Three Essays on Environmental and Health Economics

by

Mohammad Sameer Ali Al-Azzam

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the Doctorate in Philosophy degree in Economics

Department of Economics
Faculty of Social Sciences
University of Ottawa

Dedication

This thesis is dedicated to the three most precious "S" in my life. My dad, Sameer, my mom, Siham, and my wife Susan, my endless love.

Abstract

This thesis consists of three essays in applied microeconomics: First, heat and school absence: evidence from a survey of Indian children. Second, heat-waves and short-term morbidity in India. Third, intra-family marriage and pregnancy outcomes.

Chapter 1. India is the second most populous country on earth with young people representing a significant number of the population resulting in data that indicates such figures at 38 per cent. With such high numbers, major consideration must be given into developing informed and targeted policies to ensure positive educational outcomes for young people. This paper contributes to existing literature by investigating the impact of high temperatures on students' rates of absenteeism, relying on the short-term exogenous variation in daily maximum temperatures. The paper highlights the heterogeneity of the effect of temperature by climate zone. To create data, we link information on children's school absences from the India Human Development Survey (IHDS-II) in combination with meteorological data from the ERA-Interim archive taken from a thirty day period prior to individual interview dates. Our findings suggest that high temperatures have a substantial negative impact on students' attendance in rural areas. However, limited evidence of such an effect is found in urban areas. Our results therefore indicate a need to implement future in-depth studies.

Chapter 2. Within this chapter, we investigate the impact of prolonged heat exposure on individuals short-term morbidity rates over a thirty day pe-

riod prior to the interview date. We work with a broad dataset and use an econometric model that utilizes plausibly exogenous variation in high weather temperatures. We implement the percentile-based approach and three different heatwave metrics as a innovative way of defining and capturing the impact of heatwaves on health outcomes. Our results show that heatwave intensity of the eighty-fifth percentile over the duration of three consecutive days of extensive heat has a significant adverse effect on individuals short-term morbidity. More so, our findings indicate a disparity between genders in relation to the impact of heatwaves. Finally, it can be suggested that individuals having to travel an extensive distance in order to access water are most affected by high temperatures.

Chapter 3. Millions of people worldwide are married to their blood relatives, yet the resulting impact on offspring health continues to be debated. Within this paper, we provide evidence around this debate by studying the birth outcomes from a large, representative sample of Indian women in varying marital circumstances. We explore the impact of intra-family marriage on negative pregnancy outcomes including stillbirth, miscarriage, and child death after birth. We utilize an ordinary least squares model (OLS), which controls for a wide range of financial and family factors. The results show that being a woman related to her husband by blood increases the probability of experiencing negative pregnancy outcomes by 2.8 percentage points. Our finding is robust using the instrumental variable approach (IV). The instrumental variable represents the ambient level of violence against women, which positively affects the probability of consanguineous marriage. The IV approach leads to a slightly smaller adverse impact of 2.2 percentage points. In addition, the OLS results provide suggestive evidence that intra-family marriage has no heterogeneous impact across religion types.

Acknowledgement

I am deeply grateful to my supervisor Dr.Anthony Heyes for his unwavering support and guidence through this journey. His extensive knowledge within this field of research assisted me to identify my research questions and utilize rigorous methodologies.

I thank my thesis committee including the two internal examiners, Dr.Ba Chu and Dr.Jason Garred for their invaluable mentoring, ideas and feedback during my first and second thesis workshops and thesis defence. I thank Dr.Jonathan Créchet and Dr.Casey Warman for being on the thesis committee as the third internal examiner and the external examiner, respectively. As well, on their useful comments. In addition, I would like to thank the University of Ottawa for granting me the admission scholarship and for giving me the opportunity to achieve my lifelong dream of obtaining a Ph.D.

I would like to express my gratitude to my friends Jérôme Archambent for all those times he stood by me and Angéline Mayano for her revision.

Finally, I express my warmest gratitude to my parents and my wife as well as my brothers and sisters for their unconditional support and sacrifices.

General Introduction

Globalization has shifted the ways in which we understand trade and commodities globally. Such changes have heavily impacted upon the movement of labor, trade investments, and other channels throughout the world. Therefore since we live in a global marketplace, the impact of cultural and economic shocks within any given country can dramatically affect another nation, independent of its geographical location. When analyzing global data, India stands out significantly within this area as one of the largest and most diverse countries in the world. It must also be noted that India is the third largest global player in the global economy in terms of labor movement and trade, in particular agricultural products and other linkages (Shah and Tapari, 2019).[1] Despite this, India remains a relatively impoverished nation with low level of human capital, compromised health systems and reduced education outcomes. It might be argued that India's climate plays a significant contributing role towards lower economic outcomes and levels of poverty (high temperature, low level of rain, and drought). Contemporary data highlights that within the field of social sciences India in particular produces substantial amounts of open and available data which provides a unique opportunity for researchers to investigate localised topics. In this instance, such data provides great insight into the happenings within India in particular but more importantly has the potential to impact upon a range of emerging nations. As such, for the

[1] This paper focuses on the changing nature and impact of India in the world economy and can be found using this link `https://www.researchgate.net/publication/339782597_Rising_Importance_of_India_in_the_World_Economy_-A_GVAR_Analysis`

purpose of this study, we investigate and explore three localised key research areas which in turn constitute the three main chapters of this thesis.

In the first chapter, we explore the effect of temperature variation on school attendance of young people. In the second chapter, we look at the impact of heatwaves and the link between health-related outcomes. In the third chapter, we study the impact of consanguineous marriage on pregnancy outcomes. Throughout this thesis we utilize data from the nationally representative survey, India Human Development Survey (IHDS-II). This report provides us with sound data around health behaviours as well as socioeconomic and demographic factors that enable us to inform the three research questions of this thesis.

Within the first chapter, the driving force behind this research aims to analyze the potential impact of the climate on the educational outcomes of students. It must be noted that as a country, India showcases great variations in climate resulting substantial differences in temperature. For instance, in summer temperature can get as high as 45^0C while in winter it could be as low as -6^0C. Since a large proportion of the population lives in rural areas, children often travel long distances on foot. In such cases, high temperatures have been shown to impact upon the willingness or ability of these students to attend school. It has been well documented that absence negatively impacts test scores in different subjects, which in turn negatively influences economic outcomes and health well-being. It is therefore of significant importance to study the impact of climate on school attendance. Since the variation of temperature in the short-term is theoretically random, it allows us to estimate the effect of temperature on attendance.

In India, extended periods of heat or high temperatures are common and often result in heatwaves. In the second chapter of this thesis, we estimate the impact of prolonged heat exposure on the probability of negative health related illnesses. In order to determine what is defined as a heatwave, we implement a novel approach based on percentile and three different heat metrics. Our results show that a heatwave intensity of eighty-fifth percentile in combination with three consecutive hot days have a significant adverse effect on individuals' short-term probability of negative health related illnesses. We find a distinct gender-related divisions of health outcomes, in particular within the female population which is more likely to develop health related illnesses than their male counterparts.

In the third chapter, we study the effect of consanguineous marriage on negative pregnancy outcomes in India. According to the India Human Development Survey (IHDS-II), 8% of all marriages in India are consanguineous marriages. While the rate of this marriage varies across communities, negative pregnancy outcomes could be determined by several determinants related to consanguineous marriage. Therefore, to obtain a reliable impact of consanguineous marriage on negative pregnancy outcomes we utilize an ordinary least squares model (OLS) working with a large, representative, good quality sample of eligible Indian young women. The dataset allows us to control for a wide range of financial and family characteristics. As a supporting exercise to the OLS model, we would implement an instrumental variable strategy. We instrument for the likelihood that a young female marries a family member by the ambient level of violence towards women in the area in which she resides. The OLS findings add to existing literature on consanguineous marriage by providing a result that could be interpreted as an explicit impact.

As such, the overall aim of our research is to provide greater insight into what are the driving forces that impact upon human capital in emerging economies. In realizing this, governments must recognize this key understanding in order to provide informed and targeted healthcare and education policies. The main focus of this work is therefore to provide rigorous evidence that substantiates existing research that supports the development of future longstanding frameworks and initiatives.

Table of Contents

List of Figures

List of Tables

Chapter 1

Heat and School Absence: Evidence from a Survey of Indian Children

1.1 Introduction

It has been widely documented that high temperatures affect productivity, the environment and economies adversely. As such, a one degree F increase in average temperature is associated with a 4.5% decrease in GDP per capita for a given country (Dell et al., 2009).

The focus of this paper is to determine the influence of heat on human capital by studying its impact on school attendance. In order to identify such impacts, we use a large-scale survey of school children from a representative sample of Indian households and investigate the causal link between periods of high temperatures and student absenteeism. To our knowledge, this paper is the first to address the effects of temperature variations on students' attendance.

Education is essential to the process of human capital accumulation (Burgess, 2016). In understanding this, current research supports the idea that human capital investment to improve an individual's education and cognitive functions must occur during childhood (Bleakley, 2010). Therefore, it may be argued that current barriers in accessing school effectively compromises global educational outcomes and attainments.

When analyzing global data, India stands out significantly as a heavily populated country that experiences extreme heat. Through such recognition, it must be noted that India already has a large proportion of its population living in poverty, earning less than $1.25 a day. The 2011 Census report of India shows that 69% of the population lives in rural areas.[1] Barro (2000) highlights that human capital accumulation is the primary mean to escape poverty. As such, student attendance plays a critical role in increasing lifelong outcomes and attainments. The aim of this paper is to analyze the correlation between student attendance and the impact of high temperatures within rural communities for the period of March to November. We implement the analysis on rural and urban areas separately. The paper highlights the heterogeneity of the effect of temperature by climate zone. As our findings show no statistically significant impact on students' attendance in urban areas, and to avoid keep reporting insignificant results throughout the paper, we conclude with the fact that temperature has insignificant impact in urban areas. Meanwhile, we focus the analysis throughout the paper to mainly to cover the impact in rural areas. The results show that the students in the Hot and Dry climate zone are affected the most by heat.

This paper is structured as follows. Section 2 reviews relevant existing

[1] Census of India 2011, Ministry of Home Affairs: http://censusindia.gov.in/2011-prov-results/paper2/data_files/india/Rural_Urban_2011.pdf.

literature. Section 3 describes the data used. Section 4 explains the employed methods. Section 5 presents the results, and Section 6 includes the robustness exercises and the placebo test. Section 7 concludes with key findings.

1.2 Literature Review

1.2.1 Temperature and productivity of labor

Links between climate, productivity of labor has been a topic of interest for researchers and governments worldwide. As such, both climate change and weather fluctuations impact the population. However, climate change has long-term effects whereas weather fluctuations drive short-term effects. The overall impact of temperature variations can be observed through different policies and legislation within the areas of health, economic output and overall economic growth in diverse ways.

Heyes and Saberian (2019) argue that human decisions can be affected by a change in temperature. When analyzing 207,000 US immigration court cases, they found that a one standard deviation increase in case-day temperature reduces the likelihood of a decision in favour of the applicant by 8.56%. Ambient temperature can lead to a reduction in labor productivity and consequently induces economic losses (Cai, Lu, and Wang, 2018). The authors document an inverted-U-shaped relationship between temperature and daily indoor productivity for the industrial laborers. Zhang et al. (2018) studied the impact of temperature on total factor productivity. Relying on data based upon half a million Chinese manufacturing plants from 1998 to 2007, they found that temperature could have led to 39.5 billion dollars losses without considering additional adaptive actions by the mid-twenty-first century. On the other hand, Radakovic et al. (2007) found that heat stress reduces the percentage of correct responses by a set of soldiers; their study highlighted that extreme heat generates a mild attention deficit.

1.2.2 Temperature and schooling

Investments within education is a central action to support the development of lower and middle-income countries. However, understanding the factors that impact upon educational outcomes is critical to any implementation of effective government policies. Kingdon (2007) argues that despite India represents 22% of the world's population, it accounts for 46% of the world's illiterates and for a very large proportion of the world's out-of-school children and youth. In recognising this, such data is significant and underlines the reasons as to why India has been the focus of much contemporary research.

Since there is an understanding that extreme temperatures affect human capital, economists have also been studying the impact of temperature variations on schooling and educational outcomes. For instance, Graff Zivin et al. (2017) used meteorological data linked with a national longitudinal structure survey. They found that cognitive performance, specifically in math, declines when the temperature goes beyond 78.8^0F. Meanwhile, Colmer (2013) worked with an Ethiopian rural household survey to study the effect of climate change on child labor and schooling. He used panel data that includes children aged between six and sixteen years. The results show no evidence of climate variability impacting educational attainments or decisions associated with enrolment. Colmer also observed that parents often share the workload among their children to alleviate the impact on their formal education.

Garg et al. (2017) probed the effect of high temperatures on cognitive performance in math and reading test scores using India's agricultural income mechanism. They used information taken from a poll of 4 million students from grades one to twelve. Their findings indicate that high temperature negatively impacts test scores, especially within impoverished populations.

Existing research shows that cumulative heat exposure can negatively impact upon cognitive development. However, technological adaptive responses such as air conditioning play a significant role in mitigating such impact. Goodman et al. (2018) used 10 million American students who took the PSAT exam multiple times during the period 2001 to 2014. Their findings indicate that cumulative heat exposure reduces the rate of learning and skill formation, which eventually affects economic growth. Park (2017) used information on 4.5 million observations from the New York City high school exit exams for the period of 1998 to 2011. He found that taking an exam on the day where the temperature is 90^0F results in a 4.4% reduction in exam performance compared to taking an exam on a day where the temperature is 72^0F (room temperature).

This paper contributes to the existing literature by studying the direct impact of extreme temperatures on young people's short-term school attendance.[2] We benefit from within-month temperature variation in a district,

[2] Several recent studies have examined the empirical link between extreme temperatures and human health. As such, current research is able to locate a range of health impacts ranging in severity from mild impacts such as cold, fever, or change in mood to much more severe impacts, such as death (Baylis, 2015;

and we work with a representative large sample of Indian households that cover a wide range of socioeconomic factors. The survey provides information on childrens' school attendance and educational expenses. Believing in the importance of school attendance to educational outcomes motivates our research to link hot daily temperature to student daily absenteeism over a thirty day period. The paper identifies a novel channel for temperature affecting total factor productivity through lower school attendance and human capital losses.

1.3 Data

To inform our research, data from two separate sources is utilized namely the second wave of the India Human Development Survey (IHDS-II) 2011/2012 and meteorological data from the European Centre for Medium-Range Weather Forecasts (ECMWF). We link both sources in order to obtain cross-sectional data. As our focus is on the impact of high temperatures on students' absence, we benefit from the within-month temperature variation in a district and restrict the final sample to cover only the period from March to November.

1.3.1 India Human Development Survey (IHDS-II)

The India Human Development Survey is a large-scale survey developed through the collaboration between the University of Maryland and the Na-

Curriero et al., 2002; Deschenes and Greenstone, 2011; Han et al., 2017; Hajat et al.,2004; Chung et al., 2009). In our paper, we implement an exercise that excludes students who reported having a fever, cough, and diarrhea, or were unable to do a normal activity in the last thirty days prior to the interview date. This exercise allows us to show that our results are not driven only by sick persons.

tional Council of Applied Economics (NCAER) in New Delhi, funded by the National Institute of Health and Ford Foundation. It is a nationally representative survey which collects data from 204,569 individuals belonging to 42,152 households inclusive of 34 states and 384 districts. The dependent variable is taken from the individual-level data which includes the number of days the student has been absent from school in the thirty day period prior to the interview date, as well as individual student characteristics such as age, gender, grade level, and the district and state of residence.

The IHDS-II is also a rich source of detailed household-level data that includes socioeconomic and demographic information in a relation to students' families. These include household unique identifier, household income level, income source type and wealth, the highest level of education for males and females in a household, and the travel time from and to the educational settings. Dummy variables represent if the household has electricity and air cooling. Table 1.1 displays the summary statistics of socioeconomic and demographic key variables. Our statistics show that 48% of the population has access to indoor running water compared to 52% that have to access external sources of water for their daily living. 47% of the students in India are females and 53% are males. Households with agriculture as the main income source account for 27% of the whole population. In contrast, the remaining households depend upon a range of income sources such as non-agricultural wage labor, petty shops, salaried, rent, etc. This percentage significantly increases in rural areas where agriculture is the main source of income for 37% of the population.

1.3.2 Weather data

The second component of the data focuses on maximum and minimum daily temperatures, daily precipitation rates, wind speed, and daily relative humidity for the period of January 1st, 2011 to May 30th, 2013. We intend to use observational weather data from ground stations as it is the most reliable data (Garg et al., 2017). However, India does not have enough ground weather stations to cover weather variations in all districts. Therefore, we use reanalysis data as well. We follow Schlenker and Roberts (2009), Schlenker and Lobell (2010), and Auffhammer et al. (2013) who used reanalysis to obtain the best weather estimate.[3] We use the daily weather factors readings from the ERA-Interim archive[4] of grid cells with 1^0 X 1^0 latitude-longitude to generate the constructed weather variables for each district. We use all the grid points that lie within a circle of radius 50 km around each district's centroid to calculate the inverse distance weighted average for temperature, humidity, wind speed, and precipitations. We map these constructed variables to districts.[5] Finally, based on students' individual interview dates and their respective districts, we map a constructed maximum and minimum daily temperatures, daily relative humidity, wind speed, and daily precipitation rates to students. Figure 1.1 presents the variation in the constructed maximum temperature.[6] Table 1.1 displays the summary statistics of the key variables

[3]Reanalysis data provides a map without gaps through data assimilation. This happens by combining past short-range weather forecasts with observations come from different sources include weather stations and remote sensing. This analysis utilizes the same data assimilation to produce data reaching back several years (ECMWF). More details can be found on ECMWF `https://www.ecmwf.int`

[4]ERA-Interim archive is constructed by European Center for Medium-Range Weather Forecasting (ECMWF). `https://www.ecmwf.int/en/forecasts.`

[5]We use the Inverse Distance Weighted Interpolation Method

[6]The figure shows the distribution of the temperature at 50^{th} percentile of maximum daily temperature for the thirty day period for the whole country. The distribution shows the country experiences extreme

that are used in the analysis.

1.3.3 Climatic zones

The concept of climate refers to integrating weather conditions for a long period that reaches back decades. Since India is a subcontinent with diverse geographic features, its climate has significant spatial and temporal variation (Bhatnager et al., 2019).

Based upon meteorological factors including temperature, precipitation and relative humidity, researchers have identified climatic zones. Each climatic zone includes several regions that represents an area larger than a state and has common climate conditions. The existence of this concept helps in identifying the severity and conditions that make individuals feel comfortable or uncomfortable (Fovell, 1993; Anderson et al., 1985; Briggs et al., 2003)

India has been divided into climatic zones namely, hot and dry, warm and humid, moderate, cold and cloudy, cold and sunny, and composite. An area is included into anyone of these zones when the specified conditions persist for more than six months (Bhatnager et al., 2019; Bansal and Minke, 1988).[7] Later on, these zones faced some adjustments through combining the cold and cloudy with the cold and sunny together generating one zone known as

heat during the period March to November.

[7]The hot and dry climate zone has the mean monthly temperature greater than 30^0C, relative humidity less than 55%, precipitation level less than 5mm, and the number of clear days greater than 20 days. The warm and humid climatic zone has the mean monthly temperature around 30^0C, relative humidity greater than 55%, precipitation level greater than 5mm, and the number of clear days less than 20 days. The moderate zone has the mean monthly temperature between 25 and 30^0C, relative humidity less than 75%, precipitation level less than 5mm, and the number of clear days greater than 20 days. The cold and cloudy zone has the mean monthly temperature less than 30^0C, relative humidity greater than 55%, precipitation level greater than 5mm, and the number of clear days less than 20 days. The cold and sunny zone has the mean monthly temperature less than 25^0C, relative humidity less than 25%, precipitation level less than 5mm, and the number clear days greater than 20 days.

the cold climatic zone (Bhatnager et al., 2019; The National Building Codes, 2005). As well, the moderate climate was renamed to become known as the temperate climatic zone (Bureau of India Standards, 2005).

Considering the differences among the climatic zones leads to the fact that change in temperature in one climatic zone might impact differently than in another climatic zone. The opportunity cost for individuals might be different as well as temperature impact through health behaviours. Therefore, in this paper, we follow the previous studies and utilize what has been done in classifying India into climatic zones. We implement India's existing climatic zones, namely, the hot and day, the cold zone, and the warm and humid zone. Figure 6 shows the three climatic zones. The blue area represents the cold zone, the light and dark orange areas represent the warm and humid zone, and the red area represents the hot and dry zone. The cold zone covers the areas that experience cooler temperatures and includes a large proportion of the Himalayan Mountains. This zone displays above average cool weather for most of the year with the temperature dropping below zero in January. The dry and hot zone is known as rocky, flat and dry. The temperature is extremely hot in summer and remains high for most of the year with very limited rainfalls. The warm and humid zone includes two regions with similar features during March to November, with the exception that one is relatively more humid than the other. The humid zone is warm, and during monsoon, which normally starts in mid-June, rainfall levels and humidity increase. The pink area represents the temperate zone; we add it to the humid zone as we find that excluding this area does not alter the regression's estimate. In Tables 1.2, 1.3, and 1.4, we display summary statistics of the variables for each of the climatic zone. The percentage of female students is the lowest in the dry climate zone (45%), and only marginally higher in the other two climatic zones (48%). Agriculture is the main source of income for 32% of families living in the hot and dry zone, 22% living in the cold zone, and 26% living in the warm and humid zone. The average walking distance to and from school is common across the three zones, which is around 2.5km.

