0.0015) | (0.0017) | (0.0015) | (0.0009) |
| Pct white | -0.2796 | 0.4962 | -0.0408 | 0.174 | -0.1218 | -0.3013 | -0.3556 | -0.3434 | 3.0513 | 1.9564 | 0.3624 | -0.1898 |
| | (0.1214) | (0.1554) | (0.0395) | (0.1338) | (0.1109) | (0.1583) | (0.1439) | (0.1325) | (0.113) | (0.1547) | (0.0953) | (0.0851) |
| Pct bl/hisp | -0.2156 | 0.2973 | -0.0829 | 0.2213 | -0.3743 | -0.0475 | 0.0087 | -0.0931 | 1.717 | 3.4987 | 0.3887 | -0.0569 |
| | (0.1093) | (0.1421) | (0.0383) | (0.1189) | (0.1002) | (0.143) | (0.1334) | (0.1244) | (0.1353) | (0.157) | (0.1018) | (0.0734) |

All interaction coefficients ($\beta_1$) from the main specification of the model. Standard errors in parentheses, clustered by individual.

Table A4: Instrumenting for price

| | (p 10) rent | (p 30) rent | (p 50) rent | (p 70) rent | (p 90) rent |
|---|---|---|---|---|---|
| Detached houses | 1688.449 | 2026.744 | 2750.045 | 3420.962 | 2105.363 |
| | (1047.96) | (1782.96) | (2611.97) | (3925.84) | (7684.94) |
| Number of floors | 730.177*** | 1220.401*** | 1844.874*** | 3164.600*** | 7128.164*** |
| | (101.66) | (172.97) | (253.39) | (380.85) | (745.53) |
| Building age | 26.669 | 50.502* | 85.234** | 147.157** | 317.806** |
| | (16.67) | (28.35) | (41.54) | (62.43) | (122.21) |
| House size ($m^2$) | 29.669* | 46.153* | 76.689** | 135.981** | 303.938*** |
| | (15.45) | (26.28) | (38.50) | (57.86) | (113.26) |
| Constant | -3850.267 | -6498.425 | -1.21e+04** | -2.40e+04*** | -5.98e+04*** |
| | (2420.09) | (4117.46) | (6031.93) | (9066.11) | (17747.17) |
| Obs. | 127 | 127 | 127 | 127 | 127 |
| $R^2$ | 0.35 | 0.36 | 0.38 | 0.45 | 0.54 |
| F-test | 16.07 | 16.95 | 18.69 | 25.18 | 35.74 |

The explanatory variables (instruments) are exogenous characteristics of adjacent neighborhoods. For example, rent of neighborhood A is regressed on the average number of floors of all the neighborhoods that share a border with A. Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: mean MWTP for one SD change (in USD)

| | Std. dev. | Typical | Healthy |
|---|---|---|---|
| Distance CBD | 20.3 | -36085 | -43527 |
| Distance Park | 0.7 | -6928 | -15461 |
| Detached houses | 0.2 | 22836 | 30250 |
| Number of floors | 1.8 | -2423 | 13530 |
| Building age | 13.7 | 41621 | 35545 |
| House size | 17.8 | 21968 | 33082 |
| Commercial area | 2.0 | 46687 | 42033 |
| Industrial area | 0.1 | -759 | -2846 |
| Crime rate | 16.1 | -67348 | -71569 |
| Percentage white | 0.3 | 145363 | 133831 |
| Percentage black/hisp | 0.3 | 1955 | -10697 |

Marginal Willingness to Pay for a change of one standard deviation in each neighborhood characteristic at the time for a *Typical* individual (non-poor, working-age, college-educated) and a typical individual who is also *Healthy* (upper half of the health index distribution).

Table A6: Robustness to the sample:
Selected interaction coefficients excluding one borough at the time

| | Bronx | Brooklyn | Manhattan | Queens | St Island |
|---|---|---|---|---|---|
| Distance CBD | 0.0000 | -0.0030 | -0.0007 | -0.0019 | 0.0001 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Distance park | -0.0356 | -0.0757 | -0.0373 | -0.0096 | -0.0417 |
| | (0.0008) | (0.0010) | (0.0009) | (0.0011) | (0.0008) |
| Detached house | 0.1188 | 0.2950 | 0.1264 | 0.1357 | 0.1260 |
| | (0.0038) | (0.0052) | (0.0036) | (0.0057) | (0.0036) |
| Number floors | 0.0349 | 0.0423 | -0.0019 | 0.0192 | 0.0367 |
| | (0.0006) | (0.0007) | (0.0025) | (0.0007) | (0.0006) |
| Comm area | -0.0059 | -0.0118 | 0.0139 | -0.0086 | -0.0077 |
| | (0.0004) | (0.0004) | (0.0008) | (0.0004) | (0.0004) |
| Indus area | -0.0314 | 0.1075 | -0.0894 | -0.1121 | -0.0071 |
| | (0.0040) | (0.0044) | (0.0046) | (0.0065) | (0.0039) |
| Crime rate | -0.0014 | -0.0009 | -0.0002 | -0.0001 | -0.0005 |
| | (0.0001) | (0.0000) | (0.0000) | (0.0001) | (0.0000) |

Standard errors in parentheses.

Table A7: Robustness to variable specification:
Selected interaction coefficients with alternatives to main variables

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Health index 1 | Health index 3 | Log dPark | Park area | Park investment | Park crime |
| Distance CBD | -0.0014 | 0.0018 | -0.0015 | -0.0012 | -0.0012 | -0.0012 |
| | (0.0011) | (0.0010) | (0.0011) | (0.0011) | (0.0011) | (0.0012) |
| Distance park | -0.0350 | -0.0182 | -0.0380 | | -0.1601 | -0.0109 |
| | (0.0143) | (0.0134) | (0.0135) | | (0.0789) | (0.0623) |
| Park area | | | | 0.0124 | | |
| | | | | (0.0101) | | |
| Detached house | 0.1249 | -0.0357 | 0.0956 | 0.0845 | 0.0977 | 0.1067 |
| | (0.0669) | (0.0643) | (0.0665) | (0.0666) | (0.0666) | (0.0591) |
| Nr floors | 0.0332 | 0.0037 | 0.0331 | 0.0387 | 0.0354 | 0.0338 |
| | (0.0120) | (0.0110) | (0.0111) | (0.0110) | (0.0111) | (0.0022) |
| House size | 0.0008 | 0.0011 | 0.0018 | 0.0017 | 0.0018 | 0.0021 |
| | (0.0008) | (0.0007) | (0.0008) | (0.0008) | (0.0008) | (0.0010) |
| Comm area | -0.0034 | 0.0058 | -0.0083 | -0.0123 | -0.0099 | -0.0061 |
| | (0.0066) | (0.0060) | (0.0063) | (0.0062) | (0.0062) | (0.0053) |
| Indus area | 0.0027 | 0.0077 | -0.0279 | -0.0433 | -0.0273 | 0.0205 |
| | (0.0740) | (0.0707) | (0.0757) | (0.0755) | (0.0759) | (0.0042) |
| Crime rate | -0.0013 | -0.0004 | -0.0006 | -0.0001 | -0.0004 | -0.001 |
| | (0.0007) | (0.0006) | (0.0007) | (0.0007) | (0.0007) | (0.0002) |

Estimation coefficients by replacing the main variables of the analysis with an alternative. In Column 1 Health index 2 is replaced with Health index 1, while in Column 2 it is replaced by an index containing only the variables *ever smoker* and *unsafe sex.* In column 3 distance to the park is replaced with its logarithm, while in 4 it is replaced with park area by neighborhood. In columns 5 and 6 distance to the park is weighted by the inverse of the investment in each park, and by the crime rate in each park respectively. Standard errors in parenthesis, with clustering at the individual level.

Table A8: Robustness on income:
Selected interaction coefficients excluding income categories and $\beta_i^{DI}$

| | cat 1 | cat 2 | cat 3 | cat 4 | cat 5 | $\beta_i^{DI}$ |
|---|---|---|---|---|---|---|
| Distance CBD | -0.0009 | -0.0008 | -0.0008 | -0.0001 | -0.0007 | -0.0003 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Distance park | -0.0407 | -0.0339 | -0.0499 | -0.0399 | -0.0367 | -0.1229 |
| | (0.0009) | (0.0009) | (0.0009) | (0.0008) | (0.0008) | (0.0042) |
| Detached house | 0.1827 | 0.1678 | 0.0992 | 0.1125 | 0.1222 | 0.1299 |
| | (0.0041) | (0.0041) | (0.0040) | (0.0039) | (0.0038) | (0.0036) |
| Number floors | 0.0359 | 0.0361 | 0.0286 | 0.0416 | 0.0094 | 0.0350 |
| | (0.0007) | (0.0007) | (0.0007) | (0.0007) | (0.0008) | (0.0007) |
| Building age | -0.0019 | -0.0018 | -0.0017 | -0.0015 | -0.0009 | -0.0012 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Comm area | -0.0077 | -0.0090 | -0.0079 | -0.0110 | 0.0070 | -0.0092 |
| | (0.0004) | (0.0004) | (0.0004) | (0.0004) | (0.0004) | (0.0004) |
| Indus area | -0.0519 | -0.1009 | -0.0156 | 0.0829 | -0.0626 | -0.0652 |
| | (0.0045) | (0.0044) | (0.0044) | (0.0043) | (0.0043) | (0.0038) |
| Crime rate | -0.0009 | -0.0008 | -0.0009 | -0.0008 | -0.0007 | -0.0006 |
| | (0.0000) | (0.0000) | (0.0000) | (0.0000) | (0.0000) | (0.0000) |
| Percentage white | -0.0840 | -0.1172 | -0.1452 | -0.1431 | -0.1256 | -0.1398 |
| | (0.0055) | (0.0055) | (0.0053) | (0.0052) | (0.0052) | (0.0048) |
| Percentage bl/hisp | 0.0566 | -0.0156 | -0.0623 | -0.0132 | 0.0179 | -0.0248 |
| | (0.0052) | (0.0051) | (0.0050) | (0.0048) | (0.0048) | (0.0045) |

The first 5 columns shows relevant $\beta_1$ coefficients excluding the column income category. The last column instead shows these coefficients allowing for heterogeneous response to disposable income ($\beta_i^{DI}$ instead of $\beta^{DI}$). Standard errors in parentheses.

# Chapter 2

# Information, Beliefs and Social Networks as determinants of Health Behavior

*with Jérôme Adda*

JEL Code: D01, D83, I12

## 2.1 Introduction

The Surgeon General report in 1964 on the harmfulness of smoking was widely covered by TV channels and newspapers, and further backed by medical information and press coverage. Yet, by 1970, still 30% of the adult population did not believe that smoking causes lung cancer. By 2016 only a small fraction of the population seems unaware of the risk but there are nonetheless differences by socioeconomic groups: the percentage of low educated individuals of 65 years of age or more who did not believe that smoking causes lung cancer was five times that of the high educated individuals between 25 and 35 years of age. Beyond the case of tobacco, there are many health issues that rely on individuals forming correct beliefs about their actions with important spillovers for the general population, such as vaccination against communicable diseases, diet or even driving.

The aim of this paper is to study the long term evolution of health behavior together with the formation and propagation of beliefs across socioeconomic groups. Better understanding this complex process, allow policy makers to optimally target health information and health interventions to various sub-populations, highlighting that it is not necessarily optimal to direct institutional health messages to groups that are defiant of authorities. We therefore explore alternative policies.

Our paper contributes to this debate in two important ways. First, we assemble disparate information on beliefs about the danger of tobacco that cover several decades and different socioeconomic groups. This information has never been analyzed together, nor in a coherent framework that allows for policy evaluation. Second, we develop a dynamic and dynastic model of tobacco consumption, beliefs and competing causes of mortality. The model follows individuals over their life-cycle and considers many overlapping generations, covering the 20th century. We consider three mechanisms through which an individual can learn about the consequences of smoking. The first mechanism is through public information, consisting of medical research through scientific publications, public health messages, or opposing messages from the tobacco industry. We allow for the possibility that individuals may put different weights on this information, depending on their socioeconomic group. The second one is introspection: beliefs on the harmfulness of smoking may change when one receives a smoking-related health shock, and conversely beliefs could be affected if the agent does not receive these signals as they age. An important factor is the age at which tobacco related diseases such as lung cancers occur, relative to other health shocks. Individuals who are at a higher risk of mortality through other channels than tobacco may not live long enough to learn the effects of tobacco. This is more likely to be the case for older birth cohorts and for individuals belonging to low economic status groups. The third mechanism and the main contribution of this paper is social learning where agents learn from information sharing within their social networks. The latter mechanism seems consistent with the gradual evolution of beliefs and with the observed differences by socioeconomic groups.

In our model, the agents update beliefs from introspection using Bayes rule. In social learning, however, they are less sophisticated, and behave as DeGroot agents, averaging the beliefs of the individuals with whom they interact. With this setting of information aggregation, individuals in a group that communicate more among themselves than with individuals of other groups can remain stuck with the wrong belief (Chandrasekhar et al., 2020). The model takes into account the heterogeneity in the taste for tobacco in the population and varying cigarette prices, which allows us to compare policies based on tax increases to those on health information.

In this preliminary draft we explore the long term effects of medical information and find that if there had been no dissemination of medical information we would have observed a much slower learning rate in the period 1960-2000. This would have primarily affected the cohort born around 1945. Smoking prevalence in this cohort would have been almost 10% higher (an increase of 25%) and 44% of the smokers in this cohort would experience a smoking related health shock by the end of their lives, an increase of 8% compared to the baseline scenario. Beliefs about the harmfulness of smoking would nevertheless converge with baseline values around the year 2060 due to the general increase in life expectancy (which increases the role of introspection) and the role of social learning.

This paper contributes to the literature addressing the effects of medical information on smoking prevalence. Initial work by Hamilton (1972) and Warner (1981, 1989) found that anti-smoking campaigns in the second half of the 20th century had important negative effects on cigarette consumption. Sloan et al. (2002), on the other hand, found that US per capita cigarette demand changed *before* information about health effects of smoking was widely distributed, and attributed the reduction to the elimination of free cigarettes to soldiers at the end of World War II. All of these papers do time series regressions with aggregate data. De Walque (2010) went one step further to understand the role of education in the response to information campaigns on smoking harms. He found that more educated individuals responded faster to information on the dangers of smoking.

Our paper also contributes to the literature studying cigarette smoking through the effects of risk perception (Viscusi, 1990; Liu and Hsieh, 1995; Viscusi et al., 1999); bounded rationality (Strulik, 2018), learning from own's health shocks or biomarkers (Smith et al., 2001; Darden, 2017), and from health shocks to close family members (Clark and Etilé, 2001; Darden and Gilleskie, 2016). Our model considers all these mechanisms and incorporates social learning.

Finally, our paper contributes to the empirical literature on social learning and health such as Sorensen (2006), Adhvaryu (2014), Hoffmann (2017) and Barili et al. (2021) who have looked at the role of social sharing of information for the choice of health plan, antimalarial treatment, healthcare utilization and cesarean sections respectively. We contribute by analyzing how social learning affects health behavior.

## 2.2 Empirical evidence

In this section we describe the data sources and we show the evolution of smoking prevalence, beliefs and mortality. This evidence provides support to our hypotheses on the mechanisms through which information had a role in determining the reduction of smoking prevalence observed in the last decades. In subsection 2.2.6 we present econometric results on the introspection mechanism: we show that individuals are more likely to quit smoking after receiving a smoking-related health shock.

### 2.2.1 Data

We have gathered data on smoking, health, mortality, beliefs on the harmfulness of smoking, income, education, cigarette prices, medical information, trust in the medical profession and data on close social networks. We have individual level data from cross-sectional surveys and panels; administrative price data and historical aggregate data. Specifically:

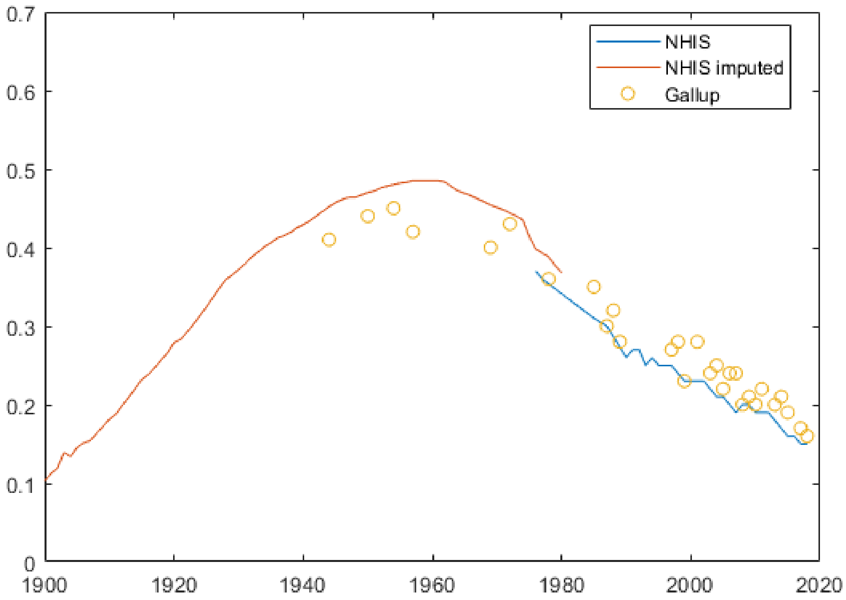
- Demographic composition
  - IPUMS USA, 1940-2018, cross-sectional data.
- Smoking
  - National Health Interview Survey (NHIS), 1970-2018, cross-sectional data.
- Health shocks
  - Health and Retirement Study (HRS), 1992-2018, panel data. Includes restricted data on cancer site.
  - National Health and Nutrition Examination Survey (NHANES), 1966-2016, cross-sectional data.
- Mortality
  - CDC death rates, 1900-1998.
  - wonder CDC death rates, 1999-2018.
  - Surgeon General Report (SGR), 2014, aggregate data on Smoking Attributable Mortality, 1965-2014.

- Beliefs
  - Gallup US poll, 1954-2018, aggregate data for given years.
  - Monitoring the Future (MTF), 1980-2018, aggregate data on a panel of young individuals.
  - Teenage Attitudes and Behavior Concerning Tobacco (TABT), 1992, cross-sectional data on adolescents.
  - Annenberg Tobacco Risk Study (ATRS), 1999, cross-sectional data on adolescents.
  - Schulman, Ronca & Bucuvalas Inc., 2000 (SRBI), cross-sectional data.
  - FFRISP 2009 (See Krosnick et al., 2017), cross-sectional data.
  - Health Information National Trends Surveys (HINTS), 2003, 2005, 2015, 2017, cross-sectional data.
  - Population Assessment of Tobacco and Health (PATH), 2013-2016, panel data.
- Trust
  - Health Information National Trends Surveys (HINTS), 2011, cross-sectional data.
- Social networks
  - General Social Survey (GSS), 1985 and 2004, cross-sectional
- Medical information
  - Web of Science (scientific publications), 1940-2018, time series.
  - ProQuest (newspaper articles), 1930-2012, time series.
- Cigarette prices
  - The Tax Burden of Tobacco (TBT), 1955-2016, panel data.

### 2.2.2 Smoking prevalence

In Figure 2.1 we show the percentage of current smokers since 1900. Data on smoking from NHIS starts in 1970. Using information on age at which smokers start and quit, we constructed smoking prevalence for years before 1970. The graph also includes some data points from Gallup survey, not so distant from our estimation.

Figure 2.1: Percentage of Current Smokers



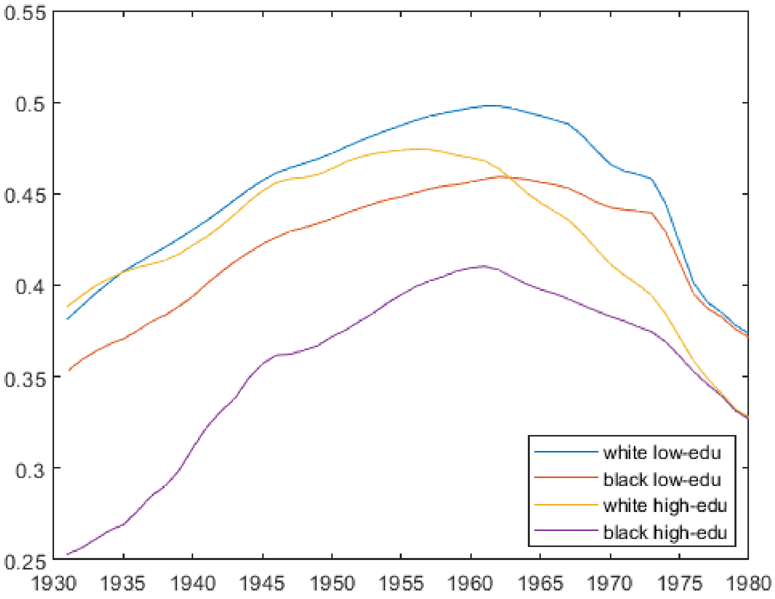
Source: NHIS, Gallup and own estimation from NHIS start and quit date

According to our estimates, smoking had a peak in 1961, when 48.5% of US adults smoked. This peak occurred a couple of years before the surgeon general report on the harmfulness of smoking, which came out after some years of scientific research on this topic.

As we can see in Figure 2.2, however, there were important differences in smoking prevalence by socioeconomic group and on the speed with which each group reduced smoking.[1] The figure shows the percentage of smokers for black and white individuals with low or high education, defined as having more than high school. Before 1960, whites of both levels of education had a similar smoking prevalence. High educated whites, however, reached a peak in smoking in 1956, some years before the other groups. Also, smoking prevalence reduced sharply for high educated whites. Blacks with low levels of education reached a peak in 1962 and did not reduce their smoking as sharply as the other groups. These differences in smoking prevalence are still present today: data from NHIS shows that 23% of individuals

[1] We omit the group latino/asian since they represented less than 4% of the US population until 1970.

Figure 2.2: Percentage of Current Smokers by group



Source: Own estimation from NHIS start and quit date

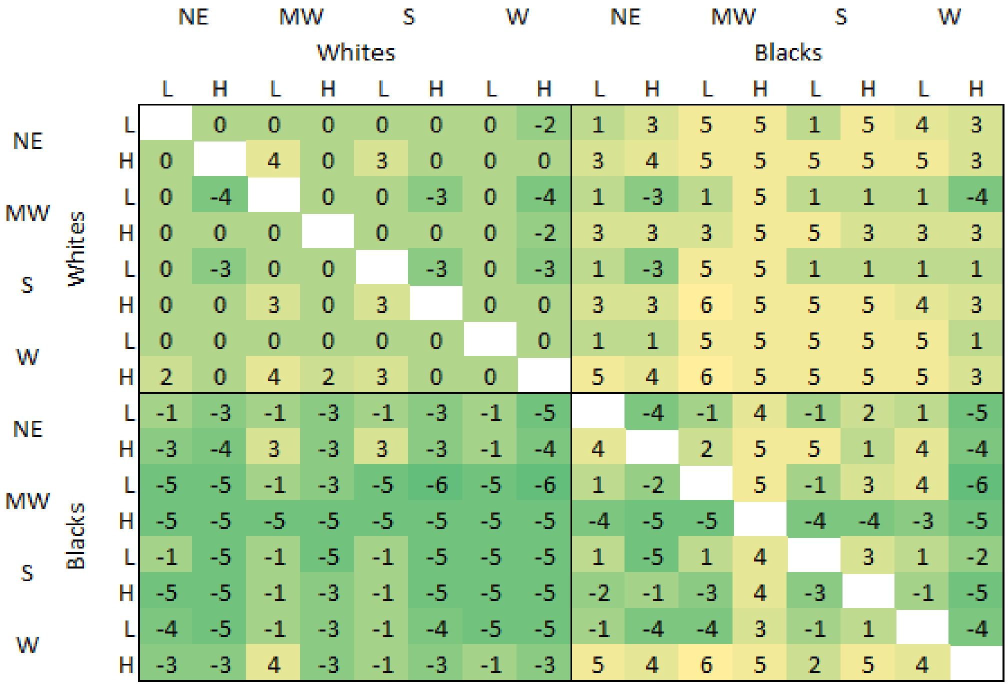
who are black, with less than 35 years of age, with at most high school education and living in the Midwest smoked in 2016, while only 12% of white individuals of the same age, more than high school and living in the Northeast smoked in the same year.

To characterize better how smoking evolves across socio-economic groups, we regress annual smoking rates of group $i$ on lags or leads of the rate for group $j$:

$$x_{i,t} = \beta_{i,j,k} x_{j,t+k} + u_t \qquad k \in \{-12, \ldots, 12\} \tag{2.1}$$

We then pick the lag $k^*_{i,j}$ with the highest $R^2$ for any pair $\{i, j\}$. The optimal lags are displayed in the matrix of Figure 2.3. The lower triangle is the mirror of the upper triangle. A positive (negative) number implies that the column group is behind (follows) the row group. For example, we find that low educated whites in the Midwest are 4 years behind high educated whites in the Northeast. The figure is color coded for better reading. A yellow color indicates that the group labelled in the column follows the one labelled in the row. A darker green indicates a group that is leading.

Figure 2.3: Optimal lag



| | | | NE | | MW | | S | | W | | NE | | MW | | S | | W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Whites | | | | | | | | Blacks | | | | | | | |
| | | | L | H | L | H | L | H | L | H | L | H | L | H | L | H | L | H |
| Whites | NE | L | | 0 | 0 | 0 | 0 | 0 | 0 | -2 | 1 | 3 | 5 | 5 | 1 | 5 | 4 | 3 |
| | | H | 0 | | 4 | 0 | 3 | 0 | 0 | 0 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | 3 |
| | MW | L | 0 | -4 | | 0 | 0 | -3 | 0 | -4 | 1 | -3 | 1 | 5 | 1 | 1 | 1 | -4 |
| | | H | 0 | 0 | 0 | | 0 | 0 | 0 | -2 | 3 | 3 | 3 | 5 | 5 | 3 | 3 | 3 |
| | S | L | 0 | -3 | 0 | 0 | | -3 | 0 | -3 | 1 | -3 | 5 | 5 | 1 | 1 | 1 | 1 |
| | | H | 0 | 0 | 3 | 0 | 3 | | 0 | 0 | 3 | 3 | 6 | 5 | 5 | 5 | 4 | 3 |
| | W | L | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 | 1 | 5 | 5 | 5 | 5 | 5 | 1 |
| | | H | 2 | 0 | 4 | 2 | 3 | 0 | 0 | | 5 | 4 | 6 | 5 | 5 | 5 | 5 | 3 |
| Blacks | NE | L | -1 | -3 | -1 | -3 | -1 | -3 | -1 | -5 | | -4 | -1 | 4 | -1 | 2 | 1 | -5 |
| | | H | -3 | -4 | 3 | -3 | 3 | -3 | -1 | -4 | 4 | | 2 | 5 | 5 | 1 | 4 | -4 |
| | MW | L | -5 | -5 | -1 | -3 | -5 | -6 | -5 | -6 | 1 | -2 | | 5 | -1 | 3 | 4 | -6 |
| | | H | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -4 | -5 | -5 | | -4 | -4 | -3 | -5 |
| | S | L | -1 | -5 | -1 | -5 | -1 | -5 | -5 | -5 | 1 | -5 | 1 | 4 | | 3 | 1 | -2 |
| | | H | -5 | -5 | -1 | -3 | -1 | -5 | -5 | -5 | -2 | -1 | -3 | 4 | -3 | | -1 | -5 |
| | W | L | -4 | -5 | -1 | -3 | -1 | -4 | -5 | -5 | -1 | -4 | -4 | 3 | -1 | 1 | | -4 |
| | | H | -3 | -3 | 4 | -3 | -1 | -3 | -1 | -3 | 5 | 4 | 6 | 5 | 2 | 5 | 4 | |

Source: Own estimation from NHIS current smoking, start and quit date

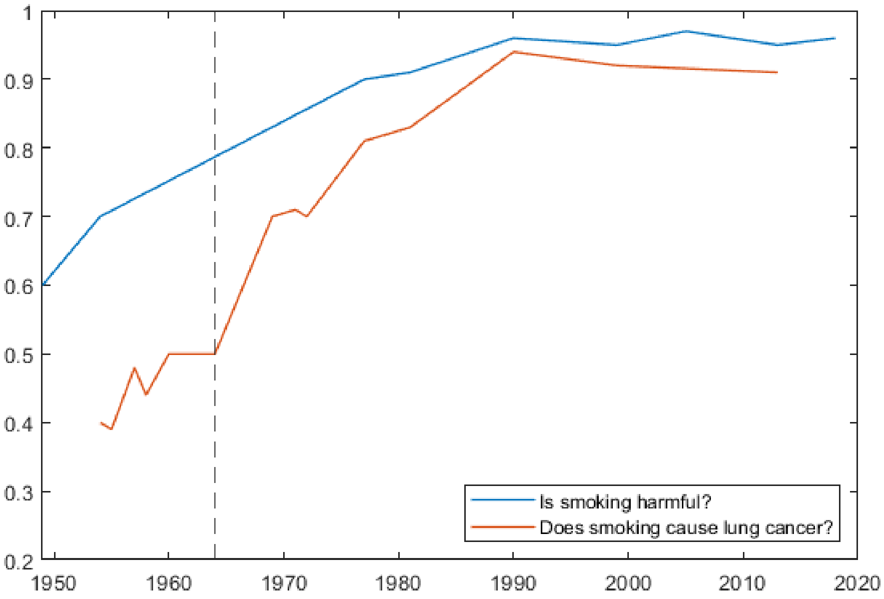
From the prevalence of green in the first and third quadrant and of yellow in the second and fourth quadrant we can see that blacks follow while whites lead. We also observe greener columns for high educated individuals and it seems that Northeast and West, so coast regions, are also leaders with respect to the Midwest and the South.

With this evidence we might expect high educated whites in the coasts to be the first group in adopting healthy behaviors and all the other groups to follow some years after.

### 2.2.3 Beliefs

Figure 2.4 shows the beliefs on the harmfulness of smoking measured by Gallup in different years since 1949. The first survey question was: *"In general, how harmful do you feel smoking is to adults who smoke"*. The blue line is the percentage of those who responded *Very harmful* or *Somewhat harmful*. The second question was *"Do you think cigarette smoking is one of the causes of lung cancer?"*. The orange line is the percentage of those who said *Yes/True.*

Figure 2.4: Beliefs on the harmfulness of smoking



Source: Gallup US poll

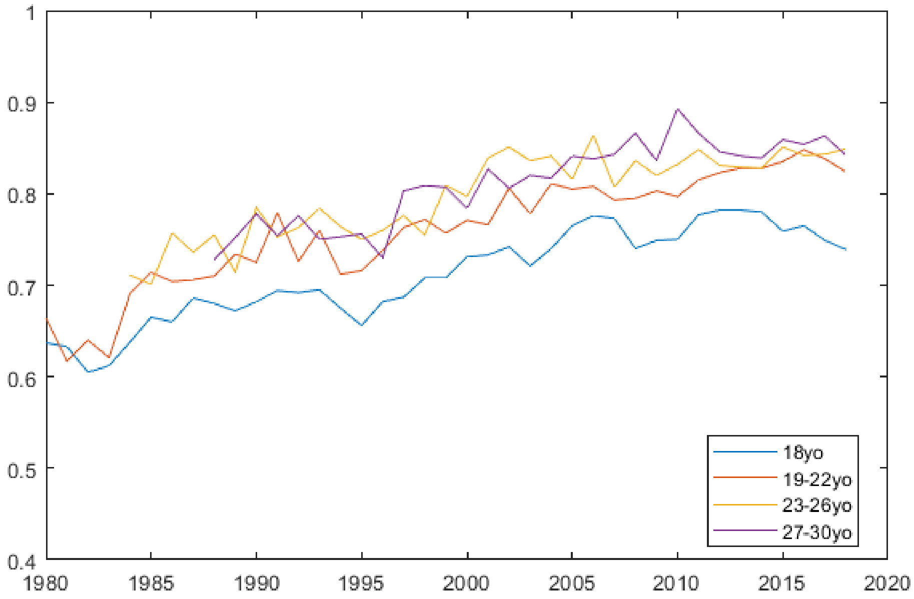
We can see that already in 1950, the first time that beliefs over the harmfulness of smoking were asked to a representative sample of the population, 60% believed that smoking was harmful. This belief increased steadily until 1990 and it has remained above 95% ever since. The awareness of the correlation among smoking and lung cancer, on the other hand, increased sharply after the surgeon general report in 1964 and gradually since 1975.

What is interesting to note is that still in the 2000s not everybody thinks that smoking is one of the causes of lung cancer or that smoking is harmful. Indeed, in 2007, 6% of respondents thought that smoking was not harmful to one's health, while they all agreed that being obese is harmful to one's health.[2]

As with smoking prevalence, these beliefs are not homogeneous by socioeconomic group nor

[2] https://news.gallup.com/poll/28177/Americans-Put-Obesity-par-Smoking-Terms-Harmful-Effects.aspx

Figure 2.5: Beliefs by age group



Source: Monitoring the Future

geographically. Survey data from PATH 2016 shows that 99% of individuals who are white, from 25 to 34 years of age, with more than high school and living in the Northeast, agree that smoking can cause lung cancer. This agreement is only 87% for individuals who are black, between 55 and 64 years of age, with at most high school education living in the Midwest. We hypothesize that these differences in beliefs among socioeconomic groups were more important in previous decades and the estimation of the model presented in section 2.3 will help us determine that.

Figure 2.5 shows the evolution of beliefs for different age groups in the period 1980-2018. Each line represents the percentage of individuals who believe that smoking one or more cigarette packs per day entails great risk. Eighteen-year-olds show lower perceived risk than individuals in other age groups even if the percentage of individuals of this age group that believe smoking entails high risk has increased by almost 10 percentage points since 1980. This figure shows that as individuals mature, they update their beliefs about the harmfulness of smoking.

In total we were able to gather data on beliefs at the individual level from six different surveys (PATH, HINTS, TABT, ATRS, SRBI, FFRISP), adding up to more than 130.000 observations for the period 1992-2017. Since there were several questions addressing the beliefs on the harmfulness of smoking in each survey and the questions differed between all surveys, we performed factor analysis in order to have one variable summarizing all the information on beliefs for each survey. Then, these factors were standardized and a single factor was created averaging individual factors for years with more than one survey.

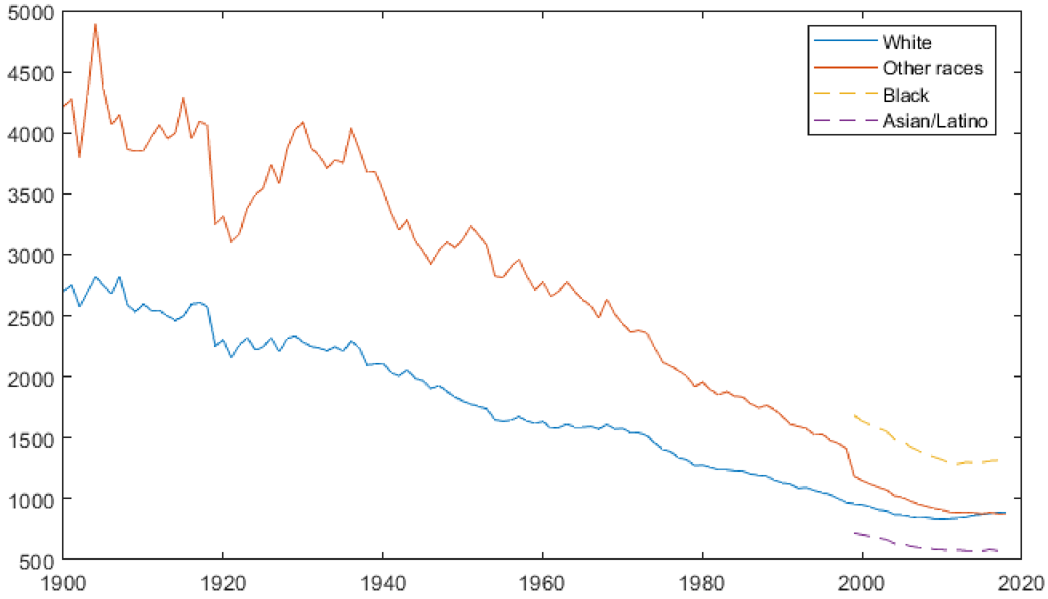
The evolution of beliefs both over the life-cycle, across birth cohorts and socio-economic groups are all features that our model detailed below addresses. We use the Gallup data in Figure 2.4 for time trends, the MTF data of Figure 2.5 to capture the evolution of beliefs in the life cycle, and the factor of beliefs to capture differences by socioeconomic groups.

### 2.2.4 Death rates

Here we present some evidence on the evolution of death rates. Figure 2.6 shows death rates for individuals from 55 to 64 years old since 1900. There was a big gap between whites and other races which gradually decreased. Until 1950, "other races" was almost exclusively composed by blacks. From 1950 onward, increasing migration to the US changed the composition of "other races" with Latinos and Asians representing 50% of this group by 1990. The figure shows that in 2018 there is no difference between the death rates of whites and those of other races but this is the combination of blacks, with a higher death rate, and Hispanics and Asians with a lower death rate.