1.3.4 Final Sample

The output file of merging the individuals and households data files linked with the calculated weather data shows that 69% of the students live in rural areas and 31% live in urban areas. We find that the temperature has no significant impact on urban-dwellers; therefore, our analysis focuses solely on rural areas. Since the aim of our research is to examine the impact of high temperatures, our final sample in the main analysis does not include the winter season.

The dependent variable is the self-report of the number of days a student was absent during the thirty day period prior to the interview. A small proportion of the reported values are not reasonable; therefore, we implement some restrictions to manage these outliers. For example, 2% of the sample provides a response greater than 21 days of absence. However, since a student can only be absent for a maximum of 21 days within a month, we nominate to exclude this 2% of students from our main analysis. Nevertheless, in the robustness check section, we show that adding these observations back to our sample does not disturb our main results.[8] Figure 1.5 shows our dependent

[8] The range of school days per year is from 200 to 230 days with the average being 215 days. Since there

variable's distribution, which is the number of non-attendance days by a student during the thirty day period prior to the interview.

1.3.5 Temperature metric

Within literature, several approaches have been used to develop a metric for temperature. An effective metric for capturing monthly heat can be developed through varied methods. In this paper, we are interested in temperature distribution as well. Therefore, we use the temperature at the 50^{th} percentile of maximum daily temperature for the thirty day period to represent our heat metric. This metric is a relative temperature measure that represents arbitrary values that vary within a district by interview date.

Indeed, India is a hot country, and as March starts, daily temperature exceeds 30^0C. Our analysis includes only the hot period of a year, which is March to November. Thus, the point at the 50^{th} percentile still represents a high value of the temperature distribution. Furthermore, the within-month temperature variation that we use is based upon the change in the maximum daily temperature. Figures 1.2 to 1.4 present the variation of maximum temperatures at the 50^{th} percentile by individuals for each climatic zone. The average values of temperature at the 50^{th} percentile for the cold, hot and dry, and the warm and humid climatic zone are 29, 32.9, and 31.7^0C, respectively.

Furthermore, we then demonstrate that when we use alternate points from the temperature distribution such as the 60^{th} and 70^{th} percentile the results remain significant. The adverse impact of hot temperature on students' ab-

is at least two months of holidays during the year, we divide 215 days into 10 months, with the answer being 21.5 days senteeism is also robust to other temperature metrics such as average monthly temperature and the number of days that maximum daily temperature exceeds 38^0C.

1.4 Methods

This paper aims to empirically examine the impact of extremely hot temperatures on school attendance. Our model's main assumption is that temperature realization is studious exogenous after accounting for temporal and spatial fixed effects. Our final sample of students is cross-sectional data. Since the dependent variable is count, our strategy allows us to implement a negative binomial regression using a simple count data model of the type

$$Y_{idt} = exp(\beta_0 + \beta_1 \, Temp_{dt} + \beta_2 \, Zone_d * Temp_{dt} + \beta_3 W_{dt} + \beta_4 X_i + \lambda_s + \theta_{dt} + \varepsilon_{idt})$$

Where, the outcome variable, Y_{idt}, represents the number of days that student i in district d reported being absent from school during the thirty day period prior to interview. To present a measure for extreme temperature, we suggest that the maximum daily temperatures at the 50^{th} percentile of the maximum temperatures distribution for the thirty day period prior to the interview is warm enough to be considered as extreme warm weather. The main source of data is collected using the period from March to November, where the daytime temperatures can be as high as 48^0C.[9] $Zone_d * Temp_{dt}$ represents interaction term between the temperature and the climate zones. It accounts for the heterogeneity of temperature impact in climatic zones. Indeed, it has

[9]See Figure 1.1

been observed that the effects of maximum temperatures vary depending upon climatic zones because individuals are able to mediate these by adapting to variations in temperature. Therefore, we control for that. W_{dt} represents weather controls, which includes average relative humidity, wind speed, and the total rainfall measured during the thirty day period. Vector X_i represents a set of household controls and individual characteristics. Household controls include the distance to school, the dummy variable represents if the household has adequate water or not and the source of drinking water. Household income level and income source types are both included in the controls, as literature showed that learning outcomes can be affected by income (Goodman et al., 2018). Dummy variables represent household with electricity or air cooling are other elements used in the vector X_i. We understand that air cooling is not randomly distributed, but it contributes to increase individual capacity to adapt (e.g., Goodman et al., 2018). Individual characteristics recorded are age, grade level, and gender. Guar and Cooper (2003) showed that when considering the gender gap, there is no significant difference students' absence. Scott and McClellen (1990) also found that the rate of absenteeism between females and males is comparable. In our paper, we find that there is a slight difference between genders. λ_s is a state fixed effect. θ_{dt} is a month interaction with climate zone fixed effect.[10] ε_{idt} is the error term.

[10] Replacing the month interaction climate zone term with the month of year fixed effect provides significant results. These are smaller in magnitude and drop to 5.5%. The same effect occurs when adding the Zone fixed effect; see Table 1.5

1.5 Main results

Table 1.5 presents the main result of a high-temperature impact on students' absence in rural areas. We build the columns by adding controls in sequence. Reading across the columns, we observe that the estimated coefficients of temperature are fairly stable and statistically significant at a 1% level of significance in most columns. Column 8 includes the full suite of controls and is the preferred specification. The average partial effect of an increase in temperature is associated with an increase of the expected count number of absent days by 0.4 days.[11]

Table 1.6 is the only table that presents the impact of temperature on students' absence in urban areas. Column 8 includes the coefficient estimate of the full regression. The coefficient estimate is -2% and does not obtain statistical significance at conventional levels, albeit on a much smaller sample. As such, an increase in temperature in urban areas does not have a significant impact on students' absence in urban areas; it can be put forward that students in urban areas live closer to school, present from higher income households, have access to better school facilities, and experience other advantages that rule out the impact of hot temperatures. As an attempt to look at heterogeneity in urban areas by income level, availability of drinkable water, and access to air-cooling, we could not document any significant impact on any one of these sub-populations. Thus, the rest of results only represent the impact in rural areas.

[11] Poisson family is a count model with exponential conditional mean function, $E[y|x] = exp(x'\beta)$. For the purpose of interpreting the results, we take the partial derivative $\partial E[y|x]/\partial xj = \beta_j E[y|x]$ $\Rightarrow \partial \mathrm{E[y|x]}/\partial \mathrm{xj} = \beta_j exp(x'\beta)$, where x_j represents temperature in Celsius, which is our variable of interest and x is a vector of all included variables. Then, when averaging the whole sample we obtain $1/N \sum_{i=1}^{N} \partial E[y_i|x]/\partial xj = \widehat{\beta}_j \bar{y}$.

Table 1.7 shows that excluding students that are seasonal migrants, students who were unable to undertake normal activities, and students who experienced breathing problems, fever, diarrhea, or cough during the thirty day period does not impact the estimate that we generated in Table 1.5. Firstly, when considering seasonal migrants, it must be noted that many students come from households who rely on agriculture as their primary source of income. In understanding this, research suggest that crop yields do not adapt to climate change (Schlenker and Roberts, 2009). Therefore, these students might be forced to temporarily move from their districts to another state due extreme weather events including flood or drought. When relocating, these students may also be unable to attend school for several days. As a result, to test whether seasonal migrants solely drive the estimate of our analysis, in column 1, we exclude this category of students. However, the estimate remains statistically significant at 1% level and around the same magnitude of impact. Secondly, in column 2, we exclude students who were unable to undertake normal activities during the study period. We assume that these students may be absent from school due to external forces or reasons independent of hot temperatures such as accidents. However, the regression estimate also remains consistent with the preferred regression estimate in Table 1.5 and significant at 1% level. Lastly, in column 3, we show that excluding students experienced breathing problems, fever, diarrhea, or cough from the sample does not change the estimate's level of significance. Despite including this category of students with these aforementioned conditions, our preferred specification estimate remains statistically significant at the 1% level. In other words, hot temperatures retain its causal impact on students' absence.

In Table 1.5, the model assumes that all individuals have an homogeneous response to hot temperature; however, categorizing students into different groups based upon several factors such as income level, income source, gender, school system level, geography, and other resources shows that students respond differently to an increase in temperature.

1.5.1 Heterogeneity by income level, income source, and gender

Table 1.8 shows heterogeneity in students' absence responses to hot temperatures in relations to income level, type of income source, and gender. Columns 1 and 2 show students' responses based on income level.[12] In our data, students are classified into poor versus non-poor based upon the Tendulkar committee's methodology.[13] This methodology uses the expenditure per capita to categorise individuals into poor versus non-poor. If expenditure per capita is less than the poverty line, the individual is classified as being poor. The estimates in these two columns show that responses to hot temperatures are statistically significant for both groups. This result is consistent with Garg et al. (2017), findings which indicate that high temperatures have a significant negative impact on test scores, in particular for the poor sub-group.

In India, agricultural labor is a source of income for many students' households. It must be noted that several of these students will eventually work on their families' farms or undertake paid work at another farm. As such,

[12] When introducing controls sequentially in the regressions by income level, we could not document any statistical differences across income groups. The coefficient estimate from the regression without controls for the poor group is 0.0886 with standard error (0.0373), and the coefficient estimate from the regression without controls for the non-poor group is 0.0958 with standard error (0.0337).

[13] See the Government of India Poverty Report: http://planningcommission.nic.in/reports/genrep/pov_rep0707.pdf

income from agriculture might be a factor driving a percentage of students' absence. Columns 3 and 4 classify students based on the primary type of income source of their households, including agricultural income source, compared to other income sources; nevertheless the estimates maintain the same level of significance and impact in both cases.

In columns 5 and 6, the estimates for male and female students are statistically significant at 1% level, and the average responses to an increase in temperature for both genders are similar. Our findings are consistent with Scott and McClellan (1990) who argue that the rate of absenteeism across genders is comparable.

1.5.2 Electricity, air cooling, and water

Socio-economic factors such as access to electricity, air cooling, and drinking water are essential indicators that may assist in understanding students' absence related to variations in temperature. When looking at Table 1.9, columns 2 and 3, it can be observed that the estimates indicate that students from rural areas living in households with electricity show a statistically significant response to temperature increase. Meanwhile, students living in households without access to electricity do not show a significant response to increases in temperature.

Data confirms that access to air cooling is not randomly distributed between households. As such, columns 4 and 5 represent the heterogeneity in students' response to changes in temperature based on air cooling availability in their homes. The impact is statistically significant for both groups; however, the magnitude of the impact is slightly higher for those who have air cooling at home.

Accessibility to drinking water in students' homes is critical during hot days. As such, it is worth investigating whether the differences in water quality and water accessibility generate a heterogeneous effect in students' absence response to heat. Columns 6 and 7 classify rural students into two groups; students living in households with access to clean drinking water and students living in households with compromised drinking water. The estimate in column 6 shows that an increase in temperatures leads to a statistically significant impact on the rate of absenteeism of students. We however find that this impact is not significant for students with compromised drinking water and the magnitude of the impact is smaller. Also, the nature of the household water supply source should impact students' responses to an increase in temperatures; however, the results do not support this. Columns 8 and 9 classify students into two groups, students living in homes with piped water sources and students accessing other sources of water such as rivers, wells, etc. Our findings show that the differences in water source supplies do not bring a heterogeneous impact on students' responses to an increase in temperatures.

1.5.3 Heterogeneity by school level

India's four school levels for children of ages six to eighteen years old are the primary level, the upper primary level, the secondary level, and higher secondary level.[14] The primary level includes grades one to five. The up-

[14] See the report of Selected Information on School Education from the Government of India, Ministry of Human Resources: `http://mhrd.gov.in/sites/upload_files/mhrd/files/statistics/SISH201112.pdf`

per primary level includes grades six to eight. The secondary level includes grades nine and ten, and the higher secondary level includes grades eleven and twelve. These four levels can be aggregated into two levels; grades one to eight constituting the elementary level and grades nine to twelve encompassing the secondary level.

In Table 1.10, we show the heterogeneous effect of high temperature by school level. Column 3 presents the effect on students at the elementary level, while column 6 presents the effect on students at the secondary school level. For both levels, an increase in temperatures has a statistically significant impact on students' rate of absenteeism. However, there is a difference in the magnitude of the temperature effect between the two levels. For the secondary level, the average partial effect of an increase in high temperatures is 0.52 days increase in the expected school days absence. Alternatively, for students at the elementary level, the average partial effect is only 0.24 days.

1.5.4 Heterogeneity by geography

Table 1.11 depicts the heterogeneous impact of students' absenteeism rate to an increase in hot temperatures based on geographical locations. As such, students are assigned to one of the three climatic zones depending on where their household is located. As discussed previously, these are the cold zone, the hot and dry zone, and the warm and humid zone that are presented in columns 1, 2, and 3, respectively. The estimate in column 2 shows that an increase in temperature has a statistically significant impact on students' absence. In contrast, the impact is not significant in the cold zone and the warm and humid zone.

In the hot and dry zone, the average partial effect of an increase in temperature is 0.17 days increase of the expected number of days a student will be absent. This impact is economically and statistically significant. While the average partial effect in the cold and the humid climatic zones are 0.01 and 0.03 days, respectively, these magnitudes are neither economically nor statistically significant. The negligible impacts in these two zones are consistent with the findings of Schlenker and Roberts (2009), which state that humans can independently adapt to increases in temperatures.

1.6 Robustness checks

This section sheds light on emerging concerns regarding the preferred (main) specification presented in Table 1.5 by answering some of the "what if" questions. These questions are related to different aspects including the sample restrictions, alternative standard errors, implementing different temperature metrics, and using Zero Inflated Negative Binomial regression. All these robustness exercises do not falsify the results of the main specification in Table 1.5 and confirm the rigidity of the main analysis results.

1.6.1 Restrictions

In Table 1.12, column 1, we present again the main specification of this paper, which implements the following restrictions: (a) The sample excluded urban areas such that the main analysis was implemented only on rural areas.

(b) The sample excluded students who missed more than twenty-one days of school during the thirty day period prior to the interview. (c) The time frame of the analysis was restricted to the period from March to November.

This section shows that relaxing some of the aforementioned restrictions by adding the excluded individuals back into the sample does not result in a considerable change in the effect, and the estimate remains significant. For example, Column 4 represents the case of adding students who missed more than twenty-one days to our sample. Although the results indicate that the estimate is slightly smaller in magnitude, it is statistically significant at a 1% level of significance.

Winsorization is another robust statistic method for dealing with outliers in a distribution of data rather than trimming. In column 7, we winsorize all the data points of values greater than twenty-one school days missed by converting them to twenty-two days of absence as the next reasonable value after twenty-one. Fortunately, those data points are just around 2% of the sample. The results show that the estimates of non-winsorized and winsorized samples have the same significance level; however, the latter has a slightly higher estimate.

Column 6 shows that including the winter period (December to February) in our sample does not reduce the statistical significance level of our preferred specification estimate. The drop in the magnitude of the impact is minimal. Students miss relatively fewer school days during winter; this is likely due to the more comfortable temperatures during the winter months. Our result is consistent with the findings of Baylis (2015) that individuals can make personal adjustments to mitigate the impact of cooler weather such as putting on more clothes, effectively making a difference between the margin of adjustment to high and low temperatures.

Column 5 shows the result of trimming all data points of the dependent variable that are greater than ten days of absenteeism.[15] The percentage of those points forms around 5% of our sample, as shown in Figure 5. Trimming these values has limited impact upon the preferred specification result.

1.6.2 Alternative standard errors

Obtaining proper standard errors is important for statistical inference. Good standard errors contribute to identifying the significance of the reported estimate and construct a more precise confidence interval. This is achieved by clustering observations into groups, where model errors are correlated within the individual group but uncorrelated across them (Colin Cameron and Miller, 2015). In our analysis, we clustered all standard errors at the district level.

In Table 1.12, we propose clustering standard errors at different levels including district by month, district by season of the year,[16] and state-level, in columns 8, 9, and 10, respectively. When comparing each result of these three

[15] See the report of Selected Information on School Education from Government of India, Ministry of Human Resources: `http://mhrd.gov.in/sites/upload_files/mhrd/files/statistics/SISH201112.pdf`. 186 days is the lowest value of the total number of school days and, on average, there are two months of holidays between semesters. Based on these variables, we get 18.6, which is the average schooling days within a given month. If we subtract the school break and total value of weekends from 365 days, we obtain 225 school days in a year. 225/30 we get 7.5 months as the valid school calendar year. According to Indian truancy law, a student cannot be absent for more than 25% of the school year. Using the maximum value of the total number of school days to calculate student disengagement threshold, we obtain 230*0.25 = 57.5 then 57.5/7.5 ≈ 8 days, which means that the maximum average monthly absence should not exceed 8 days. Based on these calculations, students should attend at least an average of ten days.

[16] Regarding temperatures and precipitations, a year is divided into winter (December to February), summer (March to May), and monsoon (June to November).

exercises with the result in the preferred specification, where we clustered at the district level, we find that the differences are not statistically significant.

1.6.3 Average temperature, two temperature bins and different temperature percentiles

We use the average maximum daily temperatures over the thirty day period as an alternate temperature metric. The result shows that an increase in average temperature has a statistically significant impact on students' absenteeism rates, as shown in Table 1.12, column 2.

Rather than temperature distribution, we use another temperature metric represents two temperature bins. In this metric, we compare the number of days where the maximum daily temperature exceeds 38^0C to the number of days where the temperature is less than 38^0C during the thirty day period. In Table 1.12, column 3, we show the results obtained using this metric. The impact is statistically significant at 1% level of significance.[17]

In Table 1.13, we show the impact of taking different points from the temperature distribution such as temperature at 40^{th}, 60^{th}, 80^{th} percentile during the thirty day period. All these temperature points have statistically significant impacts on the students attendance.

[17]This temperature metric of the number of hot days above 38^0C has a significant impact of 2 percentage point in the cold zone. Since this zone is know called all year around, then having a day with temperature above 38^0C might work as a big shock. Though our focus in this paper focuses on using a metric represents distribution of temperature over the thirty day period..

1.6.4 Placebo test

Table 1.14 shows the results of the placebo test compared to the preferred specification. In this placebo test, we use the same restrictions implemented in our final sample in Table 1.5. However, in this placebo, we are falsely redistributing the constructed temperature among individuals. Column 2 in Table 1.14 presents the result of this placebo test. The estimate is statistically insignificant, and the magnitude of the effect is neglected. The average partial effect of an increase in falsely matched temperature is 0.03 days increase in the expected student's school days absence.

The strategy of this placebo test depends upon mismatching the values of temperature among individuals. As such, it assigns the temperature value from an individual in a certain district to another individual in a different district and state. We randomly mismatch the temperature values to students by reversing the order of the districts. This false matching generates an insignificant estimate and a neglected impact on students' absenteeism, and the results support the rigidity of our work in the main analysis.

1.7 Conclusion

High weather temperatures affect children, who will be the main driver of economic growth in the near future. The impact of high temperatures is evident when further analyzing current data related to children's health outcomes, cognitive ability and levels of concentration, and overall academic performances. An increasing number of studies on the effect of variations in temperatures on academic performance have been conducted including test scores in math and reading. However, our paper is the first study that analyses the impact of extremely hot temperatures on short-term students' absenteeism rates. India was nominated as the focus of our research due to extreme variations in climate with temperatures ranging 30^0C, and peaking at 45^0C from March onwards. Simultaneously, the country suffers from a lack of infrastructure, with millions of people living without electricity and clean and accessible drinking water.

In this paper, we identify a novel channel for temperature affecting total factor productivity through lower school attendance and human capital losses. We utilize a simple count model that identifies the effect of high temperatures on rates of absenteeism in rural and urban areas separately. We shed some light on the heterogeneity of the effect of temperature by climate zone. Although, our results show that an increase in temperature is not statistically significant in urban areas, our results indicate that high ambient temperatures affect students' willingness and ability to attend school in rural areas. The average partial effect of an increase in high temperatures is 0.4 days increase in predicted rates of students' absences. Our results are also consistent with the findings from recent literature on the effect of hot temperatures on students' academic performance and indicates that the heat effect is not different across gender. This paper also highlights that those students who have access to electricity within their households have higher rates of absenteeism. Access to cooling also impacts upon students' absenteeism although we demonstrate that heat has a significant impact upon students who both live with and without air conditioning. In summary, our results show that an increase in temperature is not statistically significant in urban areas; however, the impact is significant in rural areas. It must be noted that individuals living in rural areas are already significantly disadvantaged in relation to socio-economic infrastructure, which results in reduced access to lifelong quality outcomes.

Recently, the Indian government in collaboration with not-for-profit organizations have started to focus on implementing frameworks to develop rural areas. Therefore, future research on students' absenteeism could provide more interesting and up to date findings by utilizing data currently being generated. It would also be beneficial to consider other related factors such as students' mental health as well as specific information around school environments. In impoverished communities, access to air cooling remains too expensive and as a result is not a viable option. Therefore providing information around the volume of air cooling usage might be a key angle for future work in exploring its long-term impact on students' absence.