The secular decrease in the death rates took place despite very different smoking patterns for this age group over that period and is reflecting among other things, better living standards and improved medical care. From the literature, the mean age of onset of lung cancer is around 70 (CDC, 2021). The differences in death rates imply that non Whites are more likely to experience negative health shocks from smoking than individuals of other races at any point in time as this group is more likely to experience premature deaths from other causes. Within a race group, earlier cohort are less likely to experience tobacco related health issues. This is one of the mechanisms through which learning the harmfulness of smoking could be slower in previous decades and slower for non Whites. We will model this through a competing risk framework, where in the beginning of the 20th century individuals were more likely to receive a nonsmoking-related health shock at younger ages while in the 21st century, when certain diseases (such as infectious ones) are better controlled, people are more likely

Figure 2.6: Death rates by 100.000, by race - Ages 55-64



Source: CDC (1900-2018) and wonder (1999-2018)

to receive a smoking-related health shock.

### 2.2.5 Social Networks

Data for social networks comes from the General Social Survey, a sociological survey which included questions on networks of trust including in 1985, and 2004 the contact's age, race and level of education. The wording of the question was *"From time to time, most people discuss important matters with other people. Looking back over the last six months - who are the people with whom you discussed matters important to you?"*. Table 2.1 shows that homophily is quite important: for low educated white individuals, 71% of their contacts are also low educated and 95% of their contacts are white. For high educated black individuals, 55% of their contacts are high educated and 88% are black. The race category *Other* shows a less segregated social network.

Table 2.1: Social Networks by education and race

| | Low educated | | | High educated | | |
|---|---|---|---|---|---|---|
| | White | Black | Other | White | Black | Other |
| % Contacts low-edu | 71 | 74 | 75 | 32 | 45 | 40 |
| % Contacts high-edu | 29 | 26 | 25 | 68 | 55 | 60 |
| % Contacts white | 95 | 9 | 16 | 96 | 12 | 20 |
| % Contacts black | 0 | 87 | 7 | 1 | 88 | 6 |
| % Contacts other race | 4 | 5 | 77 | 3 | 0 | 74 |
| Observations | 707 | 75 | 26 | 531 | 42 | 14 |

### 2.2.6 Responses to a Health Shock

In this subsection, we present results on how receiving a smoking-related health shock affects the likelihood of quitting smoking. Acquiring information about the harmfulness of smoking through individual experience is one of the three mechanisms that we consider. The idea is that individuals interpret a smoking-related health shock as a signal that smoking is harmful and therefore we can expect them to quit smoking after updating their beliefs.

The analysis is performed with US data from the Health and Retirement Study, a panel of individuals of 50 years of age or older and their spouses on health and longevity. The study had its first wave in 1992, with subsequent waves every two years. We got access to restricted data on cancer site to be able to differentiate individuals who get smoking-related cancer from individuals who get other types of cancer. This data covers the period 1992-2016.

Out of ever smokers, 14% (almost 3.500 individuals) quit smoking between waves and we can see whether they did so after receiving a smoking-related health shock.

Following Smith et al. (2001), we define an indicator for receiving a smoking-related health shock if the individual reports one of the following conditions (without previous history of that condition):

- Heart attack
- Congestive heart failure (if it required hospitalization)
- Stroke
- Chronic lung disease (if it limits the ability to do household chores)
- Smoking-related cancers (bladder, cervix, esophagus, larynx, lip, lung, mouth, pan-

creas, throat)

A nonsmoking-related health shock is defined as reporting one of the following:

- Diabetes[3]
- Broken hip
- Joint replacement
- Nonsmoking-related cancers

Even if the classification of health shocks is arguable since smoking can be considered a risk factor for most of them, Table 2.2 shows that ever smokers are more likely to receive a smoking-related health shock than never smokers, and the likelihood increases when individuals smoke for many years. This *signal*, however, is a noisy one since also never smokers can get a smoking-related health shock. The second column of Table 2.2 shows that the likelihood of receiving a nonsmoking-related health shock does not depend on being a smoker.

Table 2.2: Prevalence of health shocks

| | Smoking shock | Non-smoking shock |
|---|---|---|
| Never smokers | 0.21 | 0.45 |
| Ever smokers | 0.35 | 0.47 |
| if quit before 35 yo | 0.25 | 0.48 |
| if quit before 35-65 yo | 0.34 | 0.46 |
| if quit after 65 yo | 0.48 | 0.49 |

Table 2.3 shows regression results where the dependent variable is an indicator for quitting smoking and the main explanatory variables are the smoking-related health shock, the nonsmoking-related health shock and its first lags. The first column shows the regression results of a fixed effect logit model. The second and third columns show the results of estimations assuming random effects in order to estimate the coefficients of time invariant variables.

In all specifications we observe a significant effect of the smoking-related health shock on the probability of quitting. Receiving a nonsmoking-related health shock also seems to increase the probability of quitting although to a lower degree. This makes sense if doctors

[3] Smith et al. [2001] condition diabetes to cases that required hospitalization. Unfortunately, the question on hospitalization is not present after 1994.

Table 2.3: Regression Results

| | (1) | (2) | (3) |
|---|---|---|---|
| *quit smk* | Logit, fe | Logit, re | OLS, re |
| smoking shock | 1.357*** | 1.158*** | 0.245*** |
| | (967.2) | (5.04) | (4.84) |
| non-smoking shock | 0.433*** | 0.603*** | 0.0845*** |
| | (298.3) | (9.07) | (7.4) |
| L.smk shock | -0.115*** | 0.220** | 0.0208* |
| | (-66.45) | (3.21) | (2.26) |
| L.nsmk shock | -0.244*** | 0.0738 | -0.0015 |
| | (-170.06) | (1.18) | (-0.18) |
| young | | -0.329*** | -0.068*** |
| | | (-9.87) | (-16.66) |
| educated | | 0.194*** | 0.0282*** |
| | | (3.92) | (3.78) |
| race black | | 0.187*** | 0.0181** |
| | | (4.18) | (2.79) |
| race other | | 0.260*** | 0.034*** |
| | | (4.59) | (3.91) |
| region Midwest | | 0.100 | 0.012 |
| | | (1.76) | (1.47) |
| region South | | 0.0788 | 0.00908 |
| | | (1.56) | (1.26) |
| region West | | 0.189** | 0.0252** |
| | | (3.03) | (2.75) |
| smk shock*edu | | 0.096 | 0.0503 |
| | | (0.55) | (1.30) |
| smk shock*black | | -0.176 | -0.0307 |
| | | (-1.14) | (-0.73) |
| smk shock*other | | -0.296 | -0.0307 |
| | | (-1.56) | (-0.78) |
| constant | | -1.929*** | 0.183*** |
| | | (-61.26) | (26.91) |
| Obs. | 14.267 | 32.297 | 32.297 |

$*p < 0.05 **p < 0.01 ***p < 0.0001$. t-statistics in parenthesis, clustered by individual. Data from HRS, *quit smk* equals 1 when an individual is a smoker in one wave and a nonsmoker in the following.

recommend their patients to quit smoking even when they get diagnosed with diabetes or other types of cancer. The interactions at the end of the table are not statistically significant. This implies that there is no evidence that people respond different to the shocks depending on their race or level of education.

From this analysis we learn that individuals do react when they experience health shocks, with an important proportion of them quitting smoking after a smoking-related health shock. Considering that we do not see different reactions by race or education and that the overall effect of the shock is less than one (not everyone quits after receiving a shock, as seen from the OLS coefficient), it must be that there are other channels that explain quitting and the differential trends in smoking by socioeconomic group. Indeed, many individuals quit smoking before receiving any health shock.

The estimation of our model will allow us to determine the relative importance of these other factors changing the beliefs and behaviors of individuals throughout their life cycle. A simpler version of the model presented in the third column of Table 2.3 will be used in the calibration to understand how do demographics and health shocks affect the quitting probability.

## 2.3 Model

In this section we present a structural dynamic model of beliefs and smoking where beliefs evolve with information from different sources and is shared within a network of friends and family over time.

An agent belongs to a group $g \in \{1, \ldots, G\}$. A group consists of a particular birth cohort, race, education level and region of residence. We follow the agent over the life-cycle and the agent decides whether to smoke or not in each period. While smoking, the agent builds up a stock of addiction, $A_{it}$, that may shape utility and that may determine health and mortality. The agent is uncertain about the effect of tobacco on health and learns over the life-cycle from different sources, introspection, medical information and from other agents. The way the information and the beliefs of the effect of tobacco diffuses across time and socio-economic groups is the object of the study. This diffusion will shape health behavior, health and longevity and their spatial dispersion. In the model, beliefs and smoking are jointly determined and as such are both endogenous.

### 2.3.1 Beliefs and Learning

The agent hesitates between two states of the world, one in which tobacco is harmful (denoted $H$) and one in which it is not (denoted $NH$). The agent receives a private signal $\sigma_{it} \in \{0,1\}$ about the harmfulness of tobacco in the form of a diagnosis of a tobacco related disease, such as lung cancer. Denote the conditional probability of getting a smoking-related health shock as:

$$P(\sigma_{it} = 1|\sigma_{it-1} = 0) = \pi(t, A_{it}, g)$$

This probability is increasing with age, $t$, and with the stock of addiction $A_{it}$, but is not necessarily zero for non-smokers, so that the signal is an imprecise one. Serious tobacco related diseases take time to develop, so that at a young age there are no big differences between smokers and non smokers. We assume that $\sigma_{it} = 1$ is an absorbing state. We also assume

$$\pi(t, A_{it}, g|H) = \pi(t, A_{it}, g) \qquad \text{and} \qquad \pi(t, A_{it}, g|NH) = \pi(t, 0, g)$$

In the non harmful state of the world, the likelihood of contracting a tobacco related disease is the one of a non-smoker. Denote by $\lambda_{it}$ the log odds ratio of the prior of tobacco being harmful for individual $i$ at time $t$:

$$\lambda_{it} = \log\left(\frac{P_{it}(H)}{P_{it}(NH)}\right)$$

#### Introspection

Agents will use Bayes rule to update their beliefs from their own health shocks. In the absence of external influences, the log odds ratio evolves over the life-cycle as:

$$\begin{cases} \lambda_{it} = \lambda_{it-1} + \log\left(\frac{\pi(t,A_{it},g)}{\pi(t,0,g)}\right) & \text{if } \sigma_{it} = 1 \text{ and } \sigma_{it-1} = 0 \\ \lambda_{it} = \lambda_{it-1} + \log\left(\frac{1-\pi(t,A_{it},g)}{1-\pi(t,0,g)}\right) & \text{if } \sigma_{it} = 0 \text{ and } \sigma_{it-1} = 0 \end{cases} \tag{2.2}$$

At a young age, the ratio $\frac{1-\pi(t,A_{it},g)}{1-\pi(t,0,g)}$ is close to one, so that the log odds ratio does not change much from introspection. Note that never smokers do not update their prior based on introspection.

### Medical Information

Public health authorities may communicate new information about the effect of tobacco and convey a belief $\lambda_t^M$, which will affect the agent's beliefs depending on an updating parameter $\delta^M(g)$. How much the agent's beliefs are affected by new medical information will depend on the agent's socioeconomic group.

### Social Learning

The agents can also update their priors by meeting other agents and exchanging information on their evaluation of the state of the world.

In learning from others, agents behave as DeGroot agents, that is, they average the beliefs of all the other agents with whom they communicate (after everybody has received their signal). Other agents, however, have a different influence on $i$'s beliefs depending on their social proximity. For example, a young high-educated individual might not be very influenced by the beliefs of an old low-educated individual.

In particular, agents will average the Bayesian learning (introspection) of their network as well as their network's learning from medical information, where the network here is understood as the people with whom the individual interacts and trusts. For each agent, we construct a network by drawing members such as to reproduce the patterns described in Table 2.1.

Combining all the ingredients mentioned so far, beliefs evolve as follows:

$$\lambda_{it} = \lambda_{it-1} + Learn_{it}^I + Learn_{it}^M + Learn_{it}^O \tag{2.3}$$

where $Learn_{it}^I$ equals the log odds ratio of equation 2.2, $Learn_{it}^M$ equals $\delta^M(g)\lambda_t^M$ and learning from others is defined as:

$$Learn_{it}^O = \delta^{OI}(g) \sum_{j \neq i} \mu(i,j) Learn_{jt}^I + \delta^{OM}(g) \sum_{j \neq i} \mu(i,j) Learn_{jt}^M \tag{2.4}$$

with $\sum_{j \neq i} \mu(i,j) = 1$ which can be interpreted as the weight that agent $i$ places on the beliefs of the other agent or as the probability that agents $i$ and $j$ interact. In practice, $\mu(i,j) > 0$ only if $j$ is part of $i$'s network.

Finally, new generations inherit their parents' beliefs at age 15, with an additive error measured by $\delta^F$.

### 2.3.2 Mortality

Mortality is modelled in a competing risks framework as the individual may die from two separate causes in each period. Denote $\omega_{it}^0$ and $\omega_{it}^1$ two iid shocks such as $\omega_{it}^0, \omega_{it}^1 \sim \mathcal{N}(0,1)$. Death occurs either if $\omega_{it}^0 > \bar{\omega}^0(t,g)$ or $\omega_{it}^1 > \bar{\omega}^1(t,g,\sigma_{it})$. The two thresholds decrease with age and differ by group. The second threshold is such that $\bar{\omega}^1(t,g,\sigma_{it}=1) < \bar{\omega}^1(t,g,\sigma_{it}=0)$. The second shock can therefore be interpreted as a tobacco related disease as it is conditional on having the tobacco related disease. The survival probability is:

$$P(Survival_{it}|\sigma_{it}) = \Phi_{it} = \Phi(\bar{\omega}^0(t,g)) + \Phi(\bar{\omega}^1(t,g,\sigma_{it})) - \Phi(\bar{\omega}^0(t,g)).\Phi(\bar{\omega}^1(t,g,\sigma_{it}))$$

This specification allows for the fact that some population groups - for whom the general cause of death is large and have a short life expectancy irrespective of smoking- may not experience much tobacco related diseases and would therefore not learn on their own or from observing members of their own community.

### 2.3.3 Dynamic Choice

In each period the agent decides whether to smoke or not, denoted $b_{it} \in \{0,1\}$. The agent faces a static budget constraint and allocates total income between a general consumption bundle, $c_{it}$, and smoking with a relative price $p_t$:

$$c_{it} + p_t b_{it} = y(t,g)$$

where $y(t,g)$ is the income, taken as deterministic for simplicity. We write the value of being alive, with a particular belief, $\lambda_{it}$, the experience of a tobacco related health shock $\sigma_{it}$, a stock of addiction $A_{it}$ and tobacco prices $p_t$ with the following Belman equation:

$$V_t(\lambda_{it},\sigma_{it},A_{it},p_t) = \max_{b_{it}} u(c_{it},b_{it},A_{it},\sigma_{it}) + \Phi_{it}\beta E_t V_{t+1}(\lambda_{it+1},\sigma_{it+1},A_{it+1},p_{t+1}) \qquad (2.5)$$

with

$$A_{it+1} = A_{it} + b_{it}$$

The agent discounts the future with a pure discount factor $\beta$ and with the probability of being alive next period. The individual takes the expectation of the future value as the advent of tobacco related health shocks are uncertain and as prices are stochastic. We assume that the relative price of tobacco follows a Markov process. Regarding the future health shock, the expectation depends on the individual's belief. From the perspective of the agent, the probability of getting a cancer related disease is:

$$P(\sigma_{it+1} = 1 | \sigma_{it} = 0, \lambda_{it}, A_{it}) = \pi(t, A_t, g) P_{it}(H) + \pi(t, 0, g)(1 - P_{it}(H))$$

From the perspective of the agent, $E_t \lambda_{i,t+1} = \lambda_{it}$. This is because an individual cannot forecast a different belief in the future without adopting it instantaneously. Put differently, one cannot think today that tobacco is not dangerous, while knowing that tomorrow it will be seen as dangerous. Hence, the agent expects that all other agents and health authorities have the same beliefs.

This model offers a broad understanding of the decision of smoking: individuals choose to smoke or not considering their tastes, prices, beliefs on the harmfulness of smoking, their addiction stock and their life expectancy.

### 2.3.4 Implementation

We consider an overlapping generations model, where agents have offspring at age 25. Agents start at age 15 with an initial stock of addiction $A_{i0} = 0$ and with $\sigma_{i0} = 0$ as they are not born with a tobacco related disease. We take 5 cohorts, plus a cohort 0 to bequeath their beliefs to the first generation and which is discarded in the estimation. The first cohort starts with 15 years of age around 1910 (with a five year dispersion), with subsequent generations starting in 1935, 1960, 1985 and 2010.

We assume that in the initial period, $t = 0$ agents of generation 0 draw the log odds from a normal distribution, with a mean and variance to be estimated:

$$\lambda_{i0} \sim \mathcal{N}(\delta_{INIT}, \sigma_\delta)$$

Each individual belongs to a socioeconomic group, defined by their race (white, black and other), their level of education (at most high school or more than high school) and their region (Northeast, Midwest, South and West). We take 17 as the number of individuals in the social network of trust following the sociological study of DiPrete et al. (2011).

The functional forms for the utility function, mortality thresholds and the probability of getting the signal (smoking-related health shock) are given by:

$$U_{it} = \alpha_0 + exp(\alpha_1 \sigma_{it}) c_{it}^{\alpha_2} (1 + b_{it})^{\alpha_3} (1 + A_{it})^{\alpha_4} (1 + A_{it} b_{it})^{\alpha_5} (1 + b_{it}\sigma_{it})^{\alpha_6} \tag{2.6}$$

$$\begin{aligned} \bar{\omega}_{it}^0 =& \omega_{00} + \omega_{01} Age_{it} + \omega_{02} Age_{it}^2 \\ \bar{\omega}_{it}^1 =& \omega_{10} + \omega_{11} Age_{it} + \omega_{12} Age_{it}^2 + \omega_{13}\sigma_{it} \end{aligned} \tag{2.7}$$

$$P(\sigma_{it} = 1 | \sigma_{it-1} = 0) = \Phi[\pi_0 + \pi_1 Age + \pi_2 Age^2 + \pi_3 Age^3 + \pi_4 \mathbb{1}_{(A_{it}>0)} + \pi_5 \mathbb{1}_{(A_{it}>0)} Age] \tag{2.8}$$

The parameter $\alpha_3$ in the utility function is calculated using a truncated normal distribution $\mathcal{TN}(\mu_\alpha, \sigma_\alpha, 0)$ that allows individuals to have a different but always positive taste for tobacco depending on their type. We consider 5 types, so that the lowest taste equals the mean of the first fifth of the density of the truncated normal. We let the model identify the parameters of this truncated function.