References

Andersson, B., Carrol, W. L., & Martin, M. (1985). Aggregation of U.S. Population centers into climate regions. *ASHRAE Transactions* 91, 2, 109-121. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt9hw07906.pdf

Auffhammer, M., Hsiang, S. M., Schlenker, W., & Sobel, A. (2013). Using weather data and climate model output in economic analyses of climate change. *Review of Environmental Economics and Policy*, 7(2), 181-198. https://doi.org/10.1093/reep/ret016

Bhatnagar, M., Mathur, J., & Garg, V., (2019). Climate zone classification of India using new base temperature. *International Building Performance Simulation Association*. http://www.ibpsa.org/proceedings/BS2019/BS2019_211159.pdf

Bansal, N. K, & Minke, G. (1988). Climatic zones and rural housing in India, (199004). https://doi.org/citeulike-article-id:10601847

Baylis, P. (2015). Temperature and temperament: Evidence from a billion tweets. *Energy Institute Working Paper*.

Briggs, R. S., Lucas, R. G., & Taylor, Z. T. (2003). Climate classification for building energy codes and standards: Part 1-development process. *Transactions-American Society of Heating Refrigerating and Air Conditioning Engineers*, 109(1), 109-121.

Cai, X., Lu, Y., & Wang, J. (2018). The impact of temperature on manufacturing worker productivity: evidence from personnel data. *Journal of Comparative Economics*, 46(4), 889-905. https://doi.org/10.1016/j.jce.2018.06.003

Chen, D., & Chen, H. W. (2013). Using the Köppen classification to quantify climate variation and change: An example for 1901-2010. *Environmental Development*, 6, 69-79. https://doi.org/10.1016/j.envdev.2013.03.007

Chirakijja, J., Jayachandran, S., & Ong, P. (2019). Inexpensive heating reduces winter mortality. (Working paper No. w25681). Retrieved from https://doi.org/10.3386/w25681

Chung, J.-Y., Honda, Y., Hong, Y.-C., Pan, X.-C., Guo, Y.-L., & Kim, H. (2009). Ambient temperature and mortality: An international study in four capital cities of East Asia. *Science of The Total Environment*, 408(2), 390-396. https://doi.org/10.1016/j.scitotenv.2009.09.009

Colin Cameron, A., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317-372. https://doi.org/10.3368/jhr.50.2.317

Colmer, J. (2013). Climate variability, child labor and schooling: Evidence on the intensive and extensive margin. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2334819

Curriero, F. C., Heiner, K. S., Samet, J. M., Zeger, S. L., Strug, L., & Patz, J. A. (2002). Temperature and mortality in 11 cities of the Eastern United States. *American Journal of Epidemiology*, 155(1), 80-87. https://doi.org/10.1093/aje/155.1.80

Dell, M., Jones, B., & Olken, B. (2008). Climate change and economic growth: evidence from the last half century. *Cambridge, MA: National Bureau of Economic Research* (working paper No. w14132). Retrieved from https://doi.org/10.3386/w14132

Deschênes, O., & Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *American Economic Journal: Applied Economics*, 3(4), 152-185. https://doi.org/10.1257/app.3.4.152

Desai, Sonalde, and Reeve Vanneman and National Council of Applied Economic Research, New Delhi. India Human Development Survey-II (IHDS-II), 2011-12. ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-07-31. http://doi.org/10.3886/ICPSR36151.v2

Garg, T., Jagnani, M., & Taraz, V. (2017). Effects of heat stress on physiology and livelihoods: implications for human capital accumulation. Working paper.

Geruso, M., & Spears, D. (2018). Heat, humidity, and infant mortality in the developing world. *Cambridge, MA: National Bureau of Economic Research* (working paper No. w24870). Retrieved from https://doi.org/10.3386/w24870

Goodman, J., Hurwitz, M., Park, J., & Smith, J. (2018). Heat and learning. *Cambridge, MA: National Bureau of Economic Research* (working paper No. 24639). Retrieved from https://search.proquest.com/docview/2064752680?accountid=14701

Graff Zivin, J., Hsiang, S. M., & Neidell, M. (2017). Temperature and human capital in the short and long run. *Journal of the Asso-*

ciation of Environmental and Resource Economists, 5(1), 77-105. https://doi.org/10.1086/694177

Graff Zivin, J., & Neidell, M. (2010). Temperature and the allocation of time: implications for climate change. *National Bureau of Economic Research* (working paper No. 15717). Retrieved from http://www.nber.org/papers/w15717

Hajat, S., Armstrong, B., Gouveia, N., & Wilkinson, P. (2004). Comparison of mortality displacement of heat-related deaths in Delhi, Sao Paulo and London. *Epidemiology*, 15(4). doi:10.1097/00001648-200407000-00235

Han, J., Liu, S., Zhang, J., Zhou, L., Fang, Q., Zhang, J., & Zhang, Y. (2017). The impact of temperature extremes on mortality: A time-series study in Jinan, China. *BMJ Open*, 7(4), e014741. https://doi.org/10.1136/bmjopen-2016-014741

Heyes, A., & Saberian, S. (2019). Temperature and decisions: Evidence from 207,000 court cases. *American Economic Journal: Applied Economics*, 11(2):238-265.

Kanayo, O. (2013). The impact of human capital formation on economic growth in Nigeria. *Journal of Economics*, 4(2), 121-132. https://doi.org/10.1080/09765239.2013.11884972

Kingdon, G. G. (2007). The progress of school education in India. *Oxford Review of Economic Policy*, 23(2), 168-195. https://doi:10.1093/icb/grm015

Lin, T.-F., & Chen, J. (2006). Cumulative class attendance and exam performance. *Applied Economics Letters*, 13(14), 937-942. https://doi.org/10.1080/13504850500425733

Monguno, A., Jimme, M., & Monguno, H. (2017). The influence of weather on students' academic performance in Kashim Ibrahim College of Education, Maiduguri, Nigeria. *Current Journal of Applied Science and Technology*, 23(4), 1-9. https://doi.org/10.9734/CJAST/2017/33559

Park, J. (2017). Temperature, test scores, and human capital production.

Radakovic, S. S., Maric, J., Surbatovic, M., Radjen, S., Stefanova, E., Stankovic, N., & Filipovic, N. (2007). Effects of acclimation on cognitive performance in soldiers during exertional heat stress. *Military Medicine*, 172(2), 133-136. https://doi.org/10.7205/MILMED.172.2.133

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37), 15594-15598. https://doi.org/10.1073/pnas.0906865106

Schlenker, Wolfram, & Lobell, D. B. (2010). Robust negative impacts of climate change on African agriculture. *Environmental Research Letters*, 5(1), 014010. https://doi.org/10.1088/1748-9326/5/1/014010

Xu, Z., Etzel, R. A., Su, H., Huang, C., Guo, Y., & Tong, S. (2012). Impact of ambient temperature on childrens health: A systematic review. *Environmental Research*, 117, 120-131. https://doi.org/10.1016/j.envres.2012.07.002

Zhang, P., Deschenes, O., Meng, K., & Zhang, J. (2018). Temperature effects on productivity and factor reallocation: Evidence from a half million chinese manufacturing plants. *Journal of Environmental Economics and Management*, 88, 1-17. https://doi.org/10.1016/j.jeem.2017.11.001

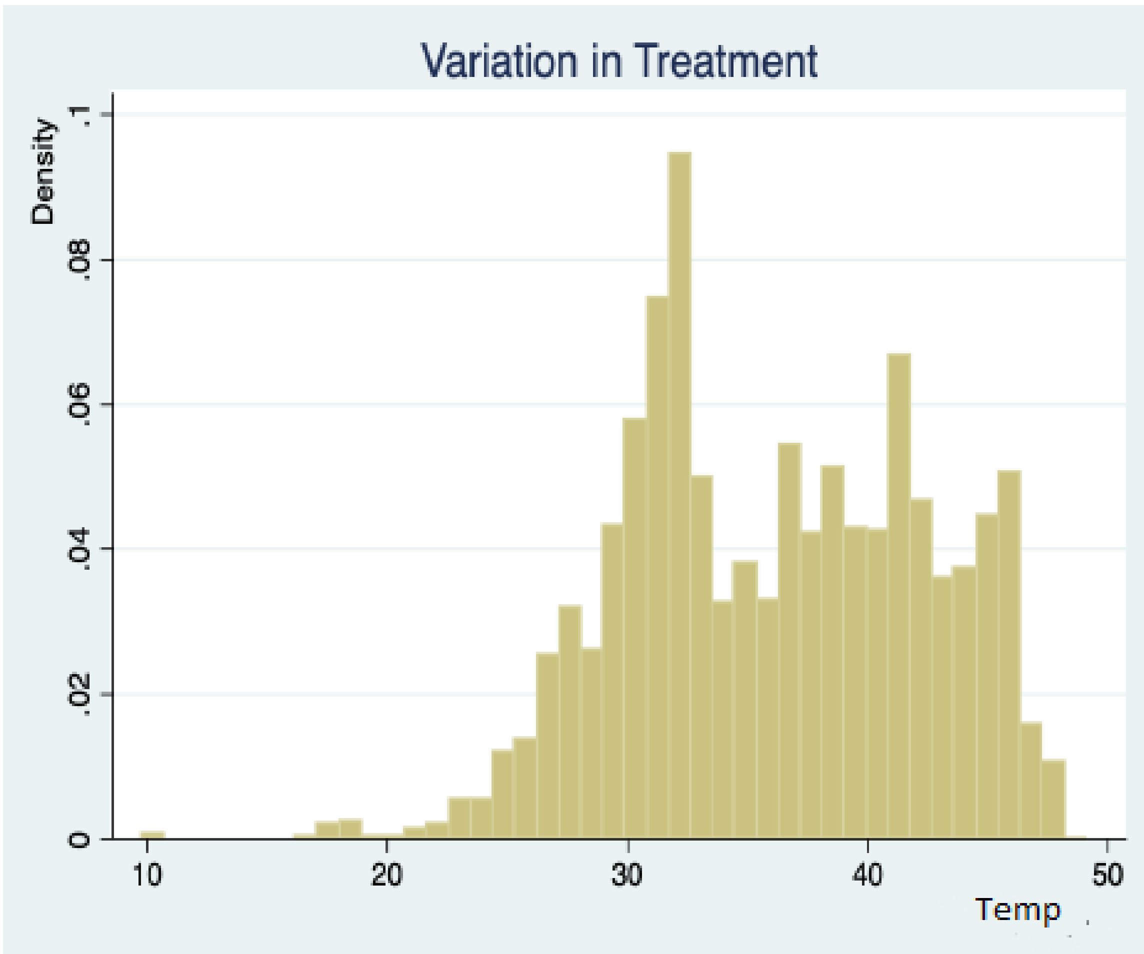

Figure 1.1: Variation in treatment.

Table 1.1: Descriptive statistics for the whole country

Variable	Obs	Mean	Std. Dev.	Min	Max
Days absence (day)	39915	3.455994	4.767978	0	30
Tmx50P (^{0}C)	39731	31.76946	6.882918	1.802808	47.41109
Tmx60P (^{0}C)	39731	32.38559	6.894887	2.482624	47.63154
Tmx70P (^{0}C)	39731	33.01276	6.916656	3.250964	48.2073
Tmx80P (^{0}C)	39731	33.71311	6.902581	3.873272	49.10981
Tmx90P (^{0}C)	39731	34.60181	6.888323	4.275944	50.2244
TmaxAvg (^{0}C)	39731	31.64196	6.779683	1.726829	47.57296
Tmax38 (day)	39731	6.704009	10.76528	0	30
Total amount of rain (mm)	39731	124.0041	249.3835	0.000617	2222.028
Average relative humidity (%)	39731	0.51882	0.219005	0.112994	0.971124
Average wind speed (km/h)	39731	2.63279	0.8639834	0.962615	7.123833
Urban	39915	0.31579	0.464838	0	1
Electricity	39896	0.86039	0.346589	0	1
Household highest male education	37403	7.39823	4.883809	0	16
Household highest female education	39717	5.08125	4.985462	0	16
Air Cooling	39888	0.18221	0.386023	0	1
Adequate water	39602	0.93291	0.250186	0	1
Gender	39915	0.52379	0.499324	0	1
Age (year)	39915	11.6421	3.438349	6	19
Distance to school (km)	39915	2.61082	5.534561	1	90
Household monthly income (Rupees)	39915	10747.54	20415.53	0	946667

Notes: Days absence represents the number of days an individual report being absent from school in the last thirty day period prior to interview date. Tmx#P represents the maximum temperature at # percentile of the thirty day period prior to the individual interview date. It is a continuous measure. TmaxAvg is a continuous measure represents the average maximum daily temperatures for the same period. Tmax38 represents the number of days that the maximum daily temperatures exceed 38^{0}C. Urban is a dummy variable represents whether an individual live in urban area. Electricity is a dummy variable represents whether an individual has access to electricity. Household highest male education (year). Household highest female education (year). Air Cooling is a dummy variable represents if an individual has air conditioning. Adequate water is a dummy variable represents availability of drinkable water. Gender represent a dummy variable male/female. Age is a continuous variable represents an individual's age (year). Distance to school is a continuous variable measures the distance to school in kilometres. Household monthly income (Indian rupees)

Table 1.2: Descriptive statistics for climate zone one "Cold"

Variable	Obs	Mean	Std. Dev.	Min	Max
Days absence (day)	2995	2.55793	4.486106	0	30
Tmx50P (^{0}C)	2994	29.52159	7.969634 (^{0}C)	1.802808	43.71073
Tmx60P (^{0}C)	2994	30.32575	8.04687	2.482624	44.25544
Tmx70P (^{0}C)	2994	31.11973	8.026803	3.250964	45.21212
Tmx80P (^{0}C)	2994	31.84803	7.978986	3.873272	45.66587
Tmx90P (^{0}C)	2994	32.83352	7.985868	4.275944	46.89494
TmaxAvg (^{0}C)	2994	29.28983	7.807661	1.726829	42.15995
Tmax38 (day)	2994	3.132265	6.313018	0	28
Total amount of rain (mm)	2994	132.2808	256.9127	1.07105	1263.691
Average relative humidity (%)	2994	0.407059	0.206471	0.151939	0.91961
Average wind speed (km/h)	2994	2.836778	0.612243	1.413077	3.774445
Urban	2995	0.277463	0.447822	0	1
Electricity	2995	0.988982	0.104406	0	1
Household highest male education	2673	8.591471	4.454775	0	16
Household highest female education	2986	6.552914	5.001847	0	16
Air Cooling	2995	0.127546	0.333639	0	1
Adequate water	2971	0.885897	0.317990	0	1
Gender	2995	0.518864	0.499727	0	1
Age (year)	2995	12.04274	3.659640	6	19
Distance to school (km)	2995	2.626377	3.071021	1	35
Household monthly income (Rupees)	2995	15232.26	18951.88	0	319833

Notes: Days absence represents the number of days an individual report being absent from school in the last thirty day period prior to interview date. Tmx#P represents the maximum temperature at # percentile of the thirty day period prior to the individual interview date. It is a continuous measure. TmaxAvg is a continuous measure represents the average maximum daily temperatures for the same period. Tmax38 represents the number of days that the maximum daily temperatures exceed 38^{0}C. Urban is a dummy variable represents whether an individual live in urban area. Electricity is a dummy variable represents whether an individual has access to electricity. Household highest male education (year). Household highest female education (year). Air Cooling is a dummy variable represents if an individual has air conditioning. Adequate water is a dummy variable represents availability of drinkable water. Gender represent a dummy variable male/female. Age is a continuous variable represents an individual's age (year). Distance to school is a continuous variable measures the distance to school in kilometres. Household monthly income (Indian rupees)

Table 1.3: Descriptive statistics for climate zone two "Hot and Dry"

Variable	Obs	Mean	Std. Dev.	Min	Max
Days absence (day)	7479	2.809333	4.349503	0	30
Tmx50P (^{0}C)	7471	32.86947	7.076773	20.43419	45.0742
Tmx60P (^{0}C)	7471	33.35633	6.992217	20.66454	45.64412
Tmx70P (^{0}C)	7471	33.88704	6.929421	21.09357	46.2275
Tmx80P (^{0}C)	7471	34.51674	6.748908	21.65361	46.77435
Tmx90P (^{0}C)	7471	35.20603	6.59768	22.44633	47.42574
TmaxAvg (^{0}C)	7471	32.73902	7.015708	20.17999	45.23304
Tmax38 (day)	7471	10.10321	12.67411	0	30
Total amount of rain (mm)	7471	22.90238	84.87623	0.000617	1048.152
Average relative humidity (%)	7471	0.383604	0.157384	0.112994	0.879228
Average wind speed (km/h)	7471	2.804192	0.793034	1.422456	5.551222
Household highest male education	7128	7.862093	4.769287	0	16
Household highest female education	7456	5.028568	4.878248	0	16
Urban	7479	0.328787	0.469804	0	1
Electricity	7475	0.922809	0.266912	0	1
Air Cooling	7477	0.254246	0.435466	0	1
Adequate water	7458	0.902789	0.296265	0	1
Gender	7479	0.548603	0.497666	0	1
Age (year)	7479	11.54218	3.388392	6	19
Distance to school (km)	7479	2.514106	5.052337	1	90
Household monthly income (Rupees)	7479	11770.83	16997.43	0	241761

Notes: Days absence represents the number of days an individual report being absent from school in the last thirty day period prior to interview date. Tmx#P represents the maximum temperature at # percentile of the thirty day period prior to the individual interview date. It is a continuous measure. TmaxAvg is a continuous measure represents the average maximum daily temperatures for the same period. Tmax38 represents the number of days that the maximum daily temperature exceeds 38^{0}C. Urban is a dummy variable represents whether an individual live in urban area. Electricity is a dummy variable represents whether an individual has access to electricity. Household highest male education (year). Household highest female education (year). Air Cooling is a dummy variable represents if an individual has air conditioning. Adequate water is a dummy variable represents availability of drinkable water. Gender represents a dummy variable male/female. Age is a continuous variable represents an individual's age (year). Distance to school is a continuous variable measures the distance to school in kilometres. Household monthly income (Indian rupees).

Table 1.4: Descriptive statistics for climate zone three "Warm and Humid"

Variable	Obs	Mean	Std. Dev.	Min	Max
Days absence (day)	29441	3.711627	4.869903	0	30
Tmx50P (^{0}C)	29266	31.71861	6.648901	16.0941	47.41109
Tmx60P (^{0}C)	29266	32.34851	6.690389	17.17082	47.63154
Tmx70P (^{0}C)	29266	32.98324	6.74826	17.71997	48.2073
Tmx80P (^{0}C)	29266	33.69875	6.784184	18.75035	49.10981
Tmx90P (^{0}C)	29266	34.62847	6.809503	19.56791	50.2244
TmaxAvg (^{0}C)	29266	31.60253	6.536641	15.58734	47.57296
Tmax38 (day)	29266	6.201667	10.38017	0	30
Total amount of rain (mm)	29266	148.9666	269.4491	0.001717	2223.028
Average relative humidity (%)	29266	0.564771	0.2153488	0.130192	0.971124
Average wind speed (km/h)	29266	2.568171	0.8936449	0.962615	7.123833
Household highest male education	27602	7.16289	4.927251	0	16
Household highest female education	29275	4.94456	4.986869	0	16
Urban	29441	0.31639	0.465077	0	1
Electricity	29426	0.83144	0.374368	0	1
Air Cooling	29416	0.16947	0.375169	0	1
Adequate water	29173	0.94539	0.227212	0	1
Gender	29441	0.521144	0.499561	0	1
Age (year)	29441	11.6266	3.425031	6	19
Distance to school (km)	29441	2.63381	5.838093	1	90
Household monthly income (Rupees)	29441	10030.01	21268.24	0	946666

Notes: Days absence represents the number of days an individual report being absent from school in the last thirty day period prior to interview date. Tmx#P represents the maximum temperature at # percentile of the thirty day period prior to the individual interview date. It is a continuous measure. TmaxAvg is a continuous measure represents the average maximum daily temperatures for the same period. Tmax38 represents the number of days that the maximum daily temperature exceeds 38^{0}C. Urban is a dummy variable represents whether an individual live in urban area. Electricity is a dummy variable represents whether an individual has access to electricity. Household highest male education (year). Household highest female education (year). Air Cooling is a dummy variable represents if an individual has air conditioning. Adequate water is a dummy variable represents availability of drinkable water. Gender represents a dummy variable male/female. Age is a continuous variable represents an individual's age (year). Distance to school is a continuous variable measures the distance to school in kilometres. Household monthly income (Indian rupees)

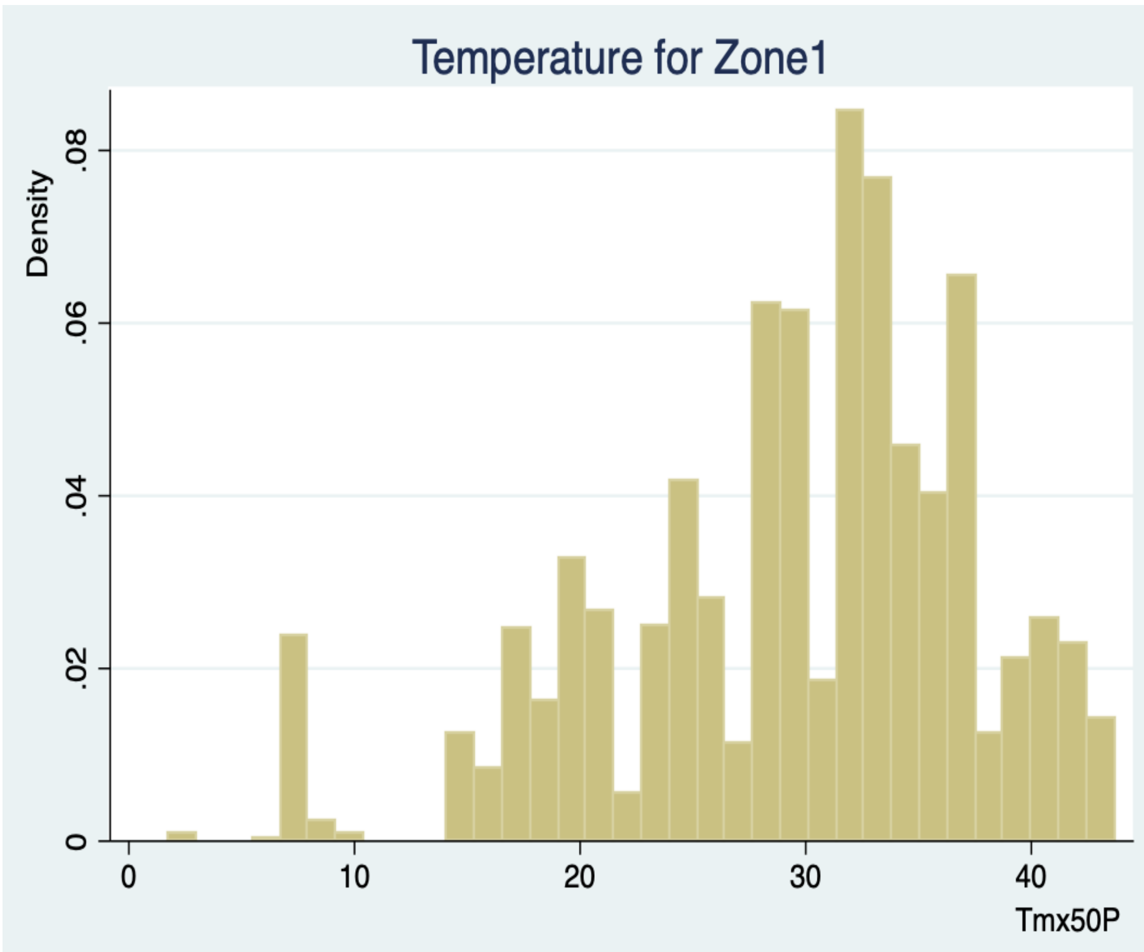

Figure 1.2: Variation in treatment for zone 1.

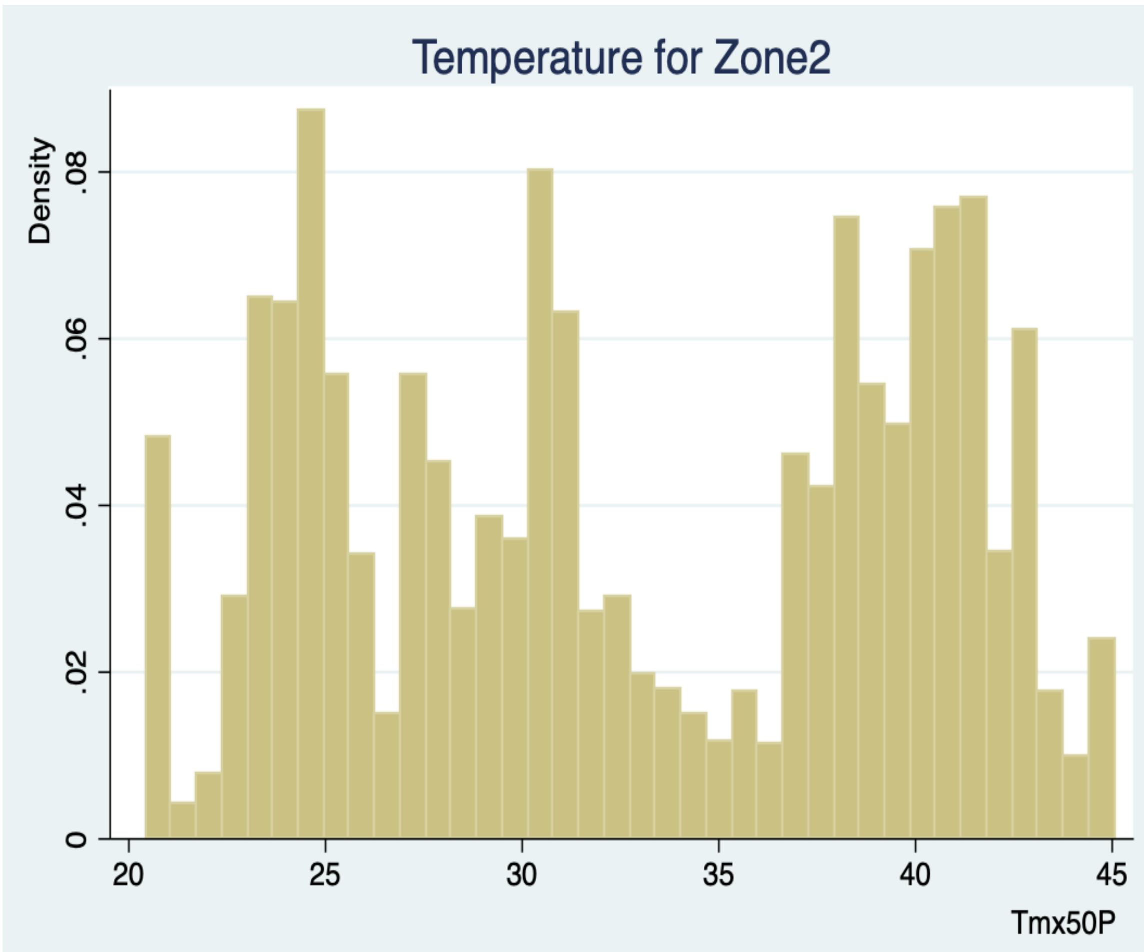

Figure 1.3: Variation in treatment for zone 2.

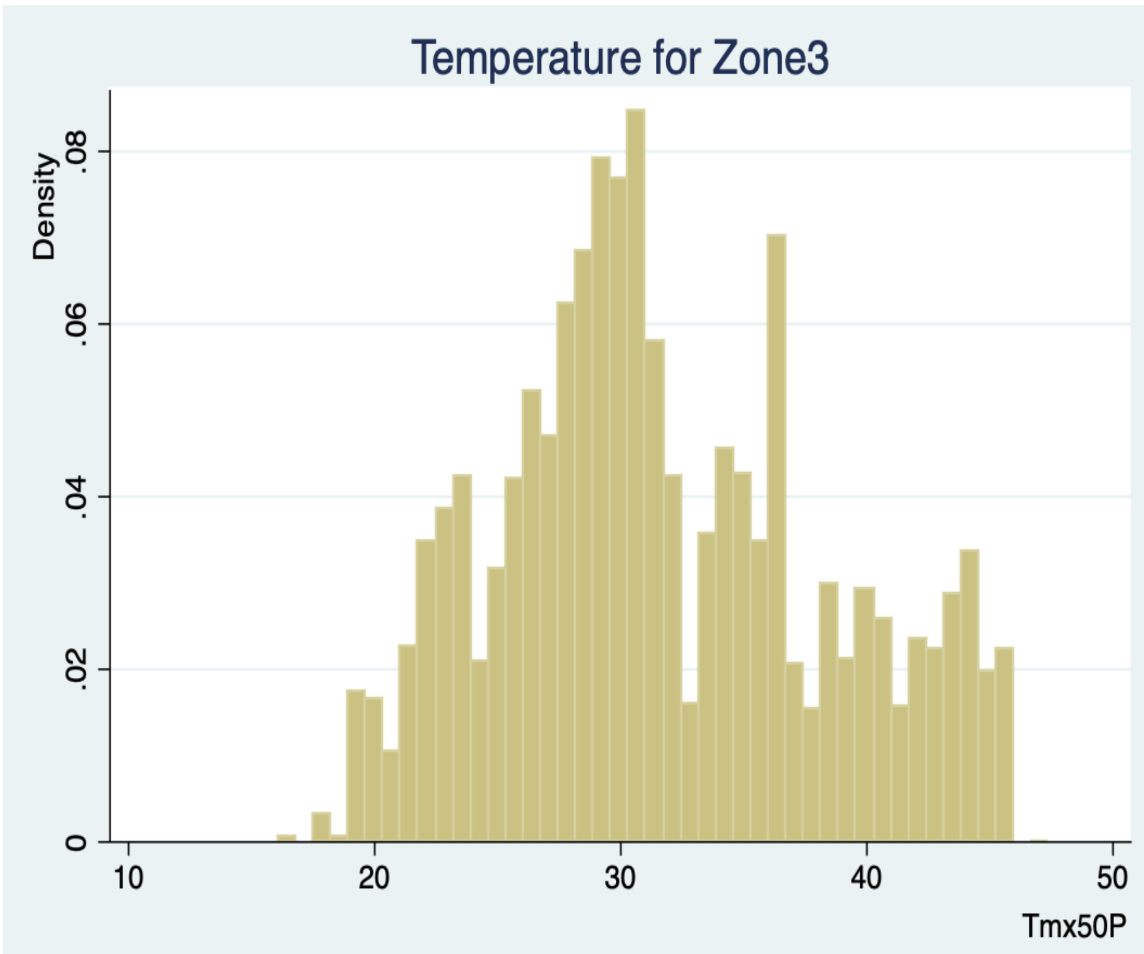

Figure 1.4: Variation in treatment for zone 3.

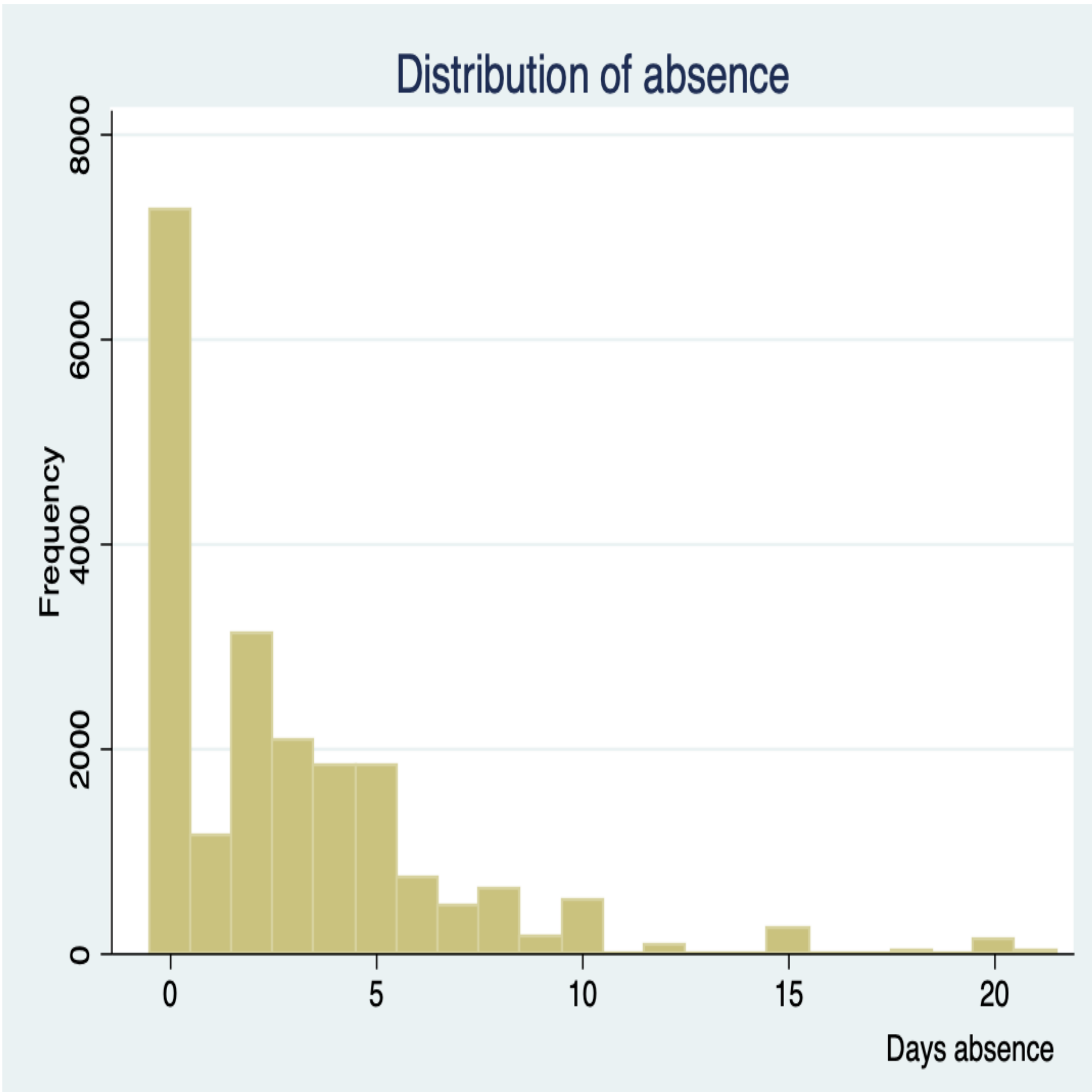

Figure 1.5: Absence distribution with the sample restrictions.

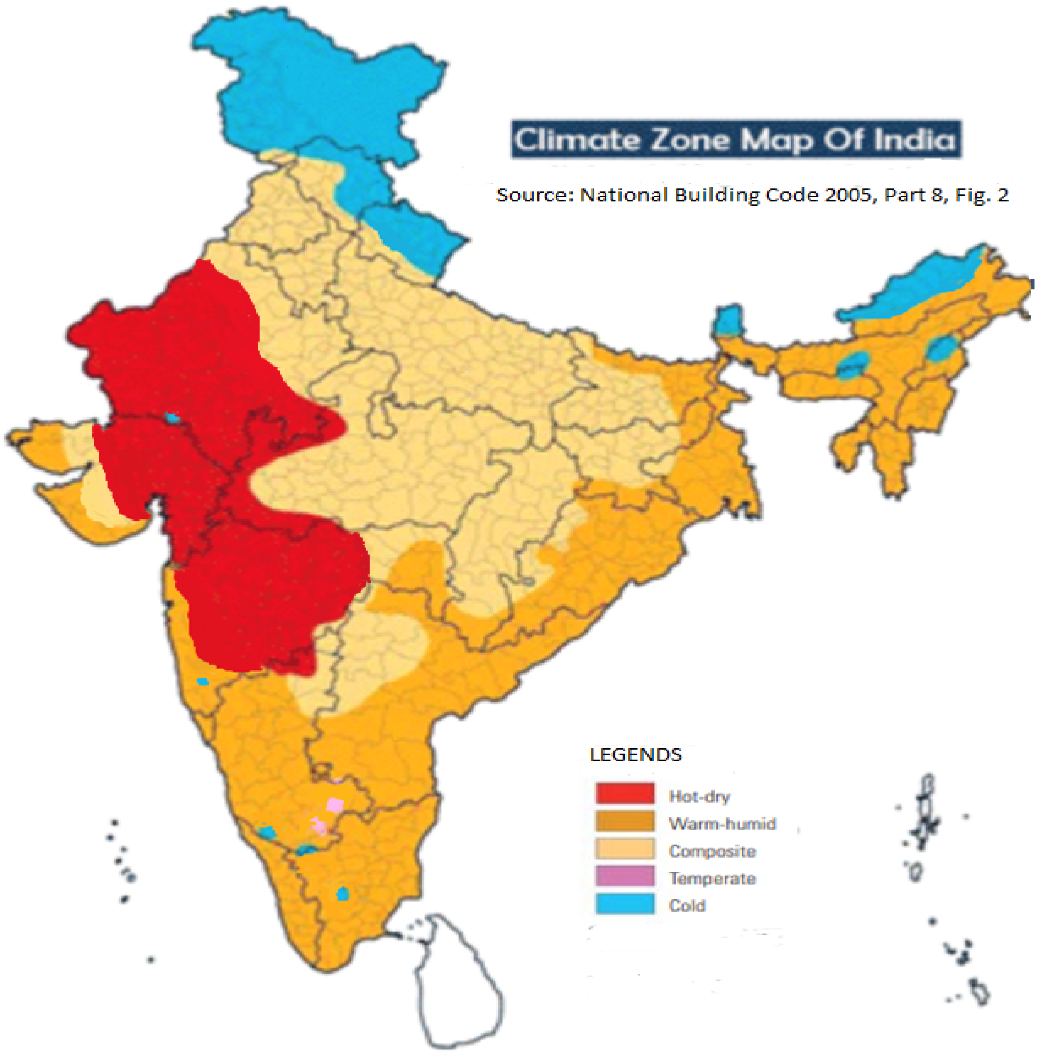

Figure 1.6: This figure shows climate zones.

Table 1.5: The effect of maximum temperatures at the 50th percentile on student absence in rural areas

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Preferred (8)	Zero_inflated (9)
DysAbs									
Temp	-0.0134	0.107***	0.109***	0.108***	0.0950***	0.0546*	0.0546*	0.0965***	0.0834***
	(0.00765)	(0.0317)	(0.0316)	(0.0307)	(0.0318)	(0.0307)	(0.0307)	(0.0309)	(0.0248)
Zone1*Temp		-0.106***	-0.108***	-0.105**	-0.0956***	-0.0576*	-0.0576*	-0.0968***	-0.0804***
		(0.0346)	(0.0343)	(0.0336)	(0.0345)	(0.0324)	(0.0324)	(0.0335)	(0.0259)
Zone3*Temp		-0.112***	-0.114***	-0.113***	-0.106***	-0.0656**	-0.0656***	-0.109***	-0.0897***
		(0.0331)	(0.0330)	(0.0321)	(0.0325)	(0.0319)	(0.0319)	(0.0317)	(0.0255)
Weather controls					√	√	√	√	√
Individual controls			√			√	√	√	√
Household controls				√		√	√	√	√
State FE		√	√	√	√	√	√	√	√
Zone*Month FE		√	√	√	√			√	√
Month FE						√	√		
Zone FE						√			
Observations	20080	20080	20080	20022	20080	20022	20022	20022	20022

Notes: Standard errors in parentheses clustered at the district level. * p<0.1, ** p<0.05, *** p<0.01.
Temp represents maximum temperatures at the 50th percentile of the maximum daily temperatures for the thirty day period prior to the individual interview date. Zone represents the climate zone, which includes many districts that have common climate features. Zone*Temp represents interaction terms of temperature with climate zones. Individual controls include age, gender and grade inclusive of grades 1 to 12. Household controls include the highest level of education for the household male and female, household income source type, distance from the household to the school, the household's monthly income, a dummy variable represents if the household has air conditioning. Weather controls include the rates of precipitations and the average relative humidity for the the thirty day period. Column 9 shows that using the zero-inflated negative binomial regression does not change the significance level of the estimate, and magnitude is slightly smaller than the estimate of the preferred specification in column 8.

Table 1.6: The effect of maximum temperature at the 50th percentile on students' absences in urban areas

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Preferred column (8)
DysAbs								
Temp	-0.0128	0.0235	0.0235	0.0201	-0.0277	-0.0277	-0.0205	-0.0242
	(0.0108)	(0.0217)	(0.0210)	(0.0193)	(0.0250)	(0.0250)	(0.0309)	(0.0298)
Zone1*Temp		0.0181	0.0204	0.0188	0.0556*	0.0556**	0.0460	0.0496
		(0.0257)	(0.0254)	(0.0245)	(0.0252)	(0.0252)	(0.0290)	(0.0287)
Zone3*Temp		-0.0362	-0.0360	-0.0359	0.00801	0.00801	-0.0197	-0.0188
		(0.0274)	(0.0269)	(0.0257)	(0.0232)	(0.0232)	(0.0298)	(0.0286)
Weather controls					√	√	√	√
Individual controls			√		√	√		√
Household controls				√	√	√		√
State FE		√	√	√	√	√	√	√
Zone*Month FE		√	√	√			√	√
Month FE					√	√		
Zone FE					√			
Observations	9109	9109	9109	9084	9084	9084	9109	9084

Notes: Standard errors in parentheses clustered at the district level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Temp represents maximum temperature at the 50th percentile of the maximum daily temperatures for the thirty day period prior to the individual interview date. Zone represents the climate zone, which includes many districts that have common climate features. Zone*Temp represents interaction terms of temperature with climate zones. Individual controls include age, gender and grade which includes grades 1 to 12. Household controls include the highest level of education for the household male and female, household income source type, distance from household to school, the household monthly income, a dummy variable represents if the household has air cooling. Weather controls include the total amount of rain and the average relative humidity for the thirty day period.