## 2.4 Calibration

The model presented in the previous section would require individual data on smoking, health shocks, mortality and beliefs for several generations, which is not available. For this reason we estimate the model with the method of simulated moments, which allows us to combine information from different data sources at different points in time.

The model is solved by backward induction (value function iterations) based on an initial set of parameters and then simulated for individuals over their life cycles. The simulated data provide a panel dataset used to construct moments that can be matched to moments obtained from the observed data. Using a quadratic loss function, the parameters of the model are then chosen such that the simulated moments are as close as possible to the observed moments. The moments chosen for calibration are presented in Table 2.4.

Table 2.4: Moments for calibration

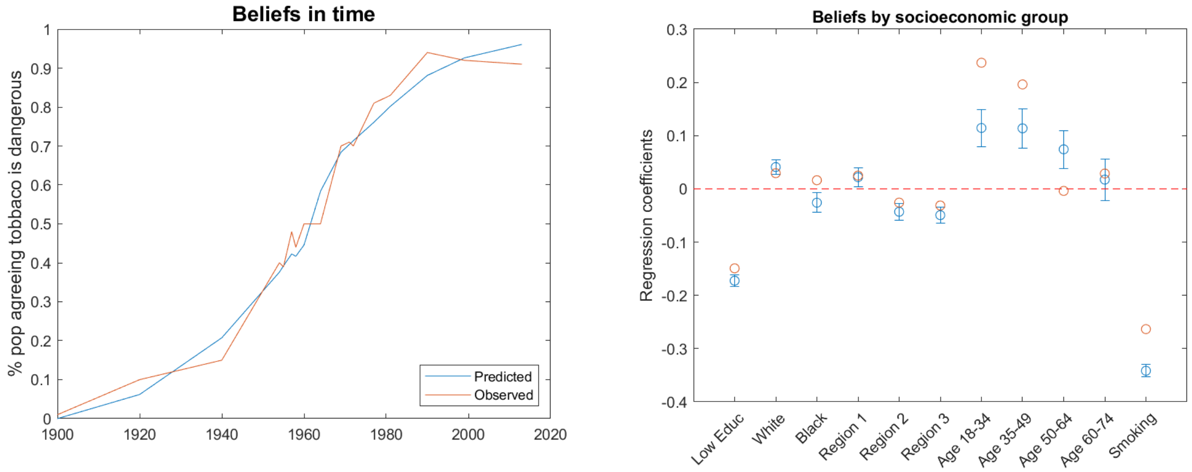
| | Moment | Dataset |
|---|---|---|
| Social networks | Proximity matrix | GSS |
| Mortality | Death rates, by socioeconomic group | CDC death rates |
| Health shock | OLS regression of smoking-related health shock on demographics and eversmoking status | NHANES |
| Beliefs | Proportion who believe smoking is harmful | Gallup |
| | Variance of relative risk, by socioeconomic group | SRBI, FFRISP |
| | Proportion who believe smoking is harmful in the life cycle, by cohort | MTF |
| | OLS regression of beliefs on demographics and smoking status | PATH, HINTS, ATRS, TABT |
| Smoking | Proportion of smokers, by socioeconomic group | NHIS |
| Quitting | RE regression of quitting on demographics and smoking-related health shock | HRS |
| Smoking attributable mortality | Proportion of smoking attributable mortality, by year and by age group | SGR |

The model calibrates the parameters for the proximity matrix, $\alpha$-parameters for the utility function, $\omega$-parameters for the mortality thresholds, $\pi$-parameters for the signal, $\delta$-parameters for beliefs, the discount factor $\beta$ and parameters for prices.

### 2.4.1 Model fitting

Figure 2.7 shows the model fitting to the data on beliefs. The model is able to capture very well the evolution of beliefs in time and fairly well the differences in beliefs by socioeconomic groups, except perhaps the differential effects by age.

Figure 2.7: Model fitting, beliefs



The left panel shows agreement with the statement *Do you think smoking is one of the causes of lung cancer?* from the Gallup poll. Data before 1950 was imputed linearly to reach 0 in 1900. The right panel shows coefficient estimates using the factor of beliefs explained in section 2.2.3 as the dependent variable.

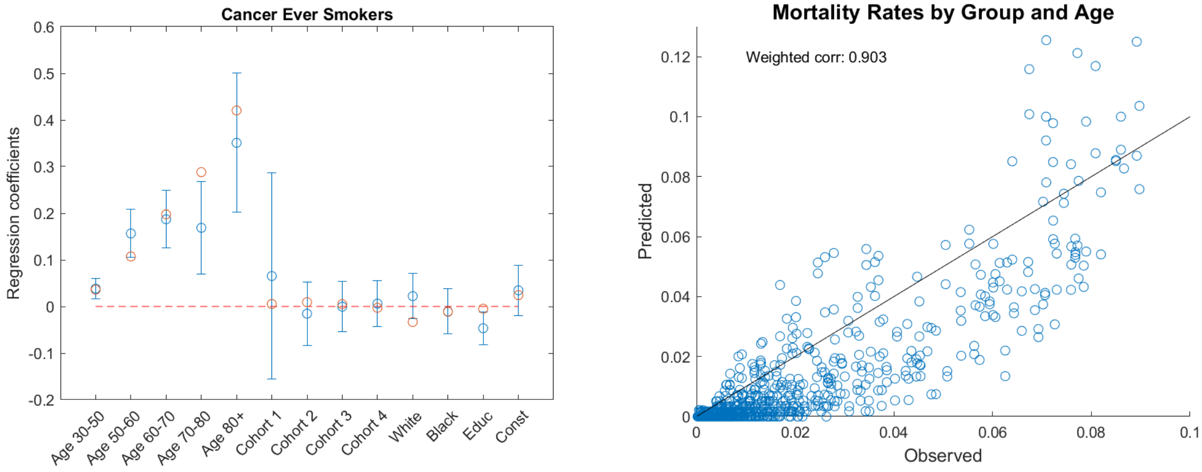
In Figure 2.8, instead we observe the model fitting the data on mortality. Here we see again that most regression coefficients related to a smoking-related health shock in ever smokers fall within the confidence bands of the model estimates. Looking at mortality in general, the correlation between the data and the model predictions is close to 90% once we take into account the observed population weights.

For some other variables, such as quitting after receiving a smoking-related health shock or the proportion of smoking attributable mortality, the model could still improve its performance, as can be seen in Figure 2.9.

## 2.5 Long term effects of Medical Information

We now use our model to quantify the long term effects of medical information in this setting. We consider the case in which both medical information proxied by the publication of scientific articles on the harms of smoking and newspaper coverage of this information (including the Surgeon General Report), are set to zero. This counterfactual tells us what would have been the effects on beliefs, smoking and smoking-related diseases if there were

Figure 2.8: Model fitting, mortality



The left panel shows coefficients from a regression of prevalence of smoking-related health shocks (from NHANES data) on the observed covariates. The right panel plots the predicted versus observed mortality rates for each socioeconomic, age and year groups.

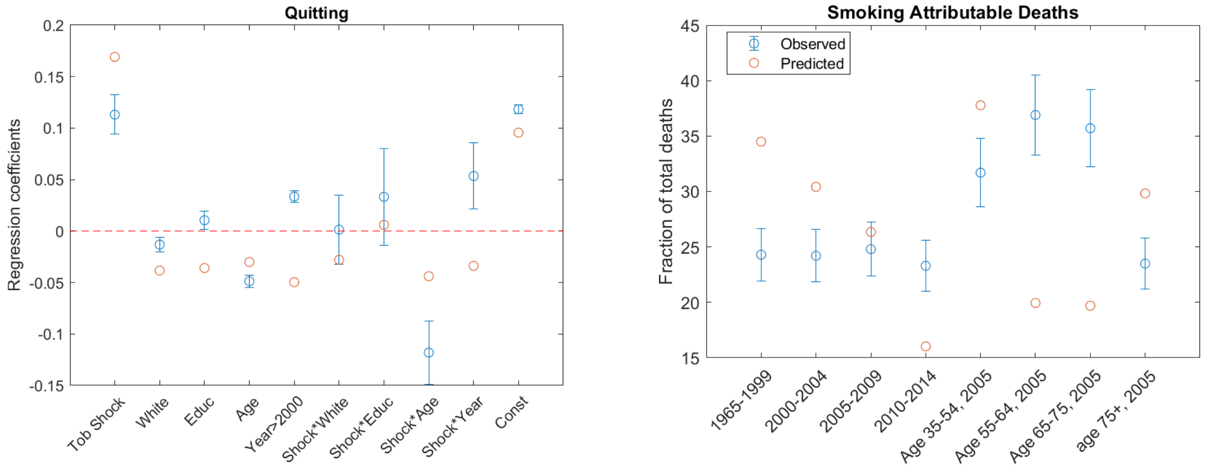
no production nor formal dissemination of information regarding the harms of tobacco. The first panel of Figure 2.10 shows that under this counterfactual the speed of learning regarding the harms of smoking would have been considerably smaller in the period 1960-1980. Then both lines are almost parallel until they converge around the year 2060.

In the right panel of Figure 2.10 we see a plot of the difference between the predicted beliefs with and without the counterfactual. In this graph we see more clearly the big drop[4] in learning about the harmfulness of smoking. This reduction in learning would have affected primarily our third cohort, born around 1945, since they could start smoking at age 15 in the year 1960.

Figure 2.11 shows the counterfactual smoking prevalence and the counterfactual prevalence of smoking-related shocks (in eversmokers) for individuals of the most affected cohort. We see that without medical information smoking prevalence in this cohort for the period 1970-1990 would have been almost 10% higher (a 26% increase with respect to baseline). This of course translates into a larger prevalence of smoking-related health shocks when these individuals reach old age. In the left panel of Figure 2.11 we see that around the year 2030, we would expect 38% of eversmokers in this cohort to get a smoking-related disease at

[4] The irregular spikes around 1980 are caused by the presence of few individuals of the first cohort who are 95 years old in this period.

Figure 2.9: Model fitting, other



The left panel shows regression coefficients with quitting smoking as dependent variable, with data from HRS. The right panel shows the fraction of total deaths that can be attributed to smoking, with data from the 2014 Surgeon General Report *The Health Consequences of Smoking*.

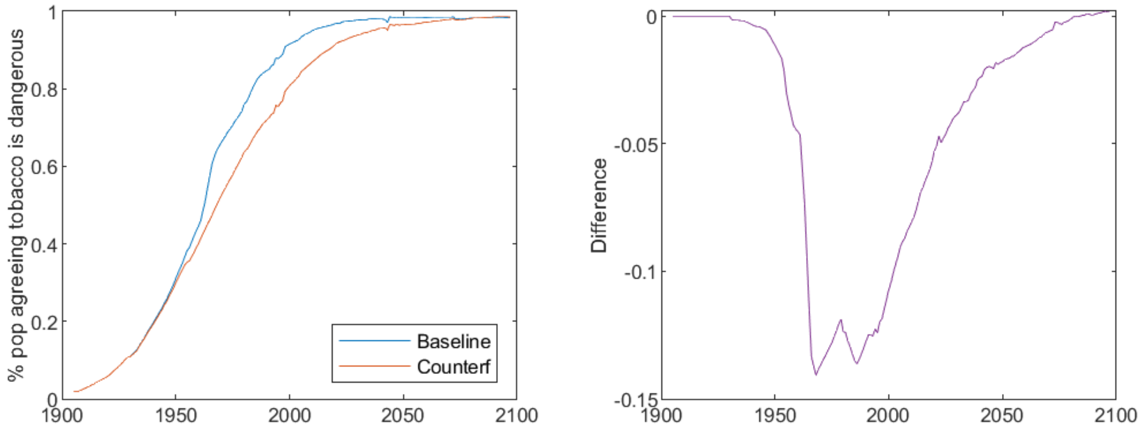
baseline, while under the counterfactual this is 42%. Since the percentage of eversmokers is also larger under the counterfactual, the yellow line shows the percentage of individuals who would get a smoking-related disease considering the eversmokers at baseline, which is almost 44% (a 21% increase with respect to baseline), implying that the burden of smoking-related diseases would be considerably bigger without the learning induced by medical information.

Notice however that even without the surgeon general report nor the subsequent public health messages and diffusion of medical information, there would be full awareness of the harmfulness of smoking around the year 2060. The increased life expectancy of the population implies that many smokers get to experience smoking-related health shocks, and this information would be transmitted to nonsmokers and future generations. According to our model then, medical information on the harmfulness of smoking had only a short term effect, reducing smoking and the onset of smoking-related diseases in the second half of the 20th century and the first half of the 21st century.

## 2.6 Conclusion

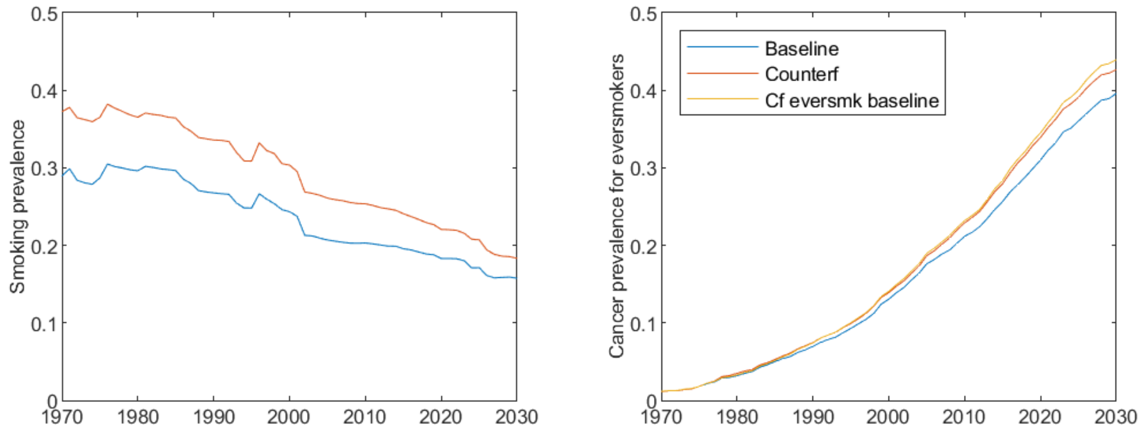
This paper studies the role of information in the evolution of beliefs about the harmfulness of smoking and the resulting reduction in smoking prevalence. We take social sharing of infor-

Figure 2.10: No medical information, beliefs



The left panel shows the predicted beliefs of Figure 2.4 and these beliefs under the counterfactual. The right panel shows the difference between these two lines.

Figure 2.11: No medical information, smoking and cancer: Cohort 3



The left panel shows the predicted smoking prevalence in the baseline scenario and in the counterfactual. The right panel shows our estimate of cancer prevalence for ever smokers under the baseline and counterfactual scenarios and under the counterfactual considering ever smokers at baseline.

mation as an important mechanism through which individuals learn from others' experiences and beliefs.

We gathered data from 20 different data sources and built a comprehensive dynamic model of smoking and beliefs that we can use to analyze the role of each learning mechanism and the long term effect of medical information. In future versions of this paper we will explore the effects of policies based on tax increases comparing them with information policies as well as identify the socioeconomic groups that would be more effective in disseminating information relevant to health behaviors.

# Chapter 3

# Can Youth Empowerment Programs Reduce Violence against Girls?

*with Selim Gulesci and Diego Ubfal*
JEL Code: J12, J13, J16, O15

## 3.1 Introduction

Throughout the world, strict containment measures caused by the COVID-19 pandemic have increased risk factors associated with domestic violence. Media and institutional reports have indicated an increase of domestic violence and in particular of violence against women and children in countries affected by COVID-19 (e.g., Reynolds, 2020; Taub, 2020; Ritz et al. 2020). Several recent studies provide evidence of this by showing increases in violence against women (VAW) due to the COVID-19 pandemic and lockdown measures (Aguero, 2021; Boserup et al., 2020; Leslie and Wilson, 2020; Mahmud and Riley, 2021; Ravindran and Shah, 2020; Silverio-Murillo et al., 2021). The reported rise in VAW due to the COVID-19 pandemic has been named by UN Women as the “Shadow Pandemic” (United Nations Women, 2020).

In this paper, we study the effects of a youth empowerment program on the prevalence of violence against youth during the COVID-19 lockdown in Bolivia. The program combines training in soft skills and technical skills with sex education, mentoring and job-finding assistance. We conducted a randomized control trial with 600 vulnerable youth who applied to the program in four cities in Bolivia. Our data include an in-person baseline survey and a follow-up survey conducted by phone due the social distancing restrictions imposed by the COVID-19 pandemic. The follow-up survey was carried out seven months after the end of the program and six months into the lockdown.

The program was designed to strengthen youth's income-generating capacity by developing their skills, and by offering job-finding assistance. Its main target outcomes were total earnings and soft skills. It was not designed to directly address violence against youth. However, given the alarming levels of violence reported at baseline, the increased relevance of violence against youth during the COVID-19 pandemic, and the possibility that both changes in soft skills and earnings could affect this outcome, we included violence as our third main outcome.

We find that the program significantly reduced violence experienced by treated girls. The prevalence of violence reported by girls fell by 10 percentage points, over a mean of 21 percent in the control group. For boys, we do not find significant reductions in violence. To address concerns about self-reporting bias, we use item list experiments, which confirm our main findings. The level of violence among girls, as measured by a list experiment included in our follow-up survey, is much lower in the treatment group than in the control group, while it is not lower for treated boys than for control boys. We present evidence that the program had a positive effect on earnings for girls, but not for boys. This is consistent with an improvement in girls' bargaining power within the household or a reduction in income-related stress, both of which may explain the decrease in violence against girls. We do not find evidence for mechanisms related to changes in soft skills as we do not see any effects of the program on a set of targeted soft skills.

The paper contributes to a burgeoning literature studying the causes of violence in general, and violence against girls in particular. Economic crises, conflicts and natural disasters are often linked to increased prevalence of violence against women and children (Anastario et al. 2013; Weitzman and Behrman, 2016; Fraser, 2020; Thurston et al. 2021). A recent study by Ravindran and Shah (2020) shows an increase in domestic violence complaints in India in districts with the strictest confinement rules. In a related study, Bandiera et al. (2020) show that temporary school closures during the 2014-16 Ebola epidemic in Sierra Leone increased teenage pregnancies and lowered school enrollment among girls, and a program that provided safe spaces (in the form of community clubs) and training in soft skills lowered these negative impacts. We contribute to this emerging literature by showing that multi-faceted youth empowerment programs can be one way to curtail the rise in violence against girls during high-risk periods, such as the COVID-19 pandemic and the ensuing lockdowns.[1]

The rest of the paper is organized as follows. Section 2 explains the details of the program. Section 3 describes the design of the experiment and sample characteristics. Section 4 presents the results. Section 5 concludes.

## 3.2 Youth Empowerment Program

The program *Adolescents: Protagonists of Development* has been implemented by the NGO *Save the Children* in Albania, Bolivia, Nepal and Uganda since 2016. Its main aim is to help vulnerable youth find a job, improve their working conditions and strengthen their income generation capacity. The target population consists of vulnerable adolescents aged 15 to 18.

To identify the sample for this study, adolescents were recruited in four cities in Bolivia. Several recruitment strategies were used: fliers in markets and other public places, Facebook ads, cooperation with neighborhood associations and schools offering night shifts, and press conferences. Interested adolescents filled a form containing information used to measure their social vulnerability (e.g., housing conditions, access to health care, violence, substance use) and their economic vulnerability (e.g., household income, child labor, lack of economic support from the family).[2] The program staff selected 600 adolescents that fulfilled at least one of the vulnerability criteria, giving priority to those with higher levels of vulnerability and who also showed willingness to participate in face-to-face interviews.

Table C.2 compares key characteristics of the sample of recruited youth with those of a sample of youth aged 15-18, living in the four cities of our study, obtained from the representative Bolivian Household Survey.[3] While we see similarities in some variables (television, computer and internet ownership and having children), and only a small difference in enrolment rates,

[1] See Kerr-Wilson et al. (2020) for a recent survey of interventions to prevent violence against women and girls. Our findings are in line with Yount et al. (2017), who conduct a systematic review of reviews on the impact of interventions to prevent violence against women and girls (VAWG) and conclude that bundled interventions with multiple components typically had more favorable impacts on VAWG than single-component interventions.

[2] Table C.1 presents the list of vulnerability criteria used by the program.

[3] Source: Instituto Nacional de Estadistica (2019).

it is clear that the study sample is more vulnerable as indicated by a higher probability of working, living in precarious dwellings, alcohol consumption, and being in a household with income below the minimum wage.

The program provided youth with soft skills and technical skills training, sex education, mentoring and support in finding a job or starting a business. In particular, it offered the following activities:

- General Training
  - Personal empowerment (Module 1, 16 hours)
  - Sexual and reproductive health (Module 2, 16 hours)
  - Economic empowerment (Module 3, 16 hours)
  - Basic competences (Module 4, 16 hours)
- Technical-skills training in predefined areas according to market demand (70 hours)
- Work insertion or business development

The four modules covering general training were taught by Save the Children. Module 1 focused on self-esteem and leadership while providing adolescents with an opportunity to get to know each other and increase trust. Module 2 discussed contraceptive methods and teen pregnancy. Module 3 covered material on market analysis, entrepreneurial soft skills, sustainable business models, workers' rights and how to prepare for a job interview. Finally, Module 4, taught basic math and literacy. These four modules were taught over the course of 4 four-hour sessions each.

After this general training, the project trained youth in specific technical skills. Adolescents were able to select from up to three training courses among a menu designed based on market demand studies by Save the Children in cooperation with private partners in each region. The training was implemented by the local partners. For example, in La Paz, the most common choices were gastronomy, customer service and graphic design. The total length of these activities was approximately 70 hours.

The final activity of the project was to help adolescents find interviews with employers offering jobs that match adolescents' skill levels and satisfy certain standards (e.g., they are compatible with schooling, they do not involve risky activities,[4] and they offer a wage no lower than the minimum wage[5]). Working conditions were monitored for up to three months from the start of the jobs. The program also offered adolescents who did not want to find a job the opportunity to start their own business. However, this happened with only one adolescent in our sample.

## 3.3 Methodology and sample characteristics

### 3.3.1 Methodology

To estimate the causal effect of the program, we conducted a randomized control trial (RCT). The RCT was designed and implemented through our collaboration with Save the Children Bolivia and the support from Save the Children Italy. The evaluation covered four metropolitan cities in Bolivia: Cochabamba, La Paz, Oruro and Santa Cruz. The first step of the evaluation involved selecting a sample of eligible youth. The program team identified 600 youth satisfying the criteria for selection into the program as explained above. All selected youth completed an in-person baseline survey.

We then conducted a private lottery using Stata to randomly select 300 youth who would be offered to be part of the program starting from 2019 (treatment group), and 300 youth who would not be offered to be part of the program in 2019, but would have the chance to participate in the program after the evaluation (control group). When conducting the randomization, we stratified the samples on region, gender, age, whether the adolescent was working at the time of the interview and, only for Cochabamba and La Paz,[6] whether s/he was a victim of violence.

The general training modules started in August 2019, and ran for about three months in the form of 4-hour weekly meetings. The technical skills training was temporarily suspended during the election-related violence in October-November 2019, which translated into four weeks of highly reduced economic activity, but it was completed by December 2019 in all sites. In February 2020, the project started offering job-finding assistance. Due to the COVID-19 pandemic, the country entered a strict lockdown on March 22, 2020 (March 16

[4] Under the Bolivian law, minors cannot work in certain tasks, such as mining or lifting heavy objects.

[5] Informal employment is common in this context and minimum wage requirements are not strictly enforced for informal jobs.

[6] In Oruro and Santa Cruz there was not enough variation to stratify on violence.

in Oruro). In two of the four cities (La Paz and Oruro), the program managed to offer job-finding assistance to most treated youth; while in the other two cities (Cochabamba and Santa Cruz), it only completed the training, with no significant job-finding assistance. In what follows, we assess the heterogeneity of treatment effects along this geographical dimension.[7]

Due to the COVID-related mobility restrictions, we conducted a follow-up survey by phone in the last two weeks of September 2020.[8] We were able to survey 511 adolescents (85% of the sample). Response rates were similar across treatment arms, and we do not find any evidence of differential attrition (see Table C.4 in the Appendix).[9]

Both the in-person baseline and the phone follow-up survey included a module to elicit sensitive questions on violence. To guarantee respondent's safety, enumerators were trained in each case by an expert on Child Safeguarding Policy following stringent ethical guidelines on how to ask these questions. Enumerators were instructed to take measures to verify the privacy of the interviews. Same-sex enumerators were used when possible. In the case of phone interviews, additional steps were taken to prevent potential perpetrators from listening to participants' answers. In particular, the interviewer provided examples of what types of actions are considered as violent; participants were asked to answer only ”yes” or ”no” and given the option of not answering the question if they did not feel comfortable with it. Respondents were also provided with a list of all the institutions where a violence victim can receive help and protection as well as the procedure to file a complaint. Finally, enumerators received an adverse event protocol explaining what they had to do in cases of abuse.

[7] Note that we are highly underpowered to detect regional heterogeneity of treatment effects by gender of the participants. As such, the analysis pertaining to regional heterogeneity should be seen as merely suggestive evidence.

[8] Due to the pandemic, the country entered a strict lockdown in March 2020. Most of the population was forced to stay at home, except for those employed in priority sectors. In the beginning of June, some of the lockdown measures were relaxed and many economic activities restarted.

[9] Due to the pandemic, we had to adapt our methodology relative to what we had pre-registered in the AEA's RCT registry. In particular, we had to switch from in-person to phone-based follow-up survey, which required using a shorter questionnaire than what we had originally planned. As a result, we were not able to collect detailed information on all the registered outcomes, such as youth's employment status. We decided to focus on outcomes that we deemed to be more relevant during the pandemic, such as violence, for which we introduced the list experiment.

### 3.3.2 Sample characteristics

Overall, 63% of the youth in the sample are girls. Table 3.1 presents summary statistics by gender on youth who completed both the baseline and follow-up surveys. Baseline characteristics of both boys and girls in the treatment and control samples are well balanced.[10] Youth in the sample are 15 to 18 years old; around 65% of them are 17 or 18 years old.

The baseline data document a high level of violence experienced by both boys and girls. More than half of these youth reported having ever experienced some type of violence. Table C.7 shows that most violence seemed to be happening at home: around a quarter of girls and boys affected by violence reported that the perpetrator was the father and/or the mother; 13% of boys and 20% of girls reported violence from siblings. This is in line with the finding in Devries et al. (2018) that the most common perpetrators of violence against both boys and girls are household members. However there is also an important role of violence perpetrated by friends (around 30%) and other relatives (16% for boys and 9% for girls). Finally, 12% of boys and 9% of girls reported suffering violence from the partner, and 7% of girls and 4% of boys from teachers in school.

Violence is classified into 3 categories; physical, psychological, and sexual. We find levels of physical and psyschological violence that are higher for boys, and levels of sexual violence that are higher for girls. The high prevalence of violence is in line with that reported in other data sources in Bolivia. The National Statistical Institute of Bolivia and UNICEF find that children are physically punished by an adult member in 83% of households.[11] Another UNICEF publication in 2014 states that 61% of girls aged 15 to 19 experienced physical violence by their current or former partner, while 22% of boys of the same age range report having experienced violence by a friend or acquaintance, and 13% by their current or former partner.[12]

Average monthly earnings are higher for boys than for girls: Bs. 441 (64 USD) and Bs. 406 (59 USD), respectively. To put this in context, the national minimum wage during this period was Bs. 2,122 (307 USD) per month. In terms of sources of income, 26% of boys and 22% of girls receive some income from wage-employment, 3% of both girls and boys receive self-employment income, around 38% get informal transfers from family or friends and close

[10] Table C.3 shows that the same holds for the full sample of youth who completed the baseline.

[11] Source: https://www.unicef.org/bolivia/biptico_estudio_violencia_ninez_bolivia.pdf

[12] Source: http://files.unicef.org/publications/files/Hidden_in_plain_sight_statistical_analysis_EN__Sept_2014.pdf

Table 3.1: Baseline Characteristics and Balance, panel sample

| | Boys | | Girls | |
|---|---|---|---|---|
| | (1)<br>Control | (2)<br>Treatment-Control | (3)<br>Control | (4)<br>Treatment-Control |
| | Mean | Diff. | Mean | Diff. |
| Age 17 or 18 | 0.638 | -0.038 | 0.652 | -0.014 |
| | (0.483) | (0.072) | (0.478) | (0.053) |
| Enrolled in school | 0.925 | -0.047 | 0.888 | -0.021 |
| | (0.265) | (0.044) | (0.316) | (0.036) |
| Violence victim (ever) | 0.574 | 0.003 | 0.528 | 0.002 |
| | (0.497) | (0.073) | (0.501) | (0.055) |
| Physical | 0.284 | 0.059 | 0.212 | 0.047 |
| | (0.454) | (0.078) | (0.410) | (0.052) |
| Psychological | 0.464 | -0.020 | 0.376 | 0.012 |
| | (0.502) | (0.078) | (0.486) | (0.057) |
| Sexual | 0.017 | -0.017 | 0.060 | 0.005 |
| | (0.130) | (0.017) | (0.239) | (0.031) |
| Income (last month) | 460.681 | -69.848 | 395.323 | 39.755 |
| | (789.323) | (111.551) | (715.819) | (76.668) |
| Employment | 0.266 | -0.044 | 0.217 | -0.037 |
| | (0.444) | (0.064) | (0.414) | (0.044) |
| Entrepreneurship | 0.032 | -0.021 | 0.031 | -0.007 |
| | (0.177) | (0.021) | (0.174) | (0.018) |
| Informal transfer | 0.383 | 0.028 | 0.379 | 0.067 |
| | (0.489) | (0.073) | (0.487) | (0.054) |
| Formal transfer | 0.021 | -0.021 | 0.019 | 0.011 |
| | (0.145) | (0.015) | (0.136) | (0.017) |
| Self concept | -0.090 | 0.001 | 0.045 | 0.017 |
| | (0.617) | (0.092) | (0.671) | (0.080) |
| Self control | -0.058 | -0.048 | 0.048 | 0.042 |
| | (0.617) | (0.093) | (0.664) | (0.073) |
| Social skills | -0.125 | -0.073 | 0.153 | -0.100 |
| | (0.717) | (0.111) | (0.667) | (0.077) |
| Communication skills | -0.075 | 0.004 | 0.032 | -0.007 |
| | (0.799) | (0.109) | (0.726) | (0.083) |
| Conflict resolution skills | 0.030 | -0.080 | 0.056 | -0.092 |
| | (0.686) | (0.102) | (0.686) | (0.078) |
| Skills to look for a job | 0.005 | -0.042 | 0.044 | -0.051 |
| | (0.734) | (0.110) | (0.767) | (0.088) |
| Observations | 94 | 184 | 161 | 327 |

**Notes:** The sample includes respondents who were successfully tracked and re-interviewed at the follow-up survey. Column (1) shows the mean and standard deviation for boys in the control group. Column (2) shows the coefficient of an OLS regression of each covariate on an indicator for treatment, robust standard errors are included in parentheses. Columns (3) and (4) show similar statistics for girls. The randomization was conducted within region and stratified on a dummy for age 17 or 18 (age 15 or 16 is the omitted category), a dummy for gender, for currently working and for reporting being victim of violence (only for Cochabamba and La Paz). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

to 2% get formal transfers by the government or an NGO. In order to measure soft skills, we used the "Employability Assessment (EA)" tool developed by Save the Children with the aim of measuring employability skills. The EA tool is a questionnaire with 24 items that produces quantitative scores for 6 categories of soft skills: self-concept, self-control, social skills, communication skills, conflict resolution skills and job-searching skills.[13]

### 3.3.3 Estimation of treatment effects

In order to estimate the average treatment effects of the program, we use the following equation:

$$y_i = \alpha + \beta \cdot T_i + \gamma \cdot T_i \cdot F_i + \lambda \cdot y_{i0} + \Theta_{i0} + \epsilon_{it}, \tag{3.1}$$

where $y_i$ is the outcome of interest for respondent $i$ at the follow-up survey, $T_i$ is a dummy variable equal to 1 if the respondent was allocated to the treatment group, $y_{i0}$ is the baseline level of the outcome for individual $i$, $F_i$ is an indicator variable that is equal to 1 if the respondent is female and 0 otherwise, and $\Theta_{i0}$ are randomization strata (dummies for each strata used in the randomization, including regional dummies). The estimate for $\beta$ corresponds to the treatment effect on males, the estimate for $\gamma$ correspond to the differential effect of the treatment on females relative to males, while the sum $\beta + \gamma$ corresponds to the treatment effect for females. We use standard errors that are robust to heteroskedasticity. For a few observations with missing values in the baseline value of the dependent variable, we replace those values with zero and add dummies for missing observations.

Our main table presents results for the three main outcomes: a dummy for any violence reported, monthly income and a soft skills index. We report p-values adjusted for the fact that we are testing for 6 main hypotheses (3 outcomes and 2 groups -boys and girls-).[14] We then look at more detailed variables for each of these outcomes: types of violence, sources of income and components of the soft-skill index.

[13] The detailed questions used in the tool are presented in Table C.8. Each component of the index is constructed by first standardizing the responses to the individual questions (by subtracting the mean and dividing by the standard deviation of the control group) and then averaging across the standardized outcomes.

[14] We obtain family-wise adjusted p-values using the implementation by Jones et al. (2018) of the free step-down procedure of Westfall and Young (1993).

## 3.4 Results

### 3.4.1 Main Outcomes

Table 3.2: Main Outcomes

| | (1)<br>Any violence (ever) | (2)<br>Income (last month) | (3)<br>Soft skills Index |
|---|---|---|---|
| Treat | 0.072 | -79.072 | 0.073 |
| | (0.049) | (100.555) | (0.065) |
| | [.0158] | [0.442] | [0.372] |
| Treat × Female | -0.168*** | 198.477 | -0.037 |
| | (0.065) | (121.704) | (0.084) |
| | [0.012] | [0.144] | [0.659] |
| Observations | 507 | 511 | 511 |
| Treatment effect for females | -0.095 ** | 119.404 * | 0.036 |
| | (0.042) | (67.434) | (0.053) |
| | [0.02] | [0.133] | [0.709] |
| Control mean, male | 0.073 | 539.794 | 0.027 |
| Control mean, female | 0.210 | 294.114 | -0.019 |

**Notes:** OLS regressions controlling for baseline value of the outcome and randomization strata. Robust standard errors in parenthesis and p-values adjusted for multiple hypotheses testing in squared brackets. We adjust for 6 hypotheses (3 outcomes and 2 groups -males and females-). At the bottom of the table we present the adjusted p-values for the effects for females, which is a linear combination of the other coefficients, and not a new hypothesis. "Control mean" refers to the mean in the control group of males/females at follow-up survey. **Any violence**: dummy variable equal to 1 if one of three types of violence during the 3 months preceding the follow-up survey (See Appendix 3.7.2 for details). **Income**: total income the month before the follow-up survey, in Bolivianos. **Soft Skills Index**: index constructed out of 24 questions reported in Table C.8 and a dummy variable =1 if the respondent overestimated the number of correct answers provided to five general knowledge questions. The significant starts are based on the p-values before adjusting for multiple hypotheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.2 presents the effects of the program on the three main outcomes: reported violence in the 3 months before the survey, monthly income during the month before the survey and soft skills. We do not see any statistically significant effect for boys. For girls, there is a marginally significant effect on monthly income and no effects on soft skills. We do see a significantly larger reduction in reported violence by girls, which survives the adjustment for multiple hypotheses testing.[15] We study this effect in more detail below.

[15] Our power calculations considering 15% attrition indicate that the minimum detectable effect to obtain 80% power at a 5% level is 0.21-0.22 standard deviations of the outcome for the full sample, 0.26-0.28 s.d. for the subsample of girls and 0.34-0.37 s.d. for the subsample of boys. Thus we are significantly underpowered to detect effects for boys.

### 3.4.2 Effects on violence: details

Table 3.3 presents more details on the effects of the program on the prevalence of violence. In the follow-up survey, enumerators asked participants about violence experienced during the previous three months.