Table 1.7: Subsamples from the final sample in rural areas

	Excluding seasonal migrants	Excluding problems with normal activity	Excluding problems with breath, fever, cough,diarrhea	Preferred
	(1)	(2)	(3)	(4)
DysAbs				
Temp	0.104**	0.102**	0.0902***	0.0955***
	(0.0316)	(0.0322)	(0.03077)	(0.0309)
Zone1*Temp	-0.104**	-0.0953**	-0.0907	-0.0968***
	(0.0342)	(0.0355)	(0.03348)	(0.0335)
Zone3*Temp	-0.115***	-0.115***	-0.102	-0.109***
	(0.0328)	(0.0332)	(0.03166)	(0.0317)
Weather controls	√	√	√	√
Individual controls	√	√	√	√
Household controls	√	√	√	√
TimeLocation FE	√	√	√	√
Observations	17950	17171	19939	20022

Notes: Standard errors in parentheses clustered at the district level. * p<0.1, ** p<0.05, *** p<0.01. Regressions in all columns are implemented on rural areas. These regressions include the period from March till November, students that had days of absenteeism ≤ 21, as well as the same household and individual controls that are used in the final sample. Temp represents maximum temperatures at the 50th percentile of the maximum daily temperatures for the thirty day period prior to individual interview date. Zone*Temp represents the interaction term of climate zones with temperature. Zone represents the climate zone, which includes many districts that have common climate features. Individual controls include age, gender and grade that includes grades 1 to 12. Household controls include the highest level of education for household males and females, household income source type, distance from household to school, the household monthly income, and a dummy variable represents if the household has air conditioning. Weather controls include the total amount of rain for the thirty day period, and the average relative humidity for the same period. TimeLocation fixed effect include climate zones interaction with months and state fixed effects. The regression in column 1 excludes students who are seasonal migrants due to extreme conditions such as drought or flood etc. Column 2 excludes any students that have been unable to undertake normal activities in the past thirty days. Column 3 excludes any students who have experienced breathing problems, fever, diarrhea, or cough in past thirty days. Column 4 does not excludes any students who reported any one of these conditions/symptoms mentioned in column 1 to 3.

Table 1.8: Heterogeneity by income level, income source type, and gender

	Income level		Income source		Gender		All
	Poor	Non poor	Agricultural source	Other sources	Male	Female	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DysAbs							
Temp	0.0812**	0.0995***	0.0928***	0.0895***	0.0957***	0.0953***	0.0955***
	(0.0364)	(0.0331)	(0.0314)	(0.0342)	(0.0329)	(0.0318)	(0.0309)
Zone1*Temp	-0.0581	-0.103***	-0.0706**	-0.104***	-0.0972**	-0.0892***	-0.0968***
	(0.0386)	(0.0358)	(0.0337)	(0.0394)	(0.0345)	(0.0356)	(0.0335)
Zone3*Temp	-0.111***	-0.106***	-0.119***	-0.0841***	-0.110***	-0.106***	-0.109***
	(0.0377)	(0.0343)	(0.0333)	(0.0346)	(0.0337)	(0.0325)	(0.0317)
Weather controls	√	√	√	√	√	√	√
Individual controls	√	√	√	√	√	√	√
Household controls	√	√	√	√	√	√	√
TimeLocation FE	√	√	√	√	√	√	√
Observations	5283	14739	10469	9553	10467	9555	20022

Notes: Standard errors in parentheses clustered at the district level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.
Temp represents maximum temperatures at the 50th percentile of the maximum daily temperatures for the thirty day period prior to the individual interview date. Zone represents the climate zone, which includes many districts that have common climate features. Zone*Temp represents interaction term of temperature with climate zones. Individual controls include age, gender and grade which includes grades 1 to 12. In columns 5 and 6, individual's controls do not include gender. Household controls include the highest level of education for households males and females, household income source type, distance from household to school, the household monthly income, a dummy variable represents if the household has air conditioning. In columns 1 and 2, household's controls do not include household monthly income. Weather controls include the total precipitation rates and the average relative humidity for the thirty day period. TimeLocation fixed effect include climate zones interaction with months and state fixed effect. All these regressions are implemented on rural areas.

Table 1.9: Electricity, air cooling, adequate water and drinking water supply source

	All	Electricity		Air cooling		Adequate water		Piped water supply	
		Yes	No	Yes	No	Yes	No	Yes	No
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DysAbs									
Temp	0.0955***	0.111***	0.0778	0.140***	0.0917***	0.105***	0.0609	0.0848***	0.0765**
	(0.0309)	(0.0302)	(0.0451)	(0.0408)	(0.0329)	(0.0346)	(0.0373)	(0.0306)	(0.0378)
Zone1*Temp	-0.0968	-0.109***	-0.0617	-0.146***	-0.0908**	-0.106***	-0.0229	-0.0559	-0.0622
	(0.0335)	(0.0327)	(0.0637)	(0.0435)	(0.0356)	(0.0369)	(0.0477)	(0.0310)	(0.0386)
Zone3*Temp	-0.109***	-0.115***	-0.119*	-0.138***	-0.108***	-0.116***	-0.0653	-0.0416	-0.0970**
	(0.0317)	(0.0320)	(0.0487)	(0.0493)	(0.0336)	(0.0349)	(0.0568)	(0.0301)	(0.0396)
Weather controls	√	√	√	√	√	√	√	√	√
Individual controls	√	√	√	√	√	√	√	√	√
Household controls	√	√	√	√	√	√	√	√	√
TimeLocation FE	√	√	√	√	√	√	√	√	√
Observations	20022	17077	2945	2361	17661	18563	1459	7639	12383

Notes: Standard errors in parentheses clustered at the district level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Temp represents maximum temperatures at the 50th percentile of the maximum daily temperatures for the thirty day period prior to the individual interview date. Zone represents the climate zone, which includes many districts that have common climate features. Zone*Temp represent interaction term of temperature with climate zones. Individual controls include age, gender and class standard or grade which include grades 1 to 12. Household controls include the highest level of education for household males and females, household income source type, distance from household to school, the household monthly income, a dummy variable represents if the household has air conditioning. In columns 3, 4 and 5 the household controls do not include the air cooling dummy variable. Weather controls include the total amount of rain and the average relative humidity for the thirty day period. TimeLocation fixed effect includes climate zones interaction with months and state fixed effect. All these regression are implemented on rural areas.

Table 1.10: Heterogeneity by school level. Elementary level versus secondary level

	Elementary level			Secondary level			The two levels together		
	Air cooling	No AC	All	Air cooling	No AC	All	Air cooling	No AC	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DysAbs									
Temp	0.120***	0.0776**	0.0779**	0.233***	0.136***	0.148***	0.140***	0.0917***	0.0941***
	(0.0441)	(0.0326)	(0.0308)	(0.0496)	(0.0406)	(0.0381)	(0.0414)	(0.0328)	(0.0308)
Zone1*Temp	-0.153***	-0.0591	-0.0666**	-0.201***	-0.148***	-0.157***	-0.143***	-0.0906**	-0.0943***
	(0.0458)	(0.0353)	(0.0332)	(0.0584)	(0.0435)	(0.0409)	(0.0440)	(0.0356)	(0.0334)
Zone3*Temp	-0.108**	-0.0933***	-0.0898***	-0.219***	-0.152***	-0.160***	-0.132***	-0.107***	-0.106***
	(0.0528)	(0.0333)	(0.0315)	(0.0505)	(0.0419)	(0.0392)	(0.0496)	(0.0336)	(0.0316)
Weather controls	√	√	√	√	√	√	√	√	√
Individual controls	√	√	√		√	√	√	√	√
Household controls	√	√	√		√	√	√	√	√
Zone*Month control	√	√	√		√	√	√	√	√
State FE	√	√	√	√	√	√	√	√	√
Month FE				√					
Observations	1585	12937	14522	776	4724	5500	2361	17661	20022

Notes: Standard errors in parentheses clustered at the district level. * p<0.1, ** p<0.05, *** p<0.01.
The education system in India is divided into elementary level and secondary level. Elementary level includes grades one to eight. Secondary level includes grades nine to twelve. Each regression in this table includes individual controls, household controls and state FE. Temp represents maximum temperatures at the 50th percentile of the maximum temperatures for the thirty day period prior to the individual interview date. Zone represents the climate zone, which includes many districts that have common climate features. Zone*Temp represents interaction term of temperature with climate zones. Individual controls include age, gender and grade which includes grades 1 to 12. Household controls include the highest level of education for household males and females, household income source type, distance from home to school, the household monthly income. In columns 3, 6 and 9, household controls do not include air cooling dummy variable. AC is abbreviation for air cooling. Weather controls include the precipitation rates for the thirty days period, and the average relative humidity for the same period. All these columns represent rural areas.

Table 1.11: Heterogeneity by geography

	Climate Zones			The preferred specification
	CZ1	CZ2	CZ3	
	Cold	Dry	Humid	
	(1)	(2)	(3)	(4)
DysAbs				
Temp	0.00811	0.110***	-0.0111	0.0965**
	(0.0105)	(0.0323)	(0.0147)	(0.0309)
Zone1*Temp				-0.0968***
				(0.03355)
Zone3*Temp				-0.109***
				(0.0317)
RainSum	√		√	√
RelHumAvg			√	√
Individual controls	√	√	√	√
Household controls	√	√	√	√
State FE	√	√	√	√
Month FE	√	√	√	
Zone*Month FE				√
Observations	2122	2998	14902	20022

Notes: Standard errors in parentheses clustered at the district level. * p<0.1, ** p<0.05, *** p<0.01.

Each column from 1 to 3 represents a separate climate zone (CZ1, CZ2, CZ3). Temp represents maximum temperature at the 50th percentile of the daily maximum temperatures for the thirty day period prior to the individual interview date. RainSum represents total amount of rain for the same period. RelHumAvg represents the average relative humidity for the same period. Individual controls include age, gender, grade. Household controls include the highest level of education for household males and females, household income source type, distance from household to school, the household monthly income, a dummy variable represents if the household has air cooling. Zone*Month is interaction between climate zones and months. All these columns are implemented on rural areas.

Table 1.12: Robustness

	Preferred	Moving Average	DayTemp+38^0C	Days absence ≤ 30	Days absence ≤ 10	Whole year	Winzorized	Clustering at different levels		
								District*month	District*season	State
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
DysAbs										
Tmax50P	0.0965***			0.0721***	0.0992***	0.0668***	0.0763***	0.0965***	0.0965***	0.0965***
	(0.0309)			(0.0248)	(0.0305)	(0.0207)	(0.0256)	(0.0260)	(0.0265)	(0.0306)
Avge max-temp		0.105***								
		(0.0288)								
DayTemp+38^0C			0.0331***							
			(0.0108)							
Zone1*Temp	-0.0968***	-0.104***	-0.00750	-0.0471	-0.104***	-0.0658***	-0.0549*	-0.0968***	-0.0968***	-0.0968***
	(0.0335)	(0.0320)	(0.0145)	(0.0277)	(0.0334)	(0.0249)	(0.0284)	(0.0288)	(0.0294)	(0.0314)
Zone3*Temp	-0.109***	-0.113***	-0.0341***	-0.0738***	-0.109***	-0.0557***	-0.0817***	-0.109***	-0.109***	-0.109***
	(0.0317)	(0.0309)	(0.0118)	(0.0268)	(0.0318)	(0.0212)	(0.0272)	(0.0261)	(0.0271)	(0.0326)
Weather controls	√	√	√	√	√	√	√	√	√	√
Individual controls	√	√	√	√	√	√	√	√	√	√
Household controls	√	√	√	√	√	√	√	√	√	√
TimeLocation FE	√	√	√	√	√	√	√	√	√	√
Observations	20022	20022	20022	20481	19494	26143	20481	20022	20022	20022

Notes: Standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

All columns are implemented on rural areas. All columns include the period from March till November, except column 6 which includes the whole year. Column 1, 2 and 3 are similar regarding controls and restrictions, and the only difference is the treatment metric. Column 1 uses the Tmax50P which represents maximum temperatures at the 50th percentile of the daily maximum temperatures for the thirty day period prior to the interview date, while column 2 uses the thirty day moving average of daily maximum temperatures for the same period. Column 3 uses two bins of the number of days with temperature greater than 38^0C compared to the days below. The three columns use the same restriction on the outcome variable, which is students with days absences ≤ 21. Column 4 contains students with days absences ≤ 30. Column 5 contains students with days absences ≤ 10. Columns 1, 2, 3, 4, 5, 6 and 7 are clustered at the district level. In column 7, any student's absence value greater than 21 days is replaced by 21. Zone*Temp represents the interaction term of climate zones with temperature. Zone represents the climate zone, which includes many districts that have common climate features. Individual controls include age, gender and grade which includes grades 1 to 12. Household controls include the highest level of education for household males and females, household income source type, distance from household to school, the household monthly income, and a dummy variable represents if the household has air conditioning. Weather controls include the total amount of rain and the average relative humidity for the thirty day period. TimeLocation fixed effect include climate zones interaction with months and state fixed effect. All regressions are implemented on rural areas.

Table 1.13: Different maximum temperature percentiles in rural areas

	DyaAbs (1)	DyaAbs (2)	DyaAbs (3)	DyaAbs (4)	DyaAbs (5)	DyaAbs (6)	DyaAbs (7)	DyaAbs (8)	DyaAbs (9)
DysAbs									
Tmax10P	0.057***								
	(0.02103)								
Tmax20P		0.082***							
		(0.02530)							
Tmax30P			0.095***						
			(0.02898)						
Tmax40P				0.096***					
				(0.03023)					
Tmax50P					**0.096*** **				
					(0.03090)				
Tmax60P						0.097***			
						(0.03147)			
Tmax70P							0.098***		
							(0.03199)		
Tmax80P								0.092***	
								(0.03095)	
Tmax90P									0.099***
									(0.02877)
Zone1*Tmx#P	-0.0607***	-0.0852**	-0.0982**	-0.0973***	-0.0968***	-0.0986***	-0.0999***	-0.0918***	-0.0907***
	(0.0249)	(0.0287)	(0.0318)	(0.0330)	(0.0335)	(0.0340)	(0.0345)	(0.0336)	(0.0320)
Zone3*Tmx#P	-0.0790***	-0.104***	-0.114***	-0.111***	-0.109***	-0.110***	-0.109***	-0.103***	-0.0960***
	(0.0230)	(0.0266)	(0.0298)	(0.0309)	(0.0317)	(0.0322)	(0.0328)	(0.0319)	(0.0302)
Weather controls	√	√	√	√	√	√	√	√	√
Individual controls	√	√	√	√	√	√	√	√	√
Household controls	√	√	√	√	√	√	√	√	√
TimeLocation FE	√	√	√	√	√	√	√	√	√
Observations	20022	20022	20022	20022	20022	20022	20022	20022	20022

Notes: Standard errors in parentheses clustered at the district level. * p<0.1, ** p<0.05, *** p<0.01.
CZ*Tmx#P in each column represents the interaction term of the maximum temperature at specific percentile value with climate zones.

Table 1.14: Placebo test

	Preferred Main analysis (1)	Placebo falsely-assigned temperature (2)
DysAbs		
Temp	0.0965***	-0.022
	(0.0309)	(0.0206)
Zone1*Temp	-0.097***	0.040
	(0.0335)	(0.0286)
Zone3*Temp	-0.109***	0.0223
	(0.0317)	(0.02097)
Weather controls	√	√
Individual controls	√	√
Household controls	√	√
TimeLocation FE	√	√
Observations	20022	20022

Notes: Standard errors in parentheses clustered at the district level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Temp represents maximum temperatures at the 50th percentile of the maximum temperatures distribution. Zone represents the climate zone, which includes many districts that have common climate features. Zone*Temp represents interaction term of temperature with climate zones. Individual controls include age, gender and grade which includes grades 1 to 12. Household controls include the highest level of education for household males and females, household income source type, distance from household to school, the household monthly income, a dummy variable represents if the household has air conditioning. Weather controls include the total amount of rainfall recorded for the past month, and the average relative humidity for the past month. TimeLocation fixed effect include climate zones interaction with months and state fixed effect.

Chapter 2

Heatwaves and Short-Term Morbidity in India

2.1 Introduction

We investigate the short-term health related impacts of heatwaves on respondents in a large representative sample of Indians using an innovative methodology.

Exposure to heat for prolonged periods endangers human health to the point that it may even result in death. According to the CNN report released by Hilary Whiteman on June 2, in 2015 only, heatwaves led to the death of 2,300 people in India.[1] The adverse impact of heat has been widely documented within mainstream media but is also part of contemporary peer reviewed research. One such example includes the work of Semenza et al. (1996) who identified that the July 1995 Chicago heatwave caused at least

[1]The CNN report by Hilary Whiteman on June 2, 2015. `https://edition.cnn.com/2015/06/01/asia/india-heat-wave-deaths/index.html`

700 excess deaths.

Exceptionally hot weather lasting for a period of time is defined as a heatwave. Such weather events are relatively rare but do generate significant economic losses as well as damages to human health and the environment. Heatwaves can destroy vegetation and lead to drought. They also have a salient adverse impact on animals and humans' health. As described by Luber and McGeehin (2008), heatwaves are responsible for more weather-related deaths than all other natural disasters such as lightning, tornadoes, floods, and earthquakes combined.

Contemporary research in health science, environmental studies and other fields have discussed the adverse impact of heatwaves. As such, current findings highlight the link between prolonged high temperatures and reduced health outcomes. This paper contributes to existing research by analyzing the short-term morbidity (fever, cough, and diarrhea) impact of heatwaves.

There is currently no worldwide specific definition of a heatwave. However, scientists have recently started implementing different scenarios of specifications and linking these with different variables of interest. It may be argued that clearly setting parameters to define a heatwave could lead to the implementation of informed policies that would reduce adverse impacts on vulnerable populations. It can also be put forward that such policies would reduce global economic losses.

Studies focusing on the impact of heatwaves generally utilize hospital admissions and emergency department visits as the outcome variable. It must be noted that the hospital admissions and the emergency visits are due to either mental and behavioral disorders or physical conditions such as cerebrovascular, cardiovascular, myocardial infarction, respiratory disease and death (Knowlton et al., 2009; Åström et al., 2011; Mastrangelo et al., 2007; Bhaskaran et al., 2010; A. L. Hansen et al., 2008; A. Hansen et al., 2008; D'Ippoliti et al., 2010). To define a heatwave, various studies simply use information and data record on heatwaves as it is given from the meteorologial department and weather forecast. Alternatively, some researchers use a given value such as 30^0C or 40^0C as a threshold of temperature over a set number of consecutive days to identify a heatwave. The percentile-based approach is more flexible and considers geographical features' diversity.

This paper contributes to the literature in two ways (i) Firstly, most of the current literature related to our topic utilizes a small sample whereas we conduct our analysis using a large representative sample of Indian residents who reported their days of short-term morbidity. (ii) We use an innovative way to measure the impacts of heatwaves based upon heat intensity and duration.

In order to collect and analyse data, we use self-reported information from a large scale survey about an individual's total number of short-term morbidity (fever, cough, and diarrhea) days for the thirty days prior to the interview as the output variable. We link it with three metrics measuring heatwaves that use a percentile-based approach. Our identifying assumption is that by controlling for variation in interview dates within months and location, realized temperature patterns, including heatwaves, are as good as randomly assigned.

India is a subcontinent that experiences high temperatures and where poverty remains a major challenge. When analyzing the impact of high temperatures using the three heat metrics, we find that the occurrence of heatwaves in this already impoverished country leads to a rise in the short-term morbidity days. Such results are consistent with existing research confirming that exposure to prolonged heat increases death rates. Although research focusing on heatwaves and death rates is essential, studying the less severe effects of heat exposure on health should not be overlooked. As such, researching these provides greater insights as to how various preventative care response strategies can be implemented to lower short-term morbidity. Such findings would also contribute to ensure that health facilities and social infrastructures are ready to tackle significant increases of individuals requiring care due to heatwaves. Finally, it provides governments further evidence around the long-term benefits of implementing innovative technologies and/or offering more affordable power supply.

According to Phelps (1980) and Barro (2000), human capital is an important economic growth factor. Therefore, by adequately protecting the population from the adverse effects of heat, economic losses are reduced as labor and production is maintained. Indeed, when an individual is unwell, the productivity of the entire household is reduced as other members may elect to take care of the person who is sick, escort them to a health care facility, etc. Therefore, human capital is automatically reduced in turn affecting industries and generating an economic loss. This paper does not focus on this particular topic, but we intend to explore this aspect within future research.

The rest of the paper is organized as follows. Section 2 is a literature review. Section 3 presents the data that we use in this study. In section 4, we outline the empirical strategy. Section 5 shows the results. Section 6 presents the robustness check. Section 7 provides a discussion. Section 8 concludes.

2.2 Literature Review

The impact of extreme heat on human morbidity and mortality is of great concern and is the focus of several recent studies around the world (Rey et al., 2009; Huang et al., 2010; Bell et al., 2008; Anderson and Bell, 2009). Studies focusing on the effects of high temperatures find that heat has a heterogeneous impact. When hot weather temperature persists for a period of time, the human body works harder to regulate its internal temperature.

Heatwaves affect both underdeveloped and developed countries. They are natural disasters that can have devastating impacts on populations. For example, the July 1995 heatwave in Chicago caused at least 700 excess deaths (Semenza et al., 1996). Despite technological advances, the adverse impact of heatwaves on human health remains a serious threat that is still not well-managed globally, even in developed countries. For instance, during its 2006 heatwave, California saw a sharp increase in department visits and hospitalizations of people aged 65 years and above (Knowlton et al., 2009). Furthermore, D'Ippoliti et al. (2010) studied the impact of the heatwave in nine European cities. Their findings indicate that heatwave is associated with an increase in mortality rate in Munich and Milan by 7% and 33.6%, respectively.