Table 3.3: Violence

| | Self reported (last month) | | | | List exp. (last week) |
|---|---|---|---|---|---|
| | (1)<br>Any violence (ever) | (2)<br>Physical | (3)<br>Psychological | (4)<br>Sexual | (5)<br>Physical |
| Treat | 0.072 | 0.010 | 0.082* | 0.019 | 0.149 |
| | (0.049) | (0.026) | (0.049) | (0.023) | (0.268) |
| Treat × Female | -0.168*** | -0.042 | -0.185*** | -0.053* | -0.501* |
| | (0.065) | (0.035) | (0.063) | (0.028) | (0.296) |
| Observations | 507 | 511 | 508 | 510 | 507 |
| Treat. effect females | -0.095** | -0.032 | -0.103** | -0.033** | -0.353 |
| | (0.042) | (0.025) | (0.042) | (0.016) | (0.235) |
| p-value no design effects | | | | | 1.000 |
| Control mean, male | 0.073 | 0.031 | 0.063 | 0.021 | 0.320 |
| Control mean, female | 0.210 | 0.057 | 0.209 | 0.032 | 0.391 |

**Notes:** OLS regressions where the dependent variable is a dummy variables =1 if the respondent reported having experienced: in column (1) any violence during the 3 months preceding the follow-up survey; in column (2) any physical violence, in column (3) any psychological violence, in column (4) any sexual violence. The last column presents the results of the list experiment: respondents were randomly allocated to a group given 5 statements (including a sensitive item) or a group given 4 statements. The sensitive item was ”You have suffered some kind of physical violence in the last week.” We run a regression of number of correct statements on the randomly allocated group, treatment, a female dummy and all interactions among these variables. The table reports the coefficient on the interaction between treatment and group, and on the triple interaction among treatment, group and female. All regressions control for randomization strata and an indicator for having ever experienced the corresponding type of violence at baseline. The test for the presence of design effects is based on the likelihood ratio test proposed by Blair and Imai (2012). "Control mean" refers to the mean in the control group of males/females at follow-up survey. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We find a negative and significant treatment effect on violence reported by girls, but not by boys. In particular, female participants are 9.5 percentage points (ppt) less likely to report suffering any violence. The point estimate for boys is positive (7.2 ppt) but not statistically significant. The differential treatment effect on girls relative to boys is 16.8 ppt and statistically significant. In terms of magnitudes, the 9.5 ppt reduction in the prevalence of violence against girls in the treatment group is a large impact, corresponding to a 46%

reduction relative to the control group where 21% of girls reported having experienced any type of violence.

The rest of Table 3.3 shows the effect on the different types of violence: physical, psychological and sexual. For girls, the program had a negative effect on all three types of violence, lowering the prevalence of physical violence by 3 ppt (56% relative to the control group), psychological violence by 10 ppt (50% reduction) and sexual violence by 3 ppt (103% reduction). The effect on physical violence is imprecisely estimated at conventional levels, while the effects on psychological and sexual violence are statistically significant at the 95% confidence level. We also find a marginally significant increase in psychological violence for treated boys. In particular, boys in the treatment group are 8 ppt more likely to report having experienced any psychological violence relative to the control group.[16]

A potential concern with measuring sensitive topics, such as the prevalence of violence, through direct survey questions is reporting bias: respondents may not want to report violence due to shame or concerns about anonymity. To avoid this problem, a strategy commonly used in the literature is to rely on indirect elicitation techniques, such as list experiments (e.g., Rosenfeld, Imai and Shapiro, 2015). During the follow-up survey, we conducted a list experiment to elicit rates of violence among youth. In particular, respondents were asked to report the number of statements from a list of 4 or 5 items that applied to them.[17] The main idea behind this methodology involves exploiting random variation in the presence of sensitive items in the lists. Every respondent is randomly assigned to one of two groups (group A or group B). Respondents in group B are presented with a list of 4 items, which does not include any sensitive item; while respondents in group A are presented with a list of 5 items, one of which is the sensitive item.[18] The only difference between the two lists is the presence of the statement "You have suffered some kind of physical violence in the last week" in the list presented to group A but not to group B. In order to calculate

[16] The violence figures are not directly comparable to those reported at baseline. First, because the baseline survey was conducted in person, while the follow-up was conducted by phone. Second, because at baseline the question was about having ever suffered violence, while the follow-up asked about the past three months. See Appendix 3.7.2 for details.

[17] Respondents were read the following script: "Now I'm going to read some statements about many different things. Some of these statements will be true and some will not. After I read all statements, please tell me *how many* of them are true for you. I don't want to know which ones, just how many."

[18] More specifically, respondents in group A were given the following 5 statements: "1) You have been to Peru, 2) You can play the guitar, 3) You have a family member who lives in La Paz, 4) You have seen the movie Avengers: Endgame" and the sensitive item "5) You have suffered some kind of physical violence in the last week"; while respondents in group B were only given the first 4 statements.

the percentage of youth for whom the sensitive item is true (i.e., the percentage of people who have suffered some kind of physical violence in the last week), we look at the difference between the average number of statements reported as true by respondents in group A relative to B. Since the assignment of individuals to group A or B is random, there is no reason why the number of true statements in the two groups should be different, other than the presence of the sensitive item. To ensure that the randomization was balanced within gender and treatment group, we randomized female and male respondents in treatment and control groups separately.

We conduct a number of validity checks for the list experiment. First, Table C.5 in the Appendix shows that Group A and Group B are balanced in terms of their baseline characteristics within gender groups. Second, we chose items whose evaluation should not be affected by the inclusion of sensitive items. To formally test for the existence of design effects, we use the likelihood ratio test proposed by Blair and Imai (2012). Table 3.3, column 5, reports the p-value of this test, which is 1. Thus we fail to reject the null hypothesis of no design effects. Third, the items in the list experiment were designed so that few respondents in the control group would answer affirmatively or negatively to all of them. Indeed, we observe few extreme responses: less than 10% of participants answered negatively or positively to all items, and then floor or ceiling effects are unlikely to affect our results.

We run a linear regression of the number of correct statements on dummies for the randomly allocated group, treatment, female, all double interactions and the triple interaction among these variables. The last column of Table 3.3 reports the coefficient on the interaction between treatment and randomly allocated group (treatment), which captures the effect of treatment on physical violence for boys, and on the triple interaction between treatment, group and female (treatment*female), which captures the differential treatment effect for girls. Two findings are of note: first, the rate of physical violence in the control group is 32% for boys and 39% for girls, which is much higher than those reported in the direct survey questions referring to even longer reference periods (6% for females and 3% for males). One explanation of this difference is that participants tend to under-report sensitive questions when asked directly about the sensitive topic. Second, consistent with the direct survey question, the effect of the treatment in reducing violence is driven entirely by females, which is confirmed by a large interaction coefficient (-50 pp) significant at the 10% level. The effect for boys is an increase of 15 pp, whereas the effect for girls is a reduction in 35 pp, both coefficients are imprecisely measured.

Overall, results from the list experiment show a similar pattern as the direct questions in the survey: the program reduced the likelihood of experiencing violence among girls, but not among boys.

### 3.4.3 Effect on violence: possible mechanisms

There are several channels through which the program may have led to a reduction in the prevalence of violence among treated girls. In this section, we discuss and test for some of the key mechanisms that have been highlighted in the literature. We discuss other potential mechanisms which we, unfortunately, do not have information on, in the discussion section.

**Effect on earnings**

To the extent that the program increased earnings for girls in the treatment group, this may have lowered the prevalence of violence against them through two mechanisms: First, a change in women's access to economic opportunities (such as employment or other earnings) may decrease or increase the prevalence of violence, depending on the initial allocation of bargaining power within the household and whose reservation utility is binding (Tauchen et al., 1991; Eswaran and Malhotra, 2011; Bloch and Rao, 2002; Anderson and Genicot, 2015).[19] If the program increased girls' earnings, this could have improved their outside options within the household and enabled them to leave abusive relationships. Second, economic insecurity and poverty-related stress caused by the lockdown measures are likely to increase the risk of domestic violence towards women and children (Peterman and O'Donnel 2020, Conrad-Hiebner and Byram 2020). An increase in earnings due to participation in the program may have mitigated the higher stress levels linked to economic insecurity and lowered the prevalence of violence.

Table 3.2 shows that girls in the treatment group earned Bs.119 more than girls in the control group and this effect is significant at the 10% level. Compared to the mean income of females in the control group (Bs. 294), this corresponds to a 41% increase in total income. We have to be cautious because if we consider that income was one of our three main outcomes and adjust for multiple hypotheses correction, the effect loses statistical significance.

[19] The existing evidence on the effects of labor market opportunities and income of women relative to men on the prevalence of domestic violence is mixed (Aizer, 2010; Andenberg et al., 2016; Angelucci, 2008; Bhalotra et al., 2021; Bobonis et al., 2013; Chin, 2012; Heath, 2014; Heise and Kotsadam, 2015; Hidrobo and Fernald, 2013).

Table 3.4: Income

| | | Income Source | | | |
|---|---|---|---|---|---|
| | (1)<br>Income (last month) | (2)<br>Wage-emp. | (3)<br>Self-emp. | (4)<br>Informal transf | (5)<br>Formal transf |
| Treat | -79.072 | -0.005 | 0.055* | 0.019 | -0.016 |
| | (100.555) | (0.063) | (0.032) | (0.061) | (0.057) |
| Treat × Female | 198.477 | -0.074 | -0.039 | -0.005 | 0.051 |
| | (121.704) | (0.077) | (0.043) | (0.075) | (0.068) |
| Observations | 511 | 511 | 511 | 511 | 511 |
| Treat. effect females | 119.404 * | -0.079 * | 0.017 | 0.014 | 0.035 |
| | (67.434) | (0.043) | (0.029) | (0.044) | (0.037) |
| Control mean, male | 539.794 | 0.289 | 0.021 | 0.206 | 0.165 |
| Control mean, female | 294.114 | 0.234 | 0.057 | 0.203 | 0.120 |

**Notes:** OLS regressions. The dependent variable in column (1) is total income of the respondent during August 2020, the month before the follow-up survey, in Bolivianos. The dependent variables in columns (2)-(5) are dummy variables =1 if the respondent had any earnings from wage-employment, self-employment, informal transfers (from family or friends) or formal transfers (from the government or NGOs), respectively. All regressions control for randomization strata and the baseline value of the dependent variable. "Control mean" refers to the mean in the control group of males/females at follow-up survey. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.4 tests if the program affected the different sources of income. Given the short length of the phone survey, we only have information on the extensive margin of these sources and not the value of earnings from each source. We find that treated girls are 8 ppt less likely to report earnings from wage labor. They are 2 ppt more likely to report having income from self-employment and 3.5 ppt more likely to have earnings from formal transfers (government or NGO provided), but these effects are imprecisely estimated. We cannot identify which source is driving the increase in earnings of females reported in column (1).

Tables C.6 presents results on the regional heterogeneity of the treatment effects on the three main outcomes. We find that the treatment effects for females on both violence and earnings are concentrated on the Cochabamba/Santa Cruz subsamples, as opposed to La Paz/Oruro. This supports the idea that in places where the program succeeded in increasing girls' earnings, it also led to a reduction in the prevalence of violence targeting them. Moreover, this provides additional evidence that the increase in earnings is not linked to the job-finding assistance provided by the program, since Cochabamba and Santa Cruz were precisely the two regions where the program did not manage to offer such kind of assistance as noted in section 3.1 above.

Overall, the results in Table 3.4 provide some evidence that the program helped girls increase their income, even if it was not through helping them find a better job as originally planned. This increase in girls' earnings may explain why they experienced lower violence – either because of their improved bargaining power within the household; or because the higher earnings lowered the prevalence of stress-related violence against girls.

### Effects on soft skills

To the extent that the program succeeded in changing girls' soft skills, such as self-confidence and expressiveness, this may have empowered them to leave or better face abusive relationships within the household. Table 3.5 presents more details on the treatment effects on soft skills.

Overall, we do not find significant effects on soft skills – the aggregate index combining the seven standardized indices included in the table shows no significant treatment effect (column 1), neither for boys nor for girls. When we examine the effects on individual components of the index, we only see significant treatment effects for girls in one of the seven soft-skills indices – that on job-searching skills. Column (8) of Table 3.5 reports the effects on a task we added in the follow-up survey to measure self-confidence.[20] The program has an effect on self-confidence only for boys. While 55% of boys in the control group are classified as over-confident, this share increases by 18 percentage points in the treatment group. For girls, we see a similar share of 52% classified as over-confident in both treatment and control groups (we do not see any effects for girls on being classified as under-confident either). Overall, the evidence suggests that the program did not have a significant effect on soft skills. This could be because soft skills are in general hard to measure, and the methods we used in the phone survey are not ideal to capture changes in soft skills.

[20] This task is a simpllified non-incentivized version of that developed in Blavatskyy (2009). Respondents were asked five general knowledge questions: "1) How many strings does an electric guitar have, 2) What country is Bad Bunny from?, 3) What kingdom do mushrooms belong to?, 4) What was the profession of Vincent van Gogh?, 5) What is the capital of Colombia?' Then, they had to guess how many of these questions they answered correctly. Overestimating the number of correct answers is taken as an indirect measure of overconfidence.

Table 3.5: Soft skills

| | (1)<br>Soft skills<br>Aggregate Index | (2)<br>Self<br>Concept | (3)<br>Self<br>Control | (4)<br>Social<br>Skills | (5)<br>Communication<br>Skills | (6)<br>Conflict<br>Resolution | (7)<br>Job Search<br>Skills | (8)<br><br>Confidence |
|---|---|---|---|---|---|---|---|---|
| Treat | 0.073 | 0.033 | 0.008 | 0.044 | 0.015 | -0.038 | 0.047 | 0.178*** |
| | (0.065) | (0.105) | (0.086) | (0.109) | (0.093) | (0.098) | (0.096) | (0.069) |
| Treat × Female | -0.037 | 0.011 | -0.026 | -0.110 | 0.005 | 0.049 | 0.221* | -0.184** |
| | (0.084) | (0.133) | (0.108) | (0.136) | (0.124) | (0.129) | (0.124) | (0.090) |
| Observations | 511 | 511 | 511 | 511 | 511 | 511 | 511 | 511 |
| Treat+Treat × Fem | 0.036 | 0.044 | -0.018 | -0.066 | 0.020 | 0.010 | 0.268 *** | -0.006 |
| | (0.053) | (0.081) | (0.065) | (0.080) | (0.081) | (0.083) | (0.079) | (0.057) |
| Control mean, male | 0.027 | 0.029 | -0.010 | -0.042 | 0.015 | 0.031 | 0.141 | 0.546 |
| Control mean, female | -0.019 | -0.018 | 0.006 | 0.026 | -0.009 | -0.019 | -0.086 | 0.519 |

**Notes:** OLS regressions with robust standard errors. All regressions control for randomization strata and the baseline value of the dependent variable. We replace missing values in covariates with zeros and include dummies for missing observations. The dependent variable in column (1) is constructed by first standardizing all outcome variables in columns (2)-(8) with respect to the control group (subtracting the mean and dividing by the standard deviation of the control group), then taking their average and standardizing again with respect to the control group. The dependent variables in columns (2)-(7) are constructed based on the EA tool developed by Save the Children, by first standardizing the responses (by subtracting the mean and dividing by the standard deviation of the control group for each outcome) and then averaging across the standardized outcomes. See Table C.8 in the Appendix for the individual components of the indices in columns (2)-(7). The dependent variable in column (8) is a dummy variable =1 if the respondent overestimated the number of correct answers provided to five general knowledge questions. "Control mean" refers to the mean in the control group of males/females at follow-up survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As noted in Section 3.4.2, we find a small increase in the prevalence of psychological violence among boys in the treatment group. This may be linked to the increase in boys' overconfidence caused by the program. One hypothesis is that as they became more confident, they may have tried to assert their opinions more strongly in their households or social networks, leading to an increase in arguments and verbal clashes. An alternative explanation could be that they did not increase earnings after participating in the program and this may have increased verbal abuse and conflict by their families or friends.

## 3.5 Discussion

Our findings indicate that a multi-faceted program designed to strengthen youth's income generating capacity can have significant downstream effects on reducing violence against girls. This finding is in line with the evidence presented in the review by Yount et al. (2017), which concludes that bundled interventions with multiple components show more favorable effects on violence against women and girls than single-component interventions. The intervention we evaluate combines various components aimed towards improving adolescents' skills, agency and social networks. Future work should study how to build the package of interventions that is most effective to reduce the prevalence of violence.

Our analysis provides some evidence that the program increased girl's earnings, which could have improved their outside options and their economic empowerment. But there are other alternative mechanisms through which the program may have diminished the prevalence of violence against girls. One is by lowering exposure to abusers. Confinement measures may put women and children living in abusive relations at even greater risk of violence because of increased exposure to their abusers. Exposure theories suggest that when perpetrators spend more time outside the home, victims are less exposed to potential abuse (Chin, 2012; Mobarak and Ramos, 2020). If the program decreased girls' chances of being exposed to their abusers during the lockdown (for instance, if they were more likely to have jobs in the "priority sectors" that were allowed to remain open or if the youth were able to find safer housing away from the perpetrators of violence), this may have reduced violence within the treatment group. However, we do not have any evidence that points in that direction. Another alternative mechanism could be improvements in girls' knowledge of and access to support services. To the extent that the program succeeded in providing support to girls in the treatment group that experienced violence prior to the lockdown, this may have increased their knowledge of and access to support services. Their abusers may have also become more aware of girls' improved access to support in case of violence. Unfortunately, we do not have any information to test this mechanism. The information we have indicates that the program did put in contact with support services those adolescents who reported abuse at baseline, but it did so equally in the treatment and control groups. Another mechanism might be improved social capital. The youth may have built better support networks during the program and this could have offered them an avenue to cope with or combat violence during the pandemic. These are all potential mechanisms that are worth exploring in future research evaluating youth empowerment programs.

## 3.6 Conclusions

In this paper, we present results on the effects of a multi-faceted youth empowerment program in Bolivia. We find that the program significantly reduced violence reported by girls during the lockdown period. While we do not have strong evidence on the mechanisms that generated this effect, we see some weak evidence for an increase on girls' earnings that are coming from activities unrelated to wage work. This increase in earnings could have improved girls' outside options and their economic empowerment and played a role in the reduction in violence against them.

The main contribution of our study is to show that multi-faceted interventions aiming to empower vulnerable youth can be particularly effective in reducing violence against girls (but not boys) during periods of heightened risk, such as the COVID-19 pandemic. Many studies have provided evidence of a sharp increase in violence against girls and women during the COVID-19 pandemic, but to our knowledge, this is the first paper reporting causal evidence of an intervention that significantly reduces violence against girls during the COVID-19 pandemic.

# 3.7 Appendix

## 3.7.1 Appendix Tables

Table C.1: Vulnerability criteria Used for Sample Selection

| Vulnerability type | Variable | Vulnerable if |
|---|---|---|
| Social | Civil state | Not single |
| | Children | Has children |
| | Residence | Does not live with family |
| | Household Size | More than 5 |
| | Housing: type | Precarious housing |
| | Housing: number of rooms | Overcrowded |
| | Housing: kitchen | No room to cook |
| | Housing: bathroom | No private bathroom |
| | Food: provider | Not provided by family |
| | Food: number of meals | Less than three per day |
| | Healthcare | Does not visit health center if health problem |
| | Health | Frequent health problems |
| | Drugs or alcohol | Used drugs or drank alcohol in the last 6 months |
| | Sexual Health Information | Not heard of contraceptive methods |
| | Conflict with the law | Has ever been arrested |
| | Violence victim | Has ever suffered violence |
| Economic | Household income | Father does not contribute to hh income |
| | Household income | Below minimum wage (Bs. 2,000) |
| | Family Support | Family does not support youth expenditures |
| | Housing costs | House is not owned or rented |
| | Child labor | Has ever worked |
| | Child labor | Worked before age 15 |
| | Child labor | Worked for family need |
| | Child labor | Currently works |
| | Child labor | Wage below minimum wage |

**Notes:** The table presents the vulnerability criteria used for sample inclusion. The condition to be included into the sample was to satisfy at least one of all these criteria. Priority was given to youth who satisfied a larger number of criteria.

Table C.2: Comparison with Representative Sample

| | (1)<br>Baseline | (2)<br>Household Survey |
|---|---|---|
| Age | 16.9 | 16.5 |
| Studies (enrolled) | 0.89 | 0.92 |
| Works | 0.30 | 0.18 |
| Child (only women) | 0.06 | 0.05 |
| Cigarettes (last twelve months) | 0.03 | 0.01 |
| Alcohol (last twelve months) | 0.16 | 0.04 |
| Precarious dwelling | 0.28 | 0.15 |
| No room only for cooking | 0.18 | 0.10 |
| Family income below minimum wage | 0.45 | 0.12 |
| Television | 0.93 | 0.96 |
| Computer | 0.40 | 0.40 |
| Internet | 0.31 | 0.32 |
| Observations | 600 | 1731 |

**Notes:** The table compares key characteristics of our study sample (column 1) with those of a sample of youth aged 15-18, living in the same four cities as study participants, which was taken from the 2019 representative national household survey of Bolvia. Source: own computations based on Instituto Nacional de Estadistica (2019).

Table C.3: Baseline Characteristics and Balance, Baseline sample

| | Boys | | Girls | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Control | Treatment-Control | Control | Treatment-Control |
| | Mean | Diff. | Mean | Diff. |
| Age 17 or 18 | 0.667 | -0.015 | 0.645 | -0.002 |
| | (0.473) | (0.063) | (0.480) | (0.050) |
| Enrolled in school | 0.929 | -0.054 | 0.882 | -0.015 |
| | (0.258) | (0.040) | (0.324) | (0.034) |
| Violence victim (ever) | 0.535 | 0.010 | 0.543 | -0.011 |
| | (0.501) | (0.067) | (0.499) | (0.052) |
| Physical | 0.261 | 0.047 | 0.227 | 0.031 |
| | (0.442) | (0.067) | (0.420) | (0.049) |
| Psychological | 0.422 | 0.000 | 0.402 | -0.015 |
| | (0.496) | (0.069) | (0.492) | (0.053) |
| Sexual | 0.013 | -0.013 | 0.066 | -0.002 |
| | (0.114) | (0.013) | (0.250) | (0.030) |
| Income (last month) | 515.561 | -116.704 | 419.833 | 4.236 |
| | (842.830) | (103.247) | (723.215) | (71.118) |
| Employment | 0.254 | -0.004 | 0.237 | -0.066 |
| | (0.437) | (0.058) | (0.426) | (0.042) |
| Entrepreneurship | 0.026 | -0.008 | 0.032 | -0.000 |
| | (0.161) | (0.020) | (0.177) | (0.018) |
| Informal transfer | 0.395 | 0.025 | 0.382 | 0.076 |
| | (0.491) | (0.066) | (0.487) | (0.051) |
| Formal transfer | 0.018 | -0.018 | 0.016 | 0.010 |
| | (0.132) | (0.012) | (0.126) | (0.015) |
| Self concept | -0.087 | 0.005 | 0.041 | 0.018 |
| | (0.599) | (0.082) | (0.664) | (0.075) |
| Self control | -0.093 | -0.009 | 0.037 | 0.038 |
| | (0.593) | (0.082) | (0.658) | (0.068) |
| Social skills | -0.142 | -0.036 | 0.145 | -0.094 |
| | (0.699) | (0.096) | (0.678) | (0.072) |
| Communication skills | -0.051 | -0.003 | 0.041 | -0.015 |
| | (0.765) | (0.095) | (0.731) | (0.077) |
| Conflict resolution skills | -0.007 | -0.025 | 0.055 | -0.084 |
| | (0.687) | (0.090) | (0.675) | (0.071) |
| Skills to look for a job | -0.005 | 0.003 | 0.032 | -0.057 |
| | (0.699) | (0.097) | (0.766) | (0.082) |
| Observations | 114 | 226 | 186 | 374 |

**Notes:** The sample includes all respondents who were part of the experiment at baseline. Column (1) shows the mean and standard deviation for boys in the control group. Column (2) shows the coefficient of an OLS regression of each covariate on an indicator for treatment, robust standard errors are included in parentheses. Columns (3) and (4) show similar statistics for girls. The randomization was conducted within region and stratified on a dummy for age 17 or 18 (age 15 or 16 is the omitted category), a dummy for gender, for currently working and for reporting being victim of violence (only for Cochabamba and La Paz). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.4: Attrition

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Treat | -0.003 | (0.029) | 0.021 | (0.138) |
| Female | | | -0.016 | (0.045) |
| Age 17 or 18 | | | 0.031 | (0.046) |
| Enrolled in school | | | -0.011 | (0.077) |
| Any violence (ever) | | | -0.044 | (0.083) |
| Physical | | | 0.046 | (0.067) |
| Psychological | | | 0.074 | (0.066) |
| Sexual | | | -0.020 | (0.088) |
| Income (last month) | | | 0.000* | (0.000) |
| Wage-emp. | | | -0.001 | (0.058) |
| Self-emp. | | | -0.089 | (0.101) |
| Informal transfer | | | 0.032 | (0.047) |
| Formal transfer | | | -0.173*** | (0.047) |
| Self concept | | | 0.026 | (0.045) |
| Self control | | | -0.060 | (0.045) |
| Social skills | | | -0.043 | (0.049) |
| Communication skills | | | 0.089** | (0.036) |
| Conflict resolution skills | | | -0.036 | (0.044) |
| Skills to look for a job | | | -0.020 | (0.031) |
| Female × Treat | | | -0.049 | (0.064) |
| Age 17 or 18 × Treat | | | 0.036 | (0.064) |
| Enrolled in school × Treat | | | 0.034 | (0.105) |
| Any violence (ever) × Treat | | | 0.014 | (0.115) |
| Physical × Treat | | | -0.040 | (0.088) |
| Psychological × Treat | | | -0.091 | (0.099) |
| Sexual × Treat | | | 0.002 | (0.116) |
| Income (last month) × Treat | | | -0.000* | (0.000) |
| Wage-emp. × Treat | | | 0.045 | (0.085) |
| Self-emp. × Treat | | | 0.332 | (0.207) |
| Informal transfer × Treat | | | 0.016 | (0.069) |
| Formal transfer × Treat | | | 0.097 | (0.062) |
| Self concept × Treat | | | -0.041 | (0.063) |
| Self control × Treat | | | 0.016 | (0.059) |
| Social skills × Treat | | | 0.040 | (0.057) |
| Communication skills × Treat | | | -0.088 | (0.058) |
| Conflict resolution skills × Treat | | | 0.076 | (0.065) |
| Skills to look for a job × Treat | | | 0.012 | (0.046) |
| Observations | 600 | | 600 | |
| Control mean | 0.148 | | 0.148 | |
| p-value joint orthogonality test | | | 0.668 | |

**Notes:** The dependent variable is a dummy =1 if the respondent could not be surveyed at the follow-up survey. Robust standard errors are shown in Coumns (2) and (4). We replace missing values in covariates with zeros and include dummies for missing observations. The p-value for the joint orthogonality test corresponds to an F-test for joint significance of the interaction terms. Robust standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.5: Balance of List Experiment

| | Boys | | Girls | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Group B | Group A-Group B | Group B | Group A-Group B |
| | Mean | Diff. | Mean | Diff. |
| Age 17 or 18 | 0.570 | 0.108 | 0.630 | 0.025 |
| | (0.498) | (0.072) | (0.484) | (0.053) |
| Enrolled in school | 0.871 | 0.051 | 0.870 | 0.025 |
| | (0.337) | (0.045) | (0.337) | (0.036) |
| Violence victim (ever) | 0.570 | 0.008 | 0.519 | 0.025 |
| | (0.498) | (0.074) | (0.501) | (0.056) |
| Physical | 0.226 | 0.030 | 0.179 | 0.025 |
| | (0.420) | (0.064) | (0.385) | (0.044) |
| Psychological | 0.409 | 0.003 | 0.333 | 0.025 |
| | (0.494) | (0.073) | (0.473) | (0.053) |
| Sexual | 0.011 | -0.011 | 0.049 | -0.006 |
| | (0.104) | (0.011) | (0.217) | (0.023) |
| Income (last month) | 439.398 | -27.009 | 427.704 | -19.914 |
| | (825.640) | (112.129) | (665.139) | (77.265) |
| Employment | 0.226 | 0.030 | 0.198 | 0.000 |
| | (0.420) | (0.064) | (0.399) | (0.044) |
| Entrepreneurship | 0.022 | -0.010 | 0.025 | 0.006 |
| | (0.146) | (0.019) | (0.156) | (0.018) |
| Informal transfer | 0.376 | 0.046 | 0.407 | 0.006 |
| | (0.487) | (0.073) | (0.493) | (0.055) |
| Formal transfer | 0.022 | -0.022 | 0.025 | -0.000 |
| | (0.146) | (0.015) | (0.156) | (0.017) |
| Self concept | -0.065 | -0.049 | 0.058 | -0.011 |
| | (0.581) | (0.093) | (0.693) | (0.080) |
| Self control | -0.045 | -0.074 | 0.051 | 0.035 |
| | (0.576) | (0.094) | (0.636) | (0.073) |
| Social skills | -0.188 | 0.053 | 0.069 | 0.059 |
| | (0.750) | (0.111) | (0.658) | (0.077) |
| Communication skills | -0.069 | -0.008 | -0.005 | 0.069 |
| | (0.765) | (0.109) | (0.712) | (0.083) |
| Conflict resolution skills | -0.054 | 0.083 | 0.011 | -0.013 |
| | (0.705) | (0.102) | (0.623) | (0.078) |
| Skills to look for a job | -0.025 | 0.020 | -0.030 | 0.086 |
| | (0.757) | (0.111) | (0.804) | (0.088) |
| Observations | 93 | 184 | 162 | 327 |

**Notes:** Columns (1) and (3) show the mean and standard deviation for the control group of the list experiment for boys and girls, respectively. Columns (2) and (4) show the coefficient of an OLS regression of each covariate on an indicator for treatment, with robust standard errors in parentheses. The randomization was stratified on gender. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.6: Main outcomes, regional heterogeneity

| | (1)<br>Any violence (ever) | (2)<br>Income (last month) | (3)<br>Soft skills Index |
|---|---|---|---|
| Treat | 0.109* | -36.993 | 0.090 |
| | (0.060) | (128.723) | (0.083) |
| Treat × Female | -0.244*** | 199.932 | -0.073 |
| | (0.079) | (154.853) | (0.106) |
| Treat × La Paz or Oruro | -0.108 | -148.189 | -0.072 |
| | (0.101) | (191.613) | (0.123) |
| Treat × Female × La Paz or Oruro | 0.280** | -28.397 | 0.143 |
| | (0.132) | (226.404) | (0.162) |
| Observations | 507 | 511 | 511 |
| Treat. effect in Cbba or Sta Cruz females | -0.135 *** | 162.939 * | 0.017 |
| | (0.051) | (85.058) | (0.066) |
| Treat. effect in La Paz or Oruro males | 0.001 | -185.182 | 0.018 |
| | (0.082) | (142.165) | (0.091) |
| Treat. effect in La Paz or Oruro females | 0.038 | -13.648 | 0.089 |
| | (0.068) | (85.135) | (0.083) |
| Control mean, male, Cbba or Santa Cruz | 0.060 | 582.500 | -0.029 |
| Control mean, female, Cbba or Santa Cruz | 0.254 | 290.588 | -0.028 |
| Control mean, male, La Paz or Oruro | 0.103 | 439.655 | 0.156 |
| Control mean, female, La Paz or Oruro | 0.077 | 304.872 | 0.010 |

**Notes:** All regressions control for randomization strata and the baseline value of the dependent variable. We replace missing values in covariates with zeros and include dummies for missing observations. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.7: Violence perpetrator, at baseline

| | Any violence | |
|---|---|---|
| | (1)<br>Boys | (2)<br>Girls |
| Father | 0.18 | 0.19 |
| Mother | 0.20 | 0.22 |
| Siblings | 0.10 | 0.16 |
| Relatives | 0.13 | 0.08 |
| Friends | 0.24 | 0.24 |
| Partner | 0.09 | 0.06 |
| Teacher | 0.02 | 0.05 |
| Observations | 122 | 201 |

**Notes:** The table shows the share of adolescents who mention a specific perpetrator conditional on having reported any type of violence. Respondents could name multiple perpetrators.

Table C.8: EA tool

| | | Strongly disagree | Disagree | Neither agree nor disagree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| Positive Self-Concept | I feel valued and appreciated by others | 1 | 2 | 3 | 4 | 5 |
| | I feel good about my future | 1 | 2 | 3 | 4 | 5 |
| | I anticipate my own needs ahead of time | 1 | 2 | 3 | 4 | 5 |
| | I can adapt to changes by learning new skills | 1 | 2 | 3 | 4 | 5 |
| Self-Control | I am able to complete assignments in time | 1 | 2 | 3 | 4 | 5 |
| | I go to work even when I feel like staying at home | 1 | 2 | 3 | 4 | 5 |
| | I feel proud when I produce high quality work | 1 | 2 | 3 | 4 | 5 |
| | I follow workplace or school dress codes | 1 | 2 | 3 | 4 | 5 |
| Social Skills | I can understand and work with people of different backgrounds | 1 | 2 | 3 | 4 | 5 |
| | I can give my opinions/suggestions to others without offending them | 1 | 2 | 3 | 4 | 5 |
| | I value the input and contributions of others | 1 | 2 | 3 | 4 | 5 |
| | I take responsibility for what I do | 1 | 2 | 3 | 4 | 5 |
| Communic. Skills | I know how to express myself in proper ways | 1 | 2 | 3 | 4 | 5 |
| | I know how to articulate my own ideas clearly | 1 | 2 | 3 | 4 | 5 |
| | I read so I can comprehend and use new information | 1 | 2 | 3 | 4 | 5 |
| | I listen actively to understand and learn | 1 | 2 | 3 | 4 | 5 |
| Problem Skills | I collect and analyze information to find the best solutions to a problem | 1 | 2 | 3 | 4 | 5 |
| | I seek many sources of information to solve a problem in school or at work | 1 | 2 | 3 | 4 | 5 |
| | I learn from my past successes and mistakes to make future decisions | 1 | 2 | 3 | 4 | 5 |
| | I can adapt to changing circumstances | 1 | 2 | 3 | 4 | 5 |
| Job Search Skills | I know how to complete a job application | 1 | 2 | 3 | 4 | 5 |
| | I have the skills and experience valued by employers | 1 | 2 | 3 | 4 | 5 |
| | I have the knowledge and skills needed to interview for jobs | 1 | 2 | 3 | 4 | 5 |
| | I know how to prepare a resume | 1 | 2 | 3 | 4 | 5 |

### 3.7.2 Violence indicators

To measure exposure to violence at baseline, we relied on a survey module developed and piloted by Save the Children with vulnerable adolescents in Bolivia. The module is in line with the Safeguarding Children Policy of Save the Children. As part of the in-person baseline survey, adolescents were asked: “Have you ever suffered from any type of violence?” If they answered positively, they were then asked a follow-up open question “What type of violence have you suffered from?,” which was coded by trained enumerators into three categories: Physical Violence (blows, slaps, pushes, hair pulling, pinches, kicks, bites, burns), Psychological/Verbal Violence (Insults, threats, verbal abuse, derision, intimidation, contempt, mockery, etc.), or Sexual violence (forced sex, transactional sex, sexual harassment, touching your body without permission). This module differs from survey instruments where the enumerator describes specific acts of violence and asks the respondent whether s/he experienced each act. Based on our discussions with colleagues from Save the Children, this method was deemed less triggering and less stressful for participants.

At the follow-up survey, we adapted the module to accommodate the change in the survey method to a phone survey. In particular, we chose questions that did not require respondents to describe the type of violence they may have experienced. This was important considering that even if the protocol included a verification on whether the respondent was alone and in a private environment, it was not possible to definitely rule out that other people, including perpetrators, could overhear the conversation. During the phone surveys, enumerators read the different types of violence and respondents would simply answer “Yes,” “No,” or “I would like to skip this question.” Moreover, we changed the time window and asked about violence experienced in the 3 months before the survey, which corresponded to the period after the onset of the COVID-19 crisis in Bolivia. The exact script read by enumerators was: “I will describe three types of violence that youth like yourself have experienced in Bolivia. I only ask you to answer yes or no when I read each statement, you can also choose not to answer. Have you suffered the following types of violence in the last 3 months?” Then the enumerator read the following three statements: “Physical violence includes cases when someone hits or slaps you, or pushes you or pulls your hair, Psychological or verbal violence includes cases when someone insults, threatens, verbally abuses, ridicules, or makes fun of you, Sexual violence includes cases when someone touches your body without your permission, or demands forceful sex or harasses you.”

As described in Section 3.4.2, the list experiment included in the follow-up survey identified physical violence experienced by respondents during the week before the survey based on the following statement: “You have suffered some kind of physical violence in the last week.” The levels of violence estimated with this method were significantly higher than those obtained with the direct question. While this could be explained by the fact that list experiments are designed to reduce reporting bias, there were two additional differences between the direct question on physical violence and the corresponding item in the list experiment: 1) The recall period: the direct question asked about physical violence during the past 3 months, while the item list asked about the last week; 2) In the direct question, enumerators gave detailed examples of physical violence, as explained above. The first difference is unlikely to lead to higher levels of physical violence when measured with the list experiment since the reference period was longer in the direct question. The second difference may have induced respondents to think about specific acts of physical violence, but the list experiment was conducted after the direct questions, and therefore once respondents had already heard the examples regarding physical violence acts. Hence we believe that any effect this may have had on the reported level of physical violence is likely to be small.