In acknowledging the impact of heat in developed countries, it must be noted that emerging nations are still the ones recording the most severe adverse effects of high temperatures. As such, it may be argued that impoverished nations that have less access to technology, higher poverty rates, and inadequate infrastructures are in a more vulnerable position. For example, as cited by Knowlton et al. (2014), the 2010 heatwaves in Ahmedabad, India killed over 1300 people. As taking part in this research field, we extend such literature by also studying short-term morbidity (fever, cough, and diarrhea) based on individuals' self-reported the number of days as being unwell in a large representative sample of Indian residents. As such, it can be argued that using a subjective assessment may impact the accuracy of the answers provided. However, in our case, the dependent variable is individual self-report of short-term morbidity days over the thirty day period prior to the interview. Therefore, we assume that individuals accurately report such information as we focus on a short period of recent history.

Whether hot days are scattered or consecutive, researchers mostly utilize standard temperature exposure measures to explore the impacts of days of extreme heat. In other words, previous studies make the assumption that temperature effects are additively separable across days (Barreca, 2012; Dell et al., 2012; Deschênes and Greenstone, 2011; Deschênes and Moretti, 2007; Graff Zivin and Neidell, 2014). However, Chua, Coggins, Miller, and Mohtadi (2019) who studied the impact of heatwaves on economic output attempted to overcome this deficit.

Within our research, we implement three different metrics for measuring the effect of the heatwaves. Although most peer reviewed papers discuss prolonged heat exposure using a similar terminology "heatwave", there is currently no formally agreed definition of a heatwave (Hajat et al., 2006; Zhang et al., 2018; Tolika, 2019; Chua et al., 2019; Peng Bi et al., 2011; Toloo et al., 2013; Lowe et al., 2011; Park and Kim, 2018). In this paper, we implement a percentile-based approach to define what constitutes a heatwave. As such, we consider accumulated heat as a heatwave when the maximum daily temperature for three consecutive days or more for specific region exceed the temperature at the 85th percentile of the hottest month in a year. As such this definition will be used throughout the main analysis of this paper. Our thresholds are based upon the work of (Mazdiyasni and Agha Kouchak, 2015; Mazdiyasni et al., 2017), where they consider heatwaves to be three or more consecutive days of temperature above the 85th percentile of the hottest month for each specific location. Our approach is still recent in measuring the accumulated heat exposure and new to the economics literature.

The proposed heatwave definition has two advantages over the heat impact literature. First, it accounts for connected heat rather than additive separable impacts of a scattered heat measure. Furthermore, using a relative threshold (percentile-based) makes it has an advantage over the connected heat measures that use absolute thresholds, where the former considers the diversity of the geographical features through accounting for the wide variation in meteorological factors. The percentile-based approach embeds an arbitrary thresholds suit each specific location. For example, if a region is known as cold, while other region known as hot, the actual value of the 85th percentile for each one of these two regions are different. In other words, it gives flexibility in defining heatwave based on each region geographical features rather than restricting all regions to absolute threshold such as 40^0C.

2.3 Data

2.3.1 Indian survey

The India Human Development Survey (IHDS-II) conducted in 2011-12 is a large-scale cross-section survey. It covers all individuals residing in 42,152 households spread across 1,420 villages and 1,042 urban neighbourhoods in India. This nationally representative survey is an output of the collaboration between researchers from the National Council of Applied Economic Research, New Delhi (NCAER) and the University of Maryland.

From this dataset, we obtain the health outcome variable. This dependent variable represents an individual's self-report of the number of days he/or she report being ill (fever, cough, and diarrhea) in the thirty day period prior to the interview date. By looking at Figure 2.1, we see a right-skewed distribution of our outcome variable, which makes a simple count data model the most effective model to be utilized. Furthermore, the dataset provides several independent variables and controls. They include an individual's age in years. In addition, a dummy variable represents an individual's gender; a dummy variable represents persevering on hand wash after defecation, continuous variable represents the number of completed years of education, and a dummy variable represents marital status.

Furthermore, we obtain household characteristics from this data, such as a dummy variable represents whether a household has electricity or not. Household size represents the number of individuals residing in a household. A dummy variable represents the household's principal source of income, and a continuous variable represents the average monthly income level. A dummy variable represents whether a household share a toilet with people from outside the family. A continuous variable represents the distance in minutes from home to a source of water. By looking at Table 2.1, we see that our sample splits into almost 50% females and 50% males. Our data show that 95% of our sample does not share a toilet with people outside the household. Approximately, 96% of the sample reported adequate hand hygiene practices after toileting. Such preventative health measures are well known for curbing spreading diseases.

Since IHDS-II includes the interview date and location for each household, it enables us to link the weather information with this big survey to obtain our final sample.

2.3.2 Meteorological data

In this section, we construct the weather variables and identify the three metrics for measuring heatwaves to link them with the IHDS-II dataset. The most suitable weather data to be used is the observational weather data from ground stations (Garg et al., 2017). However, India does not have enough weather stations on the ground that cover weather variations in all areas. Therefore, we follow Schlenker and Roberts (2009), Schlenker and Lobell (2010), and Auffhammer et al. (2013) using reanalysis data to identify the best estimated weather data.

From ERA-Interim archive of grid cells with 1^0 X 1^0 latitude-longitude,[2]

[2] ERA-Interim archive is constructed by European Center for Medium-Range Weather Forecasts (ECMWF). `https://www.ecmwf.int/en/forecasts`.

we obtain raw values of weather variables inclusive of maximum daily temperatures, minimum daily temperatures, daily average wind speed, daily average relative humidity, and daily precipitation level. To generate the constructed values of weather variables for each district, we use all the grid points that lie within 50 km of the centroid of each district and calculate the inverse distance weighted average for all weather variables (temperature, humidity, wind speed, and precipitation). Then we assign the values of these constructed variables to districts. To address the variation of interview dates within months, we assign to individuals the constructed maximum daily temperature, minimum daily temperature, daily relative humidity, daily wind speed, and daily precipitation based on their interview dates and the district they live in.

Finally, based on the percentile-based approach and using the constructed weather data assigned to individuals, we utilize three different developed metrics to measure the impact of heat waves during the thirty day period prior to the interview date. The first metric is the number of heat events that occurred in the last thirty days. The second metric represents the total number of hot days during the heatwave event. The third metric represents the accumulated cooling degree days of the heatwave periods.[3] Figure 2.2 shows the distribution of the number of heat events lasting three or more consecutive days. Figure 2.3 depicts the number of hot days' distribution. Figure 2.4 shows the accumulated cooling degree days for the same time period. We utilize these three metrics to evaluate the heatwave events because each one alone might be able to provide a full picture of the heatwave impact. The frequency

[3]We used 22^0C as the threshold to calculate the accumulated cooling degree days (ACDD). To calculate the ACDD, suppose we have the maximum daily temperature for three days 30^0C, 23^0C, and 21^0C. ((30-22)+(23-22)+ 0)= 9^0C. Therefore, the ACDD is 9. If the temperature is less than 22, we consider the point to be of zero value since we assume the day does not need to be cooled.

of heatwaves might be the same for two different regions in a year, but the duration of each one of these heatwaves might be different. For example, if we have two regions reported one heatwave each, one heatwave lasts for three days, while the other lasts for six days, this implies that the same frequency might lead to different impacts. The intensity of heatwave is a crucial metric too. For example, if we have two regions reported the same frequency and same duration of days, but the first location reported a temperature of 45^0C on each day, while the other location reported the temperature of 35^0C, then the intensity plays a significant role in determining different impact. Thus, we utilize the three metrics to capture the heatwave impact on short-term morbidity.

2.4 Methodology

We link IHDS-II to the constructed weather data, and we obtain the final cross-sectional representative sample. Our dependent variable is the total number of days that an individual reported experiencing short-term morbidity during the thirty day period prior to the interview. This variable has a skewed distribution. Using a statistical approach that does not suit the data distribution may result in inaccurate estimates of standard errors, misleading P-values, and producing estimates of dependent variables that are accurate for a subsample but inaccurate for others. For example, using the ordinary least squares model is inappropriate in this instance as our count data violates the assumptions of normality and constant variance (Du et al., 2012). Even if we attempt to modify our dependent variable and turn it into a more normal distribution by utilizing logarithm, the zero values are not taken into account and the normality of the estimated values is not achieved, and this leads to generate less accurate point estimates (Cohen et al., 2013; Manning et al., 1999; Mosteller and Tukey, 1977).

Therefore, it is more appropriate to analyze our data using a count data model. Since our outcome variable is over-dispersed, the magnitude of the conditional variance exceeds the conditional mean's magnitude,[4] preference is given to the negative binomial regression to the standard poisson model to identify the relation between heat and short-term morbidity. Our work focuses on studying the impact of prolonged heat (heatwaves) exposure on individuals' short-term morbidity (diarrhea, fever, and cough) for the thirty day period prior to the interview date. To do so, we run the negative binomial model three times and adopt one of the three different heatwave metrics each.

Our strategy allow us to estimate a simple count data model of the type

$$Y_{idmt} = exp(\alpha_0 + \alpha_1 Heat_{dmt} + \alpha_2 W_{dmt} + \alpha_3 X_i + \zeta_d + \psi_m + \eta_t + \varepsilon_{idmt}) \tag{1}$$

Where Y_{idmt} represents the dependent variable, which is an individual i from district d interviewed on date t in month m reported the number of days being sick in the thirty days prior to the interview. The coefficient of interest, α_1, on its own does not reflect the total impact. Therefore, we will consider the other regressors in the model as well to report the total effect. It will provide us with the expected change in expected amount of short-term morbidity days

[4]In the negative binomial distribution, the variance represented as a function of its mean, μ, and an additional parameter, k, represents the dispersion. The relation can be depicted as $var(Y) = \mu + \mu^2/k$

due to an increase in the number of heatwaves. $Heat_{dmt}$ represents the explanatory variable of interest, the first heatwave metric, which is the number of heatwaves during the thirty days period prior to the individual interview dates. W_{dmt} represents the constructed weather data for the same thirty day period that includes the total amount of precipitation and the thirty day average of daily relative humidity and daily wind speed. X_i represents several individual's and household's confounding factors that include individual's gender, age, marital status, and a dummy variable represents if adequate hand hygiene practices after toileting were performed. Household characteristics include the size of the household, the principal source of income, the household monthly income level, distance in minutes to a water source, a dummy variable represents whether a household has electricity, and a dummy variable represents whether a household has access to public or shared toilets. ζ_d represents control for district. ψ_m controls for month of the year. η_t controls for variation in interview dates within month. ε_{idmt} represents the error term. Standard errors are clustered at the village/ or town level.

The following mathematical equation depicts the model using the second heat metric, which estimates a simple count data type

$$Y_{idmt} = exp(\beta_0 + \beta_1 Heat_{dmt} + \beta_2 W_{dmt} + \beta_3 X_i + \zeta_d + \psi_m + \eta_t + \varepsilon_{idmt}) \qquad (2)$$

In this equation 2, the coefficient of interest is β_1. This equation is similar to equation 1 in all aspects except the variable of interest in the second equation represents the second heat metric, which is the total number of hot days during the heatwave periods over the thirty days prior to the interview date.

The coefficient of interest, β_1, identifies a percentage change in the number of short-term morbidity days due to a variations in the recorded number of hot days. However, since the impact is dependent on the other variables in the model, we will find the average partial effect.

In order to identify the heat impact using the third heat metric, we run the following equation of a simple count data type

$$Y_{idmt} = exp(\gamma_0 + \gamma_1 Heat_{dmt} + \gamma_2 W_{dmt} + \gamma_3 X_i + \zeta_d + \psi_m + \eta_t + \varepsilon_{idmt}) \tag{3}$$

In equation 3, we keep using the negative binomial regression using the same dependent variable taken from in the first two equations. In order to purge the estimates from potential confounding factors, we use several control variables, including the demographic and socio-economic variables and the individual's health behaviors that have been used in equations (1) and (2). However, in equation (3), our variable of interest represents the accumulated cooling degree days for the same period covering the thirty days prior to the interview dates, and γ_1 is the coefficient of interest, which identifies the percentage change in the number of short-term morbidity days recorded due to an increase in cooling degree days. Furthermore, we will report average partial effect for the same aforementioned reasons of the other heat metrics.

The identifying assumption for the three models is that the realization of heat metrics is random when we control for variations of interview dates within months and the districts fixed effects.

2.5 Results

Our findings establish that prolonged heat exposures (heatwave) contributes towards the significant adverse effect of increasing individual's short-term morbidity days. It can be suggested that the main analysis depends upon how the term heatwave is defined. For the purpose of this paper, a heatwave happens when the maximum daily temperature for three days or more exceeds the 85th percentile of the hottest month in the year for a given location. Not to mention that we test different time durations more and less than three days combining with lower percentile threshold. The results section presents the findings we obtain using the three heat metrics alternatives: (i) the heatwave occurrences for the thirty day period prior to the interview, (ii) the number of hot days during the heatwave periods, (iii) the accumulated cooling degree days for the heat period covering the same thirty days prior to the interview date, which is a truncated measure uses the 22^0C as the threshold, and then tested again with the 30^0C.

2.5.1 Number of heatwave events

In Table 2.2, the results demonstrate the impact of heat by adding up the recorded number of heat events. In columns 1 - 3, we add the weather, individual, and household regressors gradually. Column 4 represents the full and preferred regression. By looking at column 4, the estimate is statistically significant from zero at 10%. The average partial effect of an additional heatwave of three consecutive days or more is associated with an increase in expected recorded number of days of illness by 0.6 day.

This paper's main analysis defines a heatwave as a weather event during which temperature above the 85th percentile is recorded for three or more consecutive days. Therefore, to test whether reducing the set duration might affect the results, we redefine the heat event first by decreasing the set duration to two days or more above the 85th percentile instead of three days, and then to one hot day or more above the 85th percentile instead of three days or more.

Table 2.7 presents the results that we obtain when changing the definition of a heatwave to two be consecutive days or more above the 85th percentile. In columns 2 - 4, we add controls gradually, such that column 4 represents the full regression. The estimate in column 4 shows that by reducing the duration to two days, the impact becomes statistically insignificant and decrease by 55 percentage points.

Similarly, Table 2.8 shows the results from redefining the heatwave using one day or more. The findings show that reducing the duration to one day or more instead of three days or more above the 85th percentile leads to no statistically or economically significant impact on the recorded short-term morbidity days. Columns 2 - 4 are built gradually by adding weather, individual, and household controls. We prefer column 4, which represents the full regression. In this column, the estimate is negligible.

Human bodies sweat and increase blood flow to skin to overcome heat exposure, and these methods help to maintain having internal temperature within the regular range.[5] However, when heat is severe, an individual's health

[5]Kevin Kregel mentioned these methods as ways that human bodies use to react to hot weather exposure. For more details, see the website `https://www.research.colostate.edu/healthyagingcenter/aging-basics/body-temperature/`

is compromised, leading to illness (Madaniyazi et al., 2016). As such, our results show that three days of consecutive extreme heat can adversely impact human health and potentially lead to illness.

2.5.2 Number of hot days

In this section, we implement the number of days with high temperatures that meet the main heatwave definition requirements as another metric for measuring the prolonged heat exposure. Table 2.3 shows the impact of heat using this heat metric. Columns 1 - 4 are built gradually by adding several weather variables and individual's and household's characteristics. However, column 4 is preferred, and includes the full regression. The estimate is statistically significant from zero at the 5% level of significance. The average partial effect of an additional hot day is associated with an increase in the expected number of days of illness by 0.23 days.

Table 2.4 shows that the result is robust even when we replace the metric of the total recorded number of hot days of all heat events with an average length of heatwaves.[6] Columns 1 - 4 are built gradually by adding controls such that column 4 is preferred. It includes the the full regression. The estimate shows that the impact is statistically significant at the 5% level of significance. The average partial effect of an increase in the average heatwave duration by one day is associated with an increase in the expected number of reported days of illness by 0.21 days.

[6]We obtain the average duration of heatwaves in days by dividing the total number of days that lie within these heatwave periods by the number of heat waves events that occured over the thirty day period.

2.5.3 Accumulated cooling degree days

Accumulated cooling degree days (ACDD) represents the third developed heat metric in our analysis. Indeed, cooling degree days are an indicator of cooling load occurring due to outdoor temperature. To identify what temperature level should be used as our baseline temperature to calculate ACDD, we consider the work by Bhatnagar et al. (2018) and the room temperature range 20 - 22^0C. As such, we elect to use the 22^0C as the baseline temperature to calculate ACDD. Table 2.5 shows the heat impact in terms of the accumulated cooling degree days for the heatwave event period. Columns 1 - 4 in Table 2.5 are built gradually by adding the variables of weather, individual's and household's characteristics; however, column 4 is preferred as it represents the full regression. The estimate is statistically significant from zero at the 5% level of significance. The average partial effect of an increase in accumulated cooling degree days by one degree Celsius is associated with an expected change in short-term morbidity by 0.06 days. Since India experiences warm temperatures and that we are focusing on analyzing the impacts of extreme heat events, 22^0C might appear as high enough to substantiate our theory. As such, evidences support the fact that individuals have been shown to easily adapt to such temperature. Therefore, when we rise the cut off to 30^0C,[7] the impact on short-term morbidity increases. The result in Table 2.9 column 2 shows that an increase in ACDD by one degree Celsius above the 30^0C has a statistically significant impact on short-term morbidity. The average partial effect on the expected change in short-term morbidity days is 0.10 days.

[7] According to National Institute of Standard and Technology, a day is considered hot if the temperature is 30^0C or above.

In this previous analysis, we maintained using the baseline version of the heatwave definition. To test whether reducing the heatwave duration has the same impact on short-term morbidity using the ACDD measure, we use heatwave of two consecutive days or more instead of three consecutive days or more. Then we use one day or more instead of three consecutive days or more. In both cases, we use 30^0C as the temperature cut-off to calculate the ACDD. In the first case, when we use a heatwave of two consecutive days or more, the estimate in column 3, Table 2.9, is statistically insignificant. The average partial effect of an increase in the ACDD on the expected change in short-term morbidity days is negligible. The estimate in column 4, Table 2.9 is statistically insignificant. It represents the second case when we adjust the duration to be one hot day or more instead of three days or more. The average partial effect of an increase in the ACDD by one Celsius on short-term morbidity days is negligible.

When analyzing this data, it can be suggested that using the ACDD as heat metric assures that the three days of heat duration is of significant interest. As such, results show that three days of consecutive heat might be challenging for the human body to adapt, which lead to increase an individual's short-term morbidity days.

2.6 Heterogeneity

Individuals display different responses to extreme variation in temperature. A range of factors contribute to these various responses include age, gender and access/distance to a source of water. However, our findings indicate that consecutive heat has a statistically significant impact on short-term morbidity days for individuals of all ages (children, adults, and elderlies). It could be that children and adults are both exposed to heat based on the nature of daily tasks that are assigned to each one of these groups. For instance, walking to a water source as a task for children to bring water to their homes, while going to work such as jobs in agricultural sector for adults.

2.6.1 Gender

Previous studies provide two main theories to explain how heat impacts individuals depending upon their genders. As such, a range of studies reports no significant differences in the recorded morbidity and mortality rates across gender groups (Basu and Ostro, 2008; Stafoggia et al., 2008). However, other studies find that males and females differ in terms of their capacity to process extreme heat exposure (Vaneckova et al., 2008; D'Ippoliti et al., 2010). In our analysis, the results are consistent with the latter findings. Table 6, columns 2 and 3 show the heterogeneous impact of prolonged heat exposure in terms of the gender difference. We implement Welch's t-test to test the difference between the two groups. The test revealed that the difference is statistically significant at a 1% level such that P-value is 0.000. The average partial impact of an additional heatwave of three consecutive days or more on short-term morbidity is 0.58 days for females, and 0.36 for males. As such, the results show that females are more vulnerable to heat exposure.

2.6.2 Distance to water source

Another contributing factor to heat exposure is the distance in time taken to travel in order to access water source. As such, negative health effects are recorded due to individuals being exposed to extreme heat for considerable periods of time. In Table 2.6, columns 4 and 5, we look at the heterogeneous impact in terms of the distance in minutes to the water source. We classify observations into two groups; individuals who have to walk less than 15 minutes to access a water source versus individuals who must walk more than 15 minutes to access a water source. Our findings indicate that long walking distance during heatwaves has a significant adverse effect on short-term morbidity.

2.7 Robustness check

In this part of the analysis, shown in Table 2.10, we implement an alternate heatwave definition. We consider a heatwave when the maximum daily temperature for three consecutive hot days or more exceeds the 75th percentile of the hottest month in a year for a given location. Then, we test it when the duration is four or more projected consecutive hot days.

In other words, we lower the percentile threshold to be the 75th percentile combined with different durations of consecutive periods. Then, we observe how the results differ from our estimates in the primary analysis (baseline version of heatwave definition). Table 2.10, column 1 shows the impact of the number of recorded heatwaves. It demonstrates that lowering the intensity (percentile approach threshold) to be the 75th percentile whilst keeping the duration as three consecutive days or more yields a statistically insignificant estimate. It also must be noted that the partial effect is economically negligible. Moreover, the impact is minimal compared to the effect of the recorded number of heatwaves using the 85th percentile threshold.

In Table 2.10, columns 2 - 4, we present the coefficient estimates using the three heat metrics: the number of heatwaves, the number of hot days, and the cumulative cooling degree days, respectively. Within these three columns, we implement an alternate definition of heatwave where we use the 75th percentile combined with the duration of four or more consecutive days. The coefficient estimate in column 2 is statistically significant at 10% level of significance. The average partial effect of an additional heatwave is associated with 0.51 days increase in recorded number of days of illness. In column 3, the estimate shows a statistically significant impact; the average partial effect of an additional hot day during the heatwave is associated with 0.11 reported days of illness. Column 4 represents the coefficient estimate of the accumulated cooling degree days heat metric. The coefficient estimate is statistically insignificant. The average partial effect of one additional degree Celsius in the intensity is 0.003 days increase in recorded days of illness. Although the duration is extended to four consecutive days or more, the impact is minimal. It is salient that by reducing the percentile threshold, the effect of heat is considerably reduced compared to the baseline heatwave definition of a higher threshold.

In conclusion, our results indicate that prolonged heat exposure might cause a damage to human health such that it increases the recorded short-term morbidity days. Furthermore, lowering the percentile threshold still relatively bring adverse impact on human health when it is combined with a longer duration.

2.8 Discussion

2.8.1 Connected days versus non connected

Within our research, we utilize consecutive days to identify and locate heat impact. The reasoning behind this methodology is that the human body is able to regulate heat to some extent; however, maintaining such regulation might become increasingly challenging when heat is severe and persists for a consecutive number of days. We suggest that consecutive days of high temperatures may become intolerable for individuals and the implications of such heatwaves may result in heat-related illness. However, these adverse effects may not be found when high temperature days are scattered. Therefore, to clarify this idea, in this section, we run the same model that was utilized earlier except that for the heat regressor, we replace it with two heat regressors. The first regressor represents all the hot days in the thirty day period prior to a given interview date. The second regressor represents all the consecutive hot days. The model is depicted in the following equation

$$Y_{idmt} = exp(\alpha_0+\alpha_1 ConnDys_{dmt}+\alpha_2 AllHot_{dmt}+\alpha_3 W_{dmt}+\alpha_4 X_i+\zeta_d+\psi_m+\eta_t+\varepsilon_{idmt})$$

Where Y_{idmt}, W_{dmt}, and X_i are the same variables that are used in equation (1); however, $ConnDys_{dmt}$ represents our variable of interest, which is all consecutive days of high temperatures lasting three or more days in a row.

$AllHot_{dmt}$ represents the number of all hot days in the thirty day period prior to the interview date. The coefficient estimates show that α_1 is 0.084 and significant at 5% level, however, α_2 is 0.0052 with P-value 0.827.

2.8.2 Heat index

In all preceding sections, analyses deal with humidity by introducing a regressor in the model that represent the average value of maximum daily humidity over the thirty day period. In this section, we use the same heatwave metrics and definitions; however, we implement the heat index instead of simply using the the temperature alone. We follow the National Oceanic and Atmospheric Administration U.S. Department (NOAA) by utilizing their equation to calculate the heat index (Rothfusz, L. P., 2014).[8] Then, we keep implementing the three heatwave metrics: number of heat events, number of hot days during heatwave, and the accumulated cooling degree days for the same period of time. We implement the main heatwave definition of having three consecutive hot days or more such that the maximum daily temperatures (Heat Index calculation) exceed the 85th percentile of the hottest month in the

[8]According to the National Weather Service in the U.S., heat index equations were developed by Rothfusz and Steadman as appear in the following regression equations:

$HI = -42.379 + 2.04901523*T + 10.14333127*RH - .22475541*T*RH - .00683783*T*T - .05481717*RH*RH + .00122874*T*T*RH + .00085282*T*RH*RH - .00000199*T*T*RH*RH$. Where HI: the heat index in degrees F. T: temperature in degrees F. RH: relative humidity in percent.

If the RH is less than 13% and the temperature is between 80 and 112 degree F, then the following adjustment $ADJUST = [(13 - RH)/4] * SQRT\{[17 - ABS(T - 95.)]/17\}$ is subtracted from HI equation. Where SQRT is the square root function. ABS is the absolute value function.

However, if the RH is greater than 85% and the temperature is between 80 and 87 degrees F, then the following adjustment $ADJUST = [(RH - 85)/10] * [(87 - T)/5]$ is added to the HI equation.

The last case is when the generated value of heat index is below 80 degrees F. In this case, HI regression become the following: $HI = 0.5 * \{T + 61.0 + [(T - 68.0) * 1.2] + (RH * 0.094)\}$.

As it can be observed, the regression equations of HI uses the degrees F, while our temperatures are in degrees Celsius. Therefore, we convert all temperature values to degrees F, then implement the HI formula, and again we convert the output from the HI equation into Celsius again to be used for our heat metrics.

year. The average partial effect of an additional heatwave of three consecutive days or more is associated with an increase of projected days of illness by 0.27 days. When looking at the second heat metric, we find that the average partial effect of an extra hot day is associated with 0.07 days increase in short-term morbidity. However, the average partial effect of an increase in accumulated cooling degree days by one degree Celsius increases the expected days of illness by 0.003 days.[9] We conclude that when we implement the heat-index in our calculation, we obtain an economically significant effect. However, the effect is less severe in size than the effect we obtained in our main analysis.

2.9 Conclusion

The population of India continues to suffer from the adverse impacts of extreme heat. These impacts significantly contribute towards devastating effects upon agriculture and human health. Since reduced health outcomes directly impacts the economy of any given nation, we focus on providing new hindsight on the effects of extreme heat on human short-term morbidity in India.

We utilize a heatwave definition using the percentile-based threshold in combination with a duration of three consecutive days. We implement three heat metrics to study the impact of prolonged heat exposure "heatwave" on Indian residents' short-term morbidity. Our results provide new evidence to be added to the existing literature on extreme heat and human health. When extreme heat is scattered the adverse impacts on human health are generally

[9] 22^0C is the truncated threshold that is used to calculate the ACDD

mitigated. However, our findings show that individuals are vulnerable to a series of hot days namely three or more consecutive days of heat exposure. Furthermore, we identify a range of critical factors that have the potential to impact upon results. For example, our findings indicate that females are more sensitive to heat exposure, and that the distance to the water source impacts upon short-term morbidity during heatwaves.

This paper therefore provides evidence around the impact of extreme heat on short-term morbidity, with our study directing us towards further areas of key research namely the effect of short-term morbidity on labor supply during heatwaves.

References

Anderson, B. G., & Bell, M. L. (2009). Weather-related mortality: How heat, cold, and heat waves affect mortality in the United States. *Epidemiology*, 20(2), 205–213. https://doi.org/10.1097/EDE.0b013e318190ee08

Auffhammer, M., Hsiang, S. M., Schlenker, W., & Sobel, A. (2013). Using weather data and climate model output in economic analyses of climate change. *Review of Environmental Economics and Policy*, 7(2), 181-198. https://doi.org/10.1093/reep/ret016

Barreca, A. I. (2012). Climate change, humidity, and mortality in the United States. *Journal of Environmental Economics and Management*, 63(1), 19-34. https://doi.org/10.1016/j.jeem.2011.07.004

Barreca, A., Clay, K., Deschenes, O., Greenstone, M., & Shapiro, J. S. (2016). Adapting to climate change: the remarkable decline in the U.S. Temperature-mortality relationship over the twentieth century. *Journal of Political Economy*, 124(1), 105-159. https://doi.org/10.1086/684582

Barro, R. J. (2000). Inequality and growth in a panel of countries. *Journal of Economic Growth*, 5(1), 5-32. https://doi.org/10.1023/A:1009850119329

Basu, R., & Ostro, B. D. (2008). A multicounty analysis identifying the populations vulnerable to mortality associated with high ambient temperature in California. *American Journal of Epidemiology*, 168(6), 632-637. https://doi.org/10.1093/aje/kwn170

Bell, M. L., O'Neill, M. S., Ranjit, N., Borja-Aburto, V. H., Cifuentes, L. A., & Gouveia, N. C. (2008). Vulnerability to heat-related mortality in Latin America: A case-crossover study in São Paulo, Brazil, Santiago, Chile and Mexico City, Mexico. *International Journal of Epidemiology*, 37(4), 796-804. https://doi.org/10.1093/ije/dyn094

Bhaskaran, K., Hajat, S., Haines, A., Herrett, E., Wilkinson, P., & Smeeth, L. (2010). Short term effects of temperature on risk of myocardial infarction in England and Wales: Time series regression analysis of the Myocardial Ischaemia National Audit Project (MINAP) registry. *BMJ*, 341, c3823-c3823. https://doi.org/10.1136/bmj.c3823

Bhatnagar, M., Mathur, J., & Garg, V. (2018). Determining base temperature for heating and cooling degree-days for India. *Journal of Building Engineering*, 18, 270-280. https://doi.org/10.1016/j.jobe.2018.03.020

Chua, K., Coggins, J. S., Miller, S., & Mohtadi, H. (2019). Measuring heat waves and their effects on economic output. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3424125

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. *Routledge.*

Dell, M., Jones, B. F., & Olken, B. A. (2012). Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, 4(3), 66-95. https://doi.org/10.1257/mac.4.3.66

Desai, Sonalde, and Reeve Vanneman and National Council of Applied Economic Research, New Delhi. India Human Development Survey-II (IHDS-II), 2011-12. ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-07-31. http://doi.org/10.3886/ICPSR36151.v2

Deschênes, O., & Greenstone, M. (2011). Climate change, mortality, and adaptation: evidence from annual fluctuations in weather in the U.S. *American Economic Journal: Applied Economics*, 3(4), 152-185. https://doi.org/10.1257/app.3.4.152

Deschênes, O., & Moretti, E. (2007). Extreme weather events, mortality and migration (No.13227). *Cambridge, MA: National Bureau of Economic Research*. https://doi.org/10.3386/w13227

D'Ippoliti, D., Michelozzi, P., Marino, C., Kirchmayer, U., Analitis, A., Medina-Ramón, M., ... Perucci, C. A. (2010). The impact of heat waves on mortality in 9 European Cities: Results from the EuroHEAT project. *Environmental Health*, 9. https://doi.org/10.1186/1476-069X-9-37

Du, J., Park, Y.-T., Theera-Ampornpunt, N., McCullough, J. S., & Speedie, S. M. (2012). The use of count data models in biomedical informatics evaluation research. *Journal of the American Medical Informatics Association*, 19(1), 39-44. https://doi.org/10.1136/amiajnl-2011-000256

Garg, T., Jagnani, M., & Taraz, V. (2017). Effects of heat stress on physiology and livelihoods: implications for human capital accumulation. Working paper.

Graff Zivin, J., & Neidell, M. (2014). Temperature and the allocation of time: Implications for climate change. *Journal of Labor Economics*, 32(1), 1-26. https://doi.org/10.1086/671766

Hajat, S., Armstrong, B., Baccini, M., Biggeri, A., Bisanti, L., Russo, A., ... Kosatsky, T. (2006). Impact of high temperatures on mortality:

is there an added heat wave effect? *Epidemiology*, 17(6), 632-638. https://doi.org/10.1097/01.ede.0000239688.70829.63

Hansen, A., Bi, P., Nitschke, M., Ryan, P., Pisaniello, D., & Tucker, G. (2008). The Effect of heat waves on mental health in a temperate Australian city. *Environmental Health Perspectives*, 116(10), 1369-1375. https://doi.org/10.1289/ehp.11339

Hansen, A. L., Bi, P., Ryan, P., Nitschke, M., Pisaniello, D., & Tucker, G. (2008). The effect of heat waves on hospital admissions for renal disease in a temperate City of Australia. *International Journal of Epidemiology*, 37(6), 1359-1365. https://doi.org/10.1093/ije/dyn165

Huang, W., Kan, H., & Kovats, S. (2010). The impact of the 2003 heat wave on mortality in Shanghai, China. *Science of The Total Environment*, 408(11), 2418-2420. https://doi.org/10.1016/j.scitotenv.2010.02.009

Knowlton, K., Kulkarni, S., Azhar, G., Mavalankar, D., Jaiswal, A., Connolly, M.,... The Ahmedabad heat and climate study group. (2014). Development and implementation of South Asia's first heat-health action plan in Ahmedabad (Gujarat, India). *International Journal of Environmental Research and Public Health*, 11(4), 3473-3492.

https://doi.org/10.3390/ijerph110403473

Knowlton, K., Rotkin-Ellman, M., King, G., Margolis, H. G., Smith, D., Solomon, G., ... English, P. (2009). The 2006 California heat wave: Impacts on hospitalizations and emergency department visits. *Environmental Health Perspectives*, 117(1), 61-67. https://doi.org/10.1289/ehp.11594

Lowe, D., Ebi, K. L., & Forsberg, B. (2011). Heatwave early warning systems and adaptation advice to reduce human health consequences of heatwaves. *International Journal of Environmental Research and Public Health*, 8(12), 4623-4648. https://doi.org/10.3390/ijerph8124623

Luber, G., & McGeehin, M. (2008). Climate change and extreme heat events. *American Journal of Preventive Medicine*, 35(5), 429-435. https://doi.org/10.1016/j.amepre.2008.08.021

Manning, W. G., Liu, C.-F., Stoner, T. J., Gray, D. Z., Lurie, N., Popkin, M., & Christianson, J. B. (1999). Outcomes for medicaid beneficiaries with schizophrenia under a prepaid mental health carve-out. *The Journal of Behavioral Health Services & Research*, 26(4), 442-450. https://doi.org/10.1007/BF02287304

Mastrangelo, G., Fedeli, U., Visentin, C., Milan, G., Fadda, E., & Spolaore, P. (2007). Pattern and determinants of hospitalization during heat waves: An ecologic study. *BMC Public Health*, 7(1), 200. https://doi.org/10.1186/1471-2458-7-200

Mazdiyasni, O., & AghaKouchak, A. (2015). Substantial increase in concurrent droughts and heatwaves in the United States. *Proceedings of the National Academy of Sciences*, 112(37), 11484-11489. https://doi.org/10.1073/pnas.1422945112

Madaniyazi, L., Zhou, Y., Li, S., Williams, G., Jaakkola, J. J. K., Liang, X., Liu, Y., Wu, S., & Guo, Y. (2016). Outdoor Temperature, Heart Rate and Blood Pressure in Chinese Adults: Effect Modification by Individual Characteristics. *Scientific Reports*, 6(1), 21003. https://doi.org/10.1038/srep21003

Mazdiyasni, O., AghaKouchak, A., Davis, S. J., Madadgar, S., Mehran, A., Ragno, E., ... Niknejad, M. (2017). Increasing probability of mortality during Indian heat waves. *Science Advances*, 3(6), e1700066. https://doi.org/10.1126/sciadv.1700066

Oudin Åström, D., Bertil, F., & Joacim, R. (2011). Heat wave impact on morbidity and mortality in the elderly population: A review of recent studies. *Maturitas*, 69(2), 99-105. https://doi.org/10.1016/j.maturitas.2011.03.008

Park, J., & Kim, J. (2018). Defining heatwave thresholds using an inductive machine learning approach. *PLOS ONE*, 13(11), e0206872. https://doi.org/10.1371/journal.pone.0206872

Peng Bi, Williams, S., Loughnan, M., Lloyd, G., Hansen, A., Kjellstrom, T., ... Saniotis, A. (2011). The effects of extreme heat on human mortality and morbidity in Australia: implications for public health. *Asia Pacific Journal of Public Health*, 23(2_suppl), 27S-36S. https://doi.org/10.1177/1010539510391644

Phelps, E. S. (1980). Investment in humans, technological diffusion, and economic growth. *Studies in Macroeconomic Theory* (pp. 133-139). https://doi.org/10.1016/B978-0-12-554002-5.50015-7

Rey, G., Fouillet, A., Bessemoulin, P., Frayssinet, P., Dufour, A., Jougla, E., & Hémon, D. (2009). Heat exposure and socio-economic vulnerability as synergistic factors in heat-wave-related mortality. *European Journal of Epidemiology*, 24(9), 495-502. https://doi.org/10.1007/s10654-009-9374-3

Rothfusz, L. P. (2014). The Heat Index Equation. *National Weather Service-Weather Prediction Center-National Oceanic and Atmospheric Administration U.S. Department (NOAA)*. Retrieved June 1, 2022, from https://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37), 15594-15598. https://doi.org/10.1073/pnas.0906865106

Schlenker, Wolfram, & Lobell, D. B. (2010). Robust negative impacts of climate change on African agriculture. *Environmental Research Letters*, 5(1), 014010. https://doi.org/10.1088/1748-9326/5/1/014010

Semenza, J. C., Rubin, C. H., Falter, K. H., Selanikio, J. D., Flanders, W. D., Howe, H. L., & Wilhelm, J. L. (1996). Heat-related deaths during the July 1995 heat wave in Chicago. *New England Journal of Medicine*, 335(2), 84-90. https://doi.org/10.1056/NEJM199607113350203

Stafoggia, M., Forastiere, F., Agostini, D., Caranci, N., de'Donato, F., Demaria, M., ... Perucci, C. A. (2008). Factors affecting in-hospital heat-related mortality: a multi-city case-crossover analysis. *Journal of Epidemiology & Community Health*, 62(3), 209-215.

https://doi.org/10.1136/jech.2007.060715

Tolika, K. (2019). Assessing heat waves over Greece using the excess heat factor (EHF). *Climate*, 7(1), 9. https://doi.org/10.3390/cli7010009

Toloo, G., FitzGerald, G., Aitken, P., Verrall, K., & Tong, S. (2013). Evaluating the effectiveness of heat warning systems: Systematic review of epidemiological evidence. *International Journal of Public Health*, 58(5), 667-681. https://doi.org/10.1007/s00038-013-0465-2

Vaneckova, P., Beggs, P. J., de Dear, R. J., & McCracken, K. W. J. (2008). Effect of temperature on mortality during the six warmer months in Sydney, Australia, between 1993 and 2004. *Environmental Research*, 108(3), 361-369. https://doi.org/10.1016/j.envres.2008.07.015

Zhang, L., Zhang, Z., Ye, T., Zhou, M., Wang, C., Yin, P., & Hou, B. (2018). Mortality effects of heat waves vary by age and area: A multi-area study in China. *Environmental Health*, 17(1), 54. https://doi.org/10.1186/s12940-018-0398-6

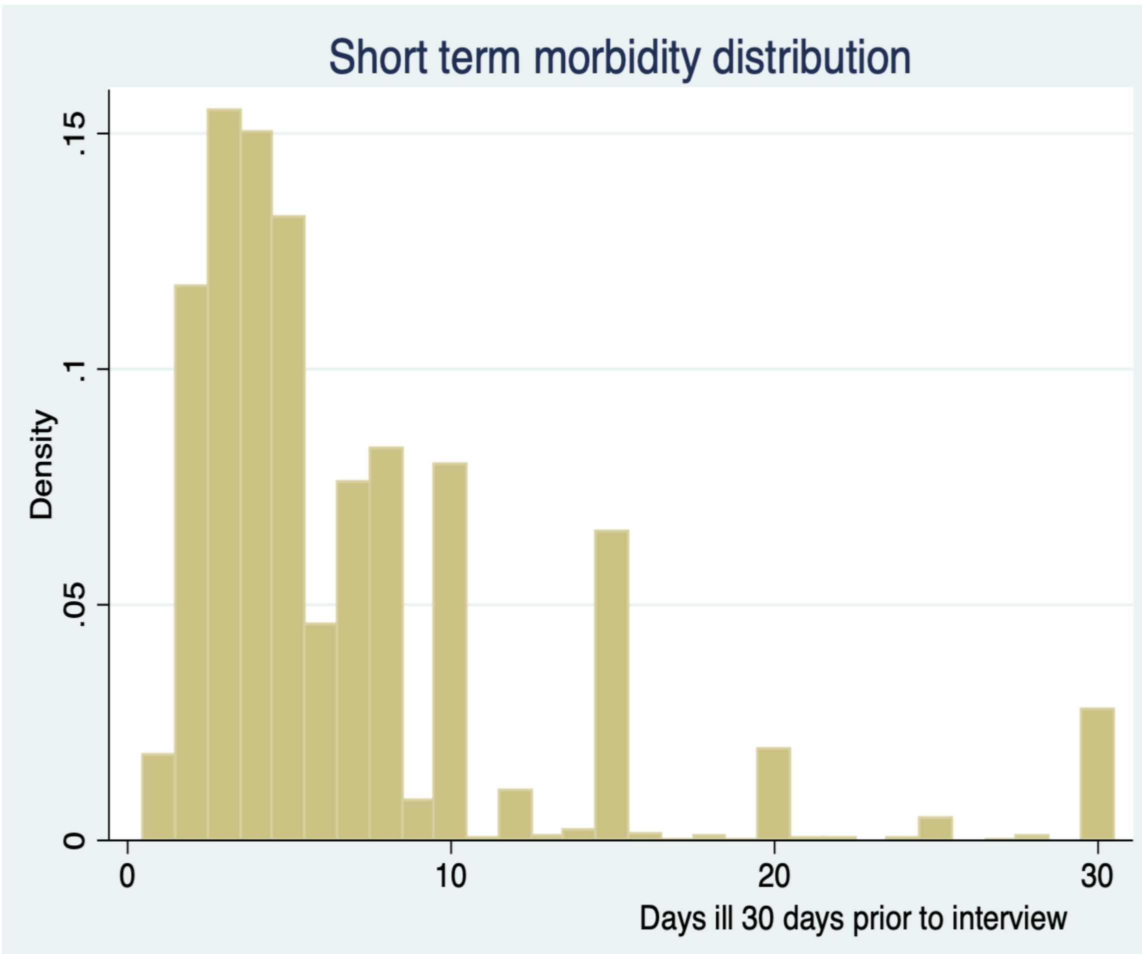

Figure 2.1: Variation in the dependent variable (excluding zeros).