# Bibliography

Adhvaryu, A. (2014). Learning, misallocation, and technology adoption: Evidence from new malaria therapy in tanzania. *Review of Economic Studies*, 81:1331–1365.

Albouy, D., Graf, W., Kellogg, R., and Wolff, H. (2016). Climate Amenities, Climate Change and American Quality of life. *Journal of the Association of Environmental and Resource Economists*, 3(1):205–246.

Albouy, D. and Stuart, B. A. (2020). Urban population and amenities: the neoclassical model of location. *International economic review*, 61(1):127–158.

Alexander, D. and Currie, J. (2017). Is it who you are or where you live? Residential segregation and racial gaps in childhood asthma. *Journal of Health Economics*, 55:186–200.

Allcott, H., Diamond, R., Dubé, J., Handbury, J., Rahkovsky, I., and Schnell, M. (2019). Food Deserts and the Causes of Nutritional Inequality. *Quarterly Journal of Economics*, 134(4):1793–1844.

Almagro, M. and Domınguez-Iino, T. (2021). Location sorting and endogenous amenities: Evidence from Amsterdam. NYU, mimeograph.

Anderson, S. T. and West, S. E. (2006). Open space, residential property values, and spatial context. *Regional science and urban economics*, 36(6):773–789.

Barili, E., Bertoli, P., and Grembi, V. (2021). Neighborhoods, networks, and delivery methods. *Journal of Health Economics*, 80:102513.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of political economy*, 115(4):588–638.

Bayer, P., Keohane, N., and Timmins, C. (2009). Migration and hedonic valuation: The case of air quality. *Journal of Environmental Economics and Management*, 58(1):1–14.

Bayer, P., McMillan, R., and Rueben, K. (2005). An Equilibrium Model of Sorting in an Urban Housing Market. Working Paper 10865, NBER.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, 63:841–890.

Bilger, M. and Carrieri, V. (2013). Health in the cities: when the neighborhood matters more than income. *Journal of Health Economics*, 32(1):1–11.

Bolitzer, B. and Netusil, N. R. (2000). The impact of open spaces on property values in Portland, Oregon. *Journal of environmental management*, 59(3):185–193.

Boone-Heinonen, J., Gordon-Larsen, P., Guilkey, D. K., Jacobs Jr, D. R., and Popkin, B. M. (2011). Environment and physical activity dynamics: the role of residential self-selection. *Psychology of sport and exercise*, 12(1):54–60.

Bruch, E. E. and Mare, R. D. (2012). Methodological issues in the analysis of residential preferences, residential mobility, and neighborhood change. *Sociological methodology*, 42(1):103–154.

Brueckner, J. K., Thisse, J.-F., and Zenou, Y. (1999). Why is central Paris rich and downtown Detroit poor?: An amenity-based theory. *European Economic Review*, 43(1):91–107.

CDC (2021). Cancer statistics data visualizations tool. *Centers for Disease Control and Prevention.*

Chandrasekhar, A. G., Larreguy, H., and Xandri, J. P. (2020). Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, 88(1):1–32.

Chay, K. Y. and Greenstone, M. (2005). Does air quality matter? evidence from the housing market. *Journal of political Economy*, 113(2):376–424.

Chetty, R. and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162.

Chyn, E. and Katz, L. F. (2021). Neighborhoods matter: Assessing the evidence for place effects. Technical report, National Bureau of Economic Research.

Clark, A. and Etilé, F. (2001). The effect of health information on cigarette consumption: Evidence from british panel data. mimeo, University of Paris.

Cutler, D. M. and Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics*, 29(1):1–28.

Damm, A. P. and Dustmann, C. (2014). Does growing up in a high crime neighborhood affect youth criminal behavior? *American Economic Review*, 104(6):1806–32.

Dang, R. (2015). Spillover Effects of Local Human Capital Stock on Adult Obesity-Evidence from German Neighborhoods. *Available at SSRN 2683509*.

Darden, M. (2017). Smoking, expectations, and health: a dynamic stochastic model of lifetime smoking behavior. *Journal of Political Economy*, 125(5):1465–1522.

Darden, M. and Gilleskie, D. (2016). The effects of parental health shocks on adult offspring smoking behavior and self-assessed health. *Health economics*, 25(8):939–954.

Darden, M. E. (2021). Cities and Smoking. *Journal of Urban Economics*, 122:103319.

De Walque, D. (2010). Education, information, and smoking decisions evidence from smoking histories in the united states, 1940–2000. *Journal of human resources*, 45(3):682–717.

Deryugina, T. and Molitor, D. (2020). Does when you die depend on where you live? Evidence from Hurricane Katrina. *American Economic Review*, 110(11):3602–3633.

Deryugina, T. and Molitor, D. (2021). The causal effects of place on health and longevity. *Journal of Economic Perspectives*, 35(4):147–70.

Diamond, R. (2016). The determinants and welfare implications of US workers' diverging location choices by skill: 1980-2000. *American Economic Review*, 106(3):479–524.

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., and Zheng, T. (2011). Segregation in social networks based on acquaintanceship and trust. *American journal of sociology*, 116(4):1234–83.

DOHMH (2009-2013). Community Health Survey. *New York City Department of Health and Mental Hygiene.*

Du, M. and Zhang, X. (2020). Urban greening: A new paradox of economic or social sustainability? *Land Use Policy*, 92:104487.

Eid, J., Overman, H. G., Puga, D., and Turner, M. A. (2008). Fat city: Questioning the relationship between urban sprawl and obesity. *Journal of Urban Economics*, 63(2):385–404.

Ellen, I. G., Mijanovich, T., and Dillman, K.-N. (2001). Neighborhood effects on health: exploring the links and assessing the evidence. *Journal of urban affairs*, 23(3-4):391–408.

Fairchild, A. L., Bayer, R., and Colgrove, J. (2015). Risky business: New York City's experience with fear-based public health campaigns. *Health affairs*, 34(5):844–851.

Gourevitch, M. (2019). Large life expectancy gaps in US cities linked to racial and ethnic segregation by neighbrohood. Published on June 5, 2019 https://nyulangone.org/news/large-life-expectancy-gaps-us-cities-linked-racial-ethnic-segregation-neighborhood.

Hamilton, J. L. (1972). The demand for cigarettes: advertising, the health scare, and the cigarette advertising ban. *The Review of Economics and Statistics*, pages 401–411.

Heblich, S., Trew, A., and Zylberberg, Y. (2021). East-Side Story: Historical Pollution and Persistent Neighborhood Sorting. *Journal of Political Economy*, 129(5):1508–1552.

Hoffmann, R. (2017). Following the peers: The role of social networks for health care utilization in the philippines.

Jo, H.-K., Kim, J.-Y., and Park, H.-M. (2019). Carbon reduction and planning strategies for urban parks in Seoul. *Urban Forestry & Urban Greening*, 41:48–54.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119.

Krosnick, J. A., Malhotra, N., Mo, C. H., Bruera, E. F., Chang, L., Pasek, J., and Thomas, R. K. (2017). Perceptions of health risks of cigarette smoking: A new measure reveals widespread misunderstanding. *PloS one*, 12(8):e0182063.

Lim, S., Chan, P. Y., Walters, S., Culp, G., Huynh, M., and Gould, L. H. (2017). Impact of residential displacement on healthcare access and mental health among original residents of gentrifying neighborhoods in new york city. *PloS one*, 12(12):e0190139.

Liu, J. T. and Hsieh, C. R. (1995). Risk perception and smoking behavior: Empirical evidence from taiwan. *Journal of Risk and Uncertainty*, 11(2):139–157.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R., and Sanbonmatsu, L. (2013). Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity. *American Economic Review*, 103(3):226–31.

Mathes, S. (2021). The Dynamics of Residential Sorting and Health: Implications of Climate Change in the U.S. *Job Market paper*.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Morancho, A. B. (2003). A hedonic valuation of urban green areas. *Landscape and urban planning*, 66(1):35–41.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy*, 9(4):513–548.

O'Mahony, D. (2019). 10-year life expectancy gap between London streets revealed by postcode inequality study. Published on March 25, 2019 https://www.standard.co.uk/news/london/10year-life-expectancy-gap-between-london-streets-revealed-by-postcode-inequality-study-a4099701.html.

Ou, S. (2019). Are some neighborhoods bad for your waistline? A test of neighborhood exposure effects on BMI. *Journal of Health Economics*, 63:52–63.

Percoco, M. (2021). Spatial health inequality and regional disparities historical evidence from Malaria in Italy. *REGION*, 8(1):53–73.

Plantinga, A. J. and Bernell, S. (2007). Can urban planning reduce obesity? the role of self-selection in explaining the link between weight and urban sprawl. *Review of Agricultural Economics*, 29(3):557–563.

Poudyal, N. C., Hodges, D. G., and Merrett, C. D. (2009). A hedonic analysis of the demand for and benefits of urban recreation parks. *Land Use Policy*, 26(4):975–983.

Sloan, F. A., Smith, V. K., and Taylor Jr, D. H. (2002). Information, addiction, and 'bad choices': lessons from a century of cigarettes. *Economics Letters*, 77(2):147–155.

Smith, V. K., Donald H. Taylor, J., Sloan, F. A., Johnson, F. R., and Desvousges, W. H. (2001). Do smokers respond to health shocks? *Review of Economics and Statistics*, 83(4):675–687.

Sorensen, A. T. (2006). Social learning and health plan choice. *The Rand Journal of Economics*, 37(4):929–945.

Stead, M., Angus, K., Langley, T., Katikireddi, S. V., Hinds, K., Hilton, S., Lewis, S., Thomas, J., Campbell, M., Young, B., et al. (2019). Mass media to communicate public health messages in six health topic areas: a systematic review and other reviews of the evidence. *Public Health Research*.

Strulik, H. (2018). Smoking kills: An economic theory of addiction, health deficit accumulation, and longevity. *Journal of health economics*, 62:1–12.

Viscusi, K. W., Magat, W. A., and Huber, J. (1999). Smoking status and public responses to ambiguous scientific risk evidence. *Southern Economic Journal*, 66(2):250–270.

Viscusi, W. K. (1990). Do smokers underestimate risks? *Journal of Political Economy*, 98(6):1253–1269.

Walsh, R. (2007). Endogenous open space amenities in a locational equilibrium. *Journal of urban Economics*, 61(2):319–344.

Warner, K. E. (1981). Cigarette smoking in the 1970's: the impact of the antismoking campaign on consumption. *Science*, 211(4483):729–731.

Warner, K. E. (1989). Effects of the antismoking campaign: an update. *American Journal of Public Health*, 79(2):144–151.

Wen, H., Zhang, Y., and Zhang, L. (2015). Assessing amenity effects of urban landscapes on housing price in Hangzhou, China. *Urban Forestry & Urban Greening*, 14(4):1017–1026.

Wong, M. (2018). A tractable approach to compare the hedonic and discrete choice frameworks. *Journal of Housing Economics*, 41:135–141.

Wood, L., Hooper, P., Foster, S., and Bull, F. (2017). Public green spaces and positive mental health–investigating the relationship between access, quantity and types of parks and mental wellbeing. *Health & place*, 48:63–71.

Zhao, Z. and Kaestner, R. (2010). Effects of urban sprawl on obesity. *Journal of Health Economics*, 29(6):779–787.

## for Chapter 3

Aguero, Jorge M. (2021). COVID-19 and The Rise of Intimate Partner Violence. World Development 137: 105217.

Aizer, A. (2010). The Gender Wage Gap and Domestic Violence. American Economic Review 100: 1847-1859.

Anderson, S. and G. Genicot. (2015). Suicide and Property Rights in India. Journal of Development Economics 114: 64-78.

Andenberg, D., Rainer, H., Wadsworth, J. and T. Wilson. (2016). Unemployment and Domestic Violence: Theory and Evidence. Economic Journal 126: 1947-1979.

Anastario, M., Shehab, N., and L. Lawry. (2013). Increased Gender-based Violence Among Women Internally Displaced in Mississippi 2 Years Post-Hurricane Katrina. Disaster Medicine and Public Health Preparedness 3(1): 18-26.

Angelucci, M. (2008). Love on the rocks: Domestic violence and alcohol abuse in rural Mexico. The BE Journal of Economic Analysis and Policy, 8(1): 1-43.

Bandiera O., Buehren, N., Goldstein, M., Rasul, I. and A. Smurra. (2020). Do school closures during an Epidemic have Persistent Effects? Evidence from Sierra Leone in the Time of Ebola. Working Paper, UCL.

Bhalotra, S., Kambhampati, U., Rawlings, S., and Z. Siddique. (2021). Intimate Partner Violence: The Influence of Job Opportunities for Men and Women. The World Bank Economic Review 35(2): 461-479.

Blair, G., and K. Imai. (2012). Statistical Analysis of List Experiments. Political Analysis, 20: 47-77.

Blavatskyy, P.R. (2009). Betting on own knowledge: Experimental test of overconfidence. Journal of Risk and Uncertainty 38: 39–49.

Bloch, F., and V. Rao. (2002). Terror as a Bargaining Instrument: A Case Study of Dowry Violence in Rural India. American Economic Review, 92 (4): 1029-1043.

Bobonis, G. J., Gonzalez-Brenes, M. and R. Castro. (2013). Public Transfers and Domestic Violence: The Roles of Private Information and Spousal Control. American Economic Journal: Economic Policy 5(1): 179-205.

Boserup, B., McKenney, M. and A. Elkbuli. (2020). Alarming Trends in US Domestic Violence During the COVID-19 Pandemic. American Journal of Emergency Medicine 38(12): 2753-2755.

Chin, Y. M. (2012). Male backlash, bargaining, or exposure reduction?: Women's working status and physical spousal violence in India. Journal of population Economics 25(1): 175-200.

Conrad-Hiebner, A. and E. Byram. (2020). The Temporal Impact of Economic Insecurity on Child Maltreatment: A Systematic Review. Trauma, Violence and Abuse 21(1): 157-178.

Devries, K., Knight, L., Petzold. M., et al. (2018) Who perpetrates violence against children? A systematic analysis of age-specific and sex-specific data. BMJ Paediatrics Open 2(1).

Eswaran, M. and N. Malhotra. (2011). Domestic violence and women's autonomy in developing countries: theory and evidence. The Canadian Journal of Economics 44(4): 1222-1263.

Fraser, E. (2020). Impact of COVID-19 Pandemic on Violence against Women and Girls. Helpdesk Research Report No. 284. London, UK: VAWG Helpdesk.

Heath, R. (2014). Women's Access to Labor Market Opportunities, Control of Household Resources, and Domestic Violence: Evidence from Bangladesh. World Development 57: 32-46.

Heise, L. L. and A. Kotsadam. (2015). Cross-national and Multilevel Correlates of Partner Violence: An Analysis of Data from Population-based Surveys. The Lancet Global Health 3(6): 332-340.

Hidrobo, M. and L. Fernald. (2013). Cash Transfers and Domestic Violence. Journal of Health Economics 32(1): 304-319.

Instituto Nacional de Estadistica. (2019). Encuesta de Hogares. Catalago de Microdatos, Instituto Nacional de Estadistica de Bolivia. http://anda.ine.gob.bo/index.php/catalog/84.

Instituto Nacional de Estadistica and United Nations Children's Fund. (2005). Violencia contra le niñez en Bolivia, La Paz – Bolivia.

Jones, D., Molitor, D. and J. Reif. (2018). What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study. Working Paper 24229, NBER.

Kerr-Wilson, A., Gibbs, A., McAslan Fraser E., Ramsoomar, L., Parke, A., Khuwaja, HMA., and R. Jewkes. (2020). A rigorous global evidence review of interventions to prevent violence against women and girls. What Works to prevent violence among women and girls global Programme. Pretoria, South Africa.

Leslie, E. and R. Wilson. (2020). Sheltering in Place and Domestic Violence: Evidence from Calls for Service During COVID-19. Journal of Public Economics 189: 104241.

Mahmud, M. and E. Riley. (2021). Household Response to an Extreme Shock: Evidence on the Immediate Impact of the Covid-19 Lockdown on Economic Outcomes and Well-being in Rural Uganda. World Development 140: 105318.

Mobarak, M. and A. Ramos. (2020). The Effects of Migration on Intimate Partner Violence: Evidence for Exposure Reduction Theory in Bangladesh. Working paper, Unpublished.

Peterman, A. and M. O'Donnel. (2020). COVID-19 and Violence Against Women and Children: a Second Research Round Up. CGD Note September 2020, Center for Global Development.

Ravindran, S., and M. Shah. (2020). Unintended consequences of lockdowns: COVID-19 and the shadow pandemic. Working Paper 27562, NBER.

Reynolds, P. (2020). 25% rise in domestic violence calls during pandemic. https://www.rte.ie/news /2020/0609/1146245-domestic-violence-gardai/

Ritz, D., O'Hare, G, and M. Burges. (2020). The Hidden Impact of COVID-19 on Child Protection and Wellbeing. London, Save the Children International.

Silverio-Murillo, A., Balmori de la Miyar, J.R. and L. Hoehn-Velasco. (2021). Families under Confinement: COVID-19, Domestic Violence. Andrew Young School of Policy Studies Research Paper Series.

Taub, A. (2020). A New Covid-19 Crisis: Domestic Abuse Rises Worldwide. New York Times, 6 April. https://www.nytimes.com/2020/04/06/world/coronavirus-domestic-violence.html

Tauchen, H., Witte, A. and S. Long. (1991). Domestic violence: A nonrandom affair. International Economic Review 32(2): 491-511.

Thurston, A. M. , Stöckl, H. and Ranganathan, M. (2021) Natural hazards, disasters and violence against women and girls: a global mixed-methods systematic review. BMJ Global Health 6(4).

United Nations Children's Fund. (2014). Hidden in Plain Sight: A statistical analysis of violence against children, UNICEF, New York. https://www.unicef.org/media/66916/file/Hidden-in-plain-sight.pdf

United Nations Women. (2020). COVID-19 and Ending Violence Against Women and Girls. UN Women.

Weitzman, A., and J.A. Behrman. (2016). Disaster, Disruption to Family Life and Intimate Partner Violence: The Case of the 2010 Earthquake in Haiti. Sociological Science 3: 167-189.

Westfall, P. H. and S. S. Young. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons, vol. 279.

Yount, K. M., Krause, K. H. and Miedema, S. S. (2017) Preventing gender-based violence victimization in adolescent girls in lower-income countries: Systematic review of reviews. Social Science & Medicine 192: 1-13.