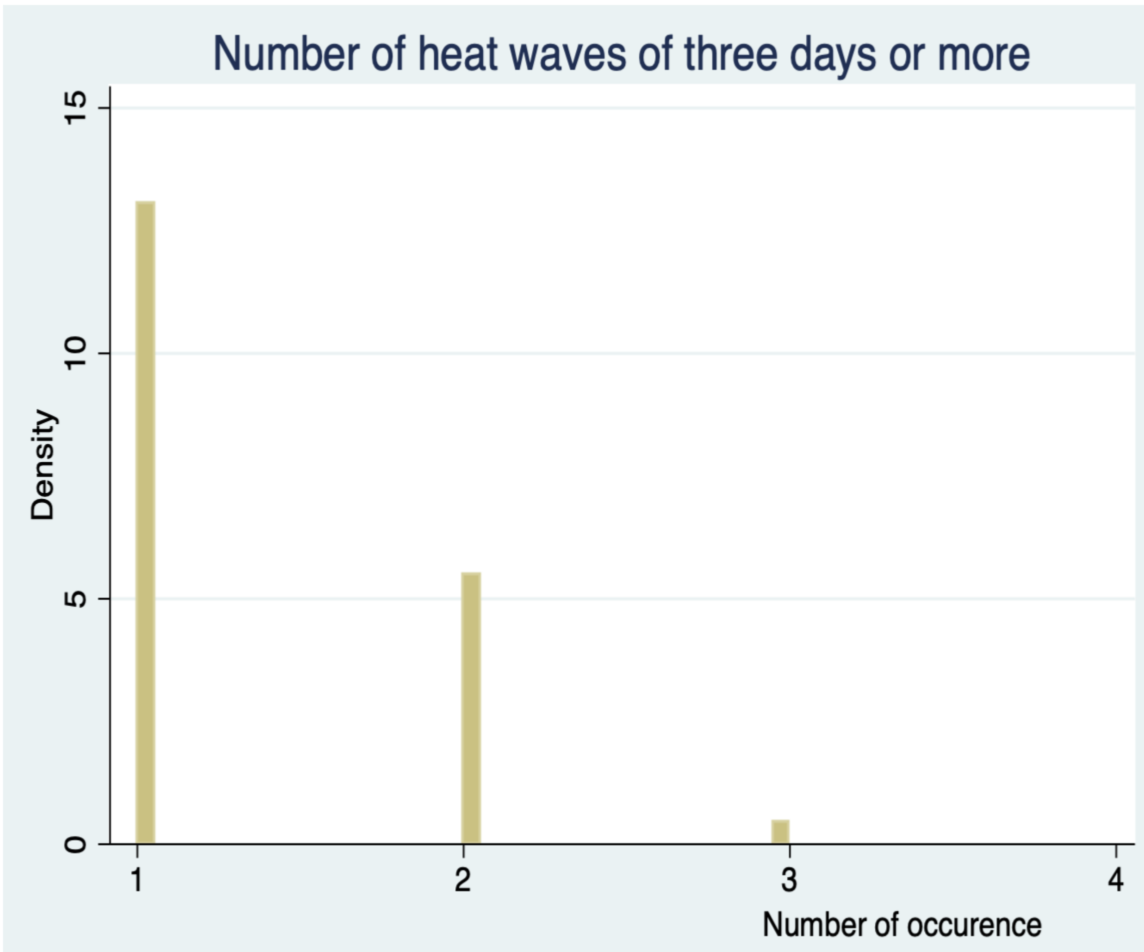

Figure 2.2: Distribution of the number of heatwaves (excluding zeros)

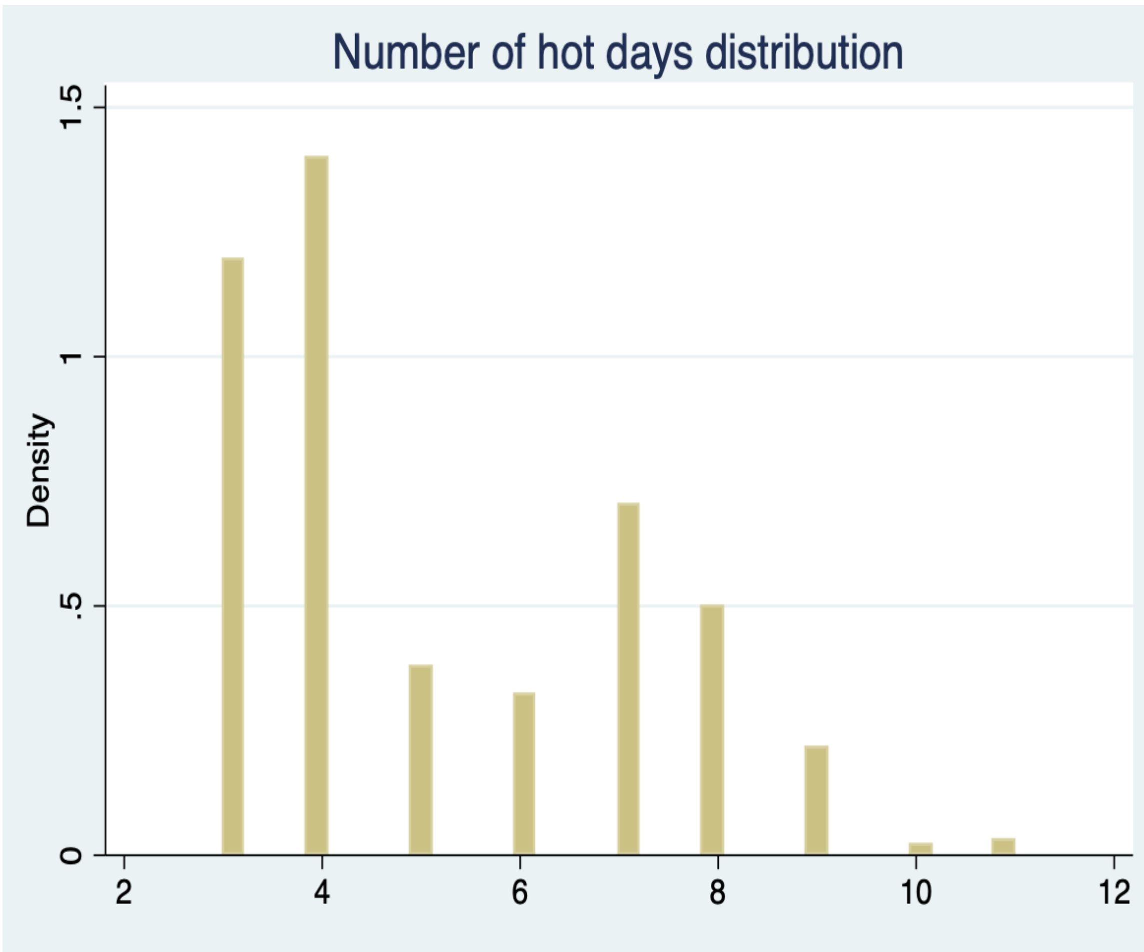

Figure 2.3: Variation in number of hot days (excluding zeros)

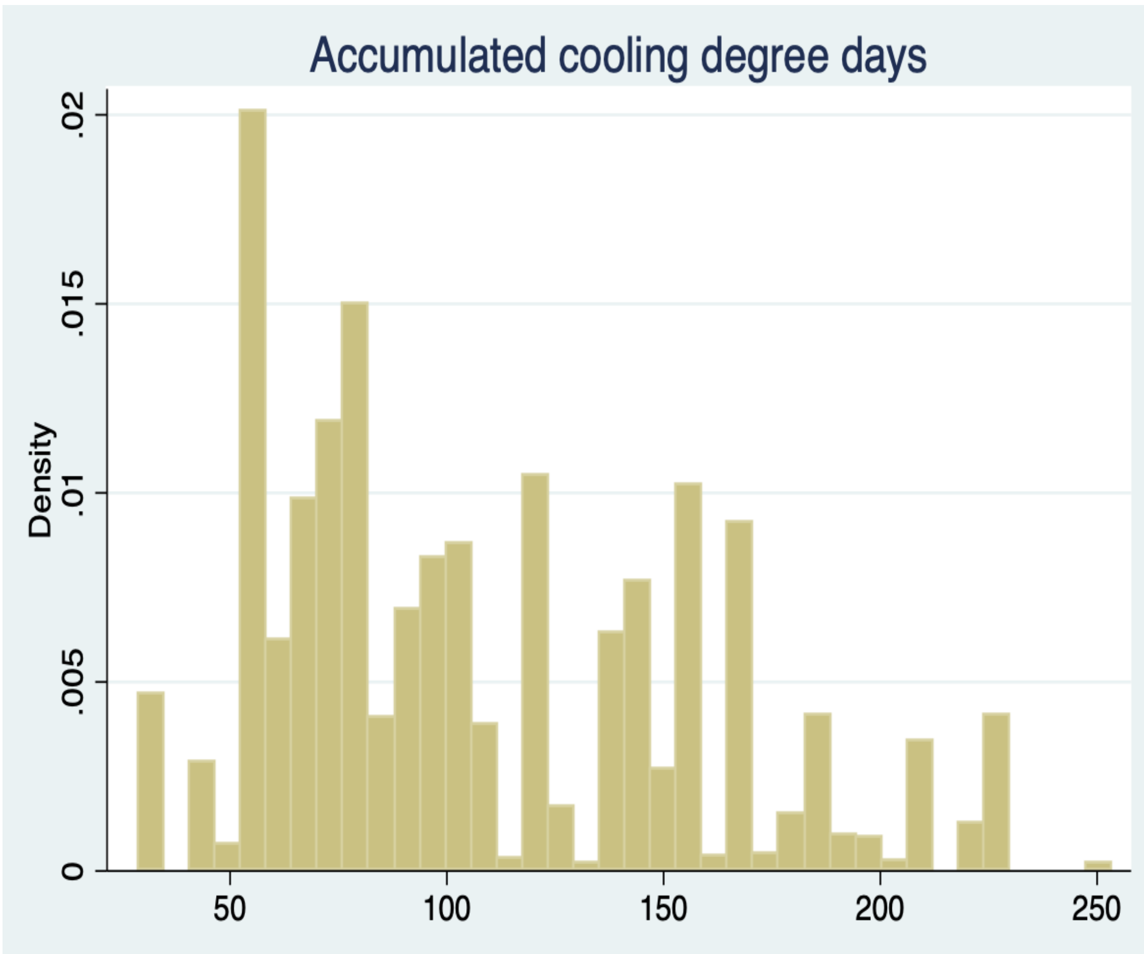

Figure 2.4: Distribution of accumulated cooling degree days (excluding zeros)

Table 2.1: Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Short term morbidity days (day)	61,823	1.315158	3.710923	0	30
Average relative humidity (%)	61,823	0.512993	0.215458	0.112994	0.965566
Average wind speed (km/h)	61,823	2.712858	0.962245	1.164031	7.054047
Total amount of rain (mm)	61,823	109.4759	181.3246	0.000617	897.0751
Household size	61,823	5.778125	2.524525	1	20
Distance to water in minutes	61,823	12.13226	14.90323	0	240
Washing hands	61,823	0.954855	0.207624	0	1
Access to electricity	61,823	0.763616	0.424864	0	1
Average monthly income (Rupees)	61,823	6115.989	6766.968	0	127590
Female	61,823	0.501545	0.500002	0	1
Age	61,823	28.07452	20.23581	0	90

Note: Short term morbidity is the number of short-term morbidity days reported during the thirty day period prior to the interview date. Household size represents number of individuals live in the same house. Washing hands is a dummy variable represents perseverance on washing hands after defecation. Access to electricity is a dummy variable represents whether a household has access to electricity. Average monthly income is a continuous variable represents household monthly income. Female is a dummy variable represent gender. Age is a continuous variable represent an individual's age in year.

Table 2.2: Number of heatwaves

	(1)	(2)	(3)	(4) Full
illness days				
Number of heatwaves	0.313**	0.321**	0.221*	0.245*
	(0.13205)	(0.13733)	(0.1328)	(0.13995)
Weather controls		√		√
Individual controls			√	√
Household controls			√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.3: Number of days in heatwave events

	(1)	(2)	(3)	(4) Preferred
illness days				
Number of hot days	0.0986***	0.099***	0.0729**	0.078**
	(0.03424)	(0.03540)	(0.034318)	(0.03439)
Weather controls		√		√
Individual controls			√	√
Household controls			√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: The second heat metric represents the number of recorded hot days during the heatwave events. * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.4: Heatwave average duration length in days

	(1)	(2)	(3)	(4) Full
illness days				
Average heatwave duration	0.10***	0.10***	0.074**	0.079**
	(0.03466)	(0.03588)	(0.03475)	(0.03508)
Weather controls		√		√
Individual controls			√	√
Household controls			√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.5: Intensity

	(1)	(2)	(3)	(4) Full
illness days				
Accumulated cooling degree days	0.017*	0.018*	0.016*	0.018**
	(0.00974)	(0.0098)	(0.00966)	(0.00946)
Weather controls		√		√
Individual controls			√	√
Household controls			√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: The third heat metric represents ACDD above 22^0C. * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.6: Heterogeneity

	Preferred	Gender		Distance to water	
	(1)	(2)	(3)	(4)	(5)
		Male	Female	< 15mints	> 15mints
illness days					
Number of heatwaves	0.245*	0.17	0.32**	0.149	0.831***
	(0.13995)	(0.20143)	(0.16186)	(0.15437)	(0.27671)
Weather controls	√	√	√	√	√
Individual controls	√	√	√	√	√
Household controls	√	√	√	√	√
Month FE	√	√	√	√	√
District FE	√	√	√	√	√
Observations	61,823	30,816	31,007	44,835	16,988

Note: * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.7: Heat of two or more consecutive days

	(1) Preferred	(2)	(3)	(4) Full
illness days				
numHWs3	0.245*			
	(0.13995)			
numHWs2		0.15*	0.099	0.11
		(0.08257)	(0.07763)	(0.07820)
Weather controls		√		√
Individual controls			√	√
Household controls			√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: numHWs3 represents the number of heatwave events of three or more consecutive hot days above the 85th percentile base. numHWs2 represents the number of heatwave events of the two or more consecutive hot days above the 85th percentile base. * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.8: Heat of one or more consecutive days

	(1) Preferred	(2)	(3)	(4) Full
illness days				
numHWs3	0.245*			
	(0.13995)			
numHWs1		-.011	-.006	-.011
		(0.03905)	(0.03733)	(0.03833)
Weather controls	√	√		√
Individual controls	√		√	√
Household controls	√		√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: numHWs3 represents the number of heatwave events of three or more consecutive hot days above the 85th percentile base. numHWs1 represents the number of heatwave events of one or more consecutive hot days above the 85th percentile base. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Standard errors in parentheses clustered at the town/or village.

Table 2.9: Intensity at 30^0C comparing to 22^0C for different heat wave durations

	(1) ACDD22D3	(2) ACDD30D3	(3) ACDD30D2	(4) ACDD30D1
illness days				
ACDD22D3	0.01882**			
	(0.009468)			
ACDD30D#		0.0289*	0.00735	0.0001
		(0.01568)	(0.01153)	(0.000565)
Weather controls	$\surd$	$\surd$		$\surd$
Individual controls	$\surd$		$\surd$	$\surd$
Household controls	$\surd$		$\surd$	$\surd$
Month FE	$\surd$	$\surd$	$\surd$	$\surd$
District FE	$\surd$	$\surd$	$\surd$	$\surd$
Observations	61,823	61,823	61,823	61,823

Note: ACDD22D3 represents the accumulated cooling degree days above 22^0C for the case of heatwave baseline definition. ACDD30D# represents the accumulated cooling degree days above 30^0C for different heatwave duration lengths such as three days or more (ACDD30D3), two days or more (ACDD30D2), and one day or more consecutively (ACDD30D1). * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Table 2.10: 75Th percentile and different heat waves durations of consecutive days

	Three days duration	Metrics based on four days duration		
	(1)	(2)	(3)	(4)
	numHWs3	numHWs4	numHotDays4	ACDD22D4
illness days				
Heat wave metric	0.018	0.21*	0.04*	0.0018
	(0.06986)	(0.11312)	(0.02445)	(0.00108)
Weather controls	√	√	√	√
Individual controls	√	√	√	√
Household controls	√	√	√	√
Month FE	√	√	√	√
District FE	√	√	√	√
Observations	61,823	61,823	61,823	61,823

Note: numHWs3 represents the number of heatwaves for three or more consecutive days above the 85th percentile (baseline heatwave definition). numHWs4 represents the number of heatwaves of four or more consecutive days above the 75th percentile. numHotDays4 represents the number of hot days during heat events periods of four or more consecutive days, where the temperatures exceed the 75th percentile. ACDD22D4 represents the accumulated cooling degree days above 22^0C for heatwave duration of four or more consecutive days exceeding the 75th percentile. * p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses clustered at the town/or village.

Chapter 3

Intra-Family Marriage and Pregnancy Outcomes

3.1 Introduction

The aim of this chapter is to investigate the impact of intra-family marriage on negative pregnancy outcomes using the ordinary least squares model (OLS) working with a large sample of young women in India. Within existing literature, marriage between blood relatives is referred to as consanguineous marriage. Marrying a blood relative is not prohibited and remains an option for around 1000 million people worldwide. The majority of this proportion comes from the north and sub-Saharan Africa, the Middle East, and west, central, and south Asia (Hamamy, 2012; Bittles and Black, 2010b). Unless stringent governmental policies curb the spread of consanguineous marriage, it is up to individual's decision which generally depends upon personal incentives shaped by costs and benefits. In many countries, consanguineous marriages are a concern due to their potential adverse effect on offspring health, which can generate significant negative economic outcomes in terms of medical costs and decreased human capital (Akyol and Mocan, 2020).

Studies in medical and public health literature have shed some light on the adverse effects of consanguineous marriage on offspring health (Shawky, 2011; Zlotogora, 1991; Al-Rahal, 2018; Bener et al., 2007; Bittles and Black, 2009; Bachir and Aouar, 2019; Al-Gazali, 1998). Furthermore, it appears that the closer the blood relation between the parents, the higher the possibility of offspring health problems (Zlotogora, 1991). Although some studies demonstrate a positive correlation between fetal loss rate and consanguineous marriage (Bittles and Black, 2010; Zlotogora, 1991), others failed to identify a significant relation (Tadmouri et al., 2009). As such, the exact connection between consanguineous marriage and offspring health is still unclear. It must be noted that the resulting impact on their children's health also depends on demographic, social, and economic factors. Hence, not accounting for these factors will likely cause the impact of consanguineous marriage on offspring health to be biased. Therefore, this study aims to provide evidence to investigate this gap in research using a rigorous dataset and empirical method.

Our study fills a gap in this existing literature by making three particular contributions. Firstly, as opposed to most health studies, we use a large, good quality representative sample of eligible Indian women aged 15-49 years old. The dataset, India Human Development Survey (IHDS-II), is a rich source of demographic, socioeconomic and health information, and covers 34 states and 384 districts in India. Secondly, a woman experiencing stillbirth, miscarriage, or child death after birth is the outcome variable, which might be easier to investigate than to assess the impact on health once adulthood has been reached, where more non-genetic confounding factors should be considered. Thirdly, we apply the ordinary linear squares model (OLS) controlling for a wide range of factors to investigate the relation. Using this rich dataset (IHDS-II) allows the OLS model to include a set of family and financial factors, which makes our study less of a concern compared to other studies in terms of omitted variable bias. As a robustness exercise, we support the OLS model by implementing the standard two-stage least squares method (2SLS). The instrumental variable is the ambient level of violence towards women in the area in which they reside. The latter approach is known for its ability to allow for explicit estimation of the causal impact. Thus, we can check the robustness of the OLS conclusion of the impact of consanguineous marriage on offspring health (pregnancy outcomes).

Our empirical findings confirm the significant adverse impact of consanguineous marriage on pregnancy outcomes. The OLS results show that a woman in a consanguineous marriage has a higher probability of experiencing perinatal and child death after birth by 2.8 percentage points. This result is consistent with the result of the 2SLS approach, where the latter shows a slightly smaller adverse impact of 2.2 percentage points. The difference is statistically insignificant. This paper provides additional evidence on the adverse impact of consanguineous marriage on offspring health.

The rest of the paper is organized as follows. Section 2 reviews the existing literature. Section 3 presents the data. Section 4 outlines the methodology applied. Section 5 discusses the results, and section 6 concludes.

3.2 Literature review

Consanguineous marriage is common in various communities and varies in proportion within and across countries (Bennett et al., 1999; Bennett et al., 2001). The rate of consanguineous marriages also differs depending on culture, religion, customs and income level. In western society, historical and religious prejudice are central to the negative popular opinion on consanguineous marriage (Bittles 2009; Reddy, 2009) which has led to changes in legislation in many countries. For instance, in the United States, first cousin marriages are banned in thirty-one states (Paul and Spencer, 2008).

Conflicting data can be identified when analyzing many studies that have focused on the connection between consanguinity and offspring health. Bittles and Black (2010a) reviewed a vast number of these papers, and observed that contradictory impacts were found. The authors identified that there is no conclusive evidence. This could be a result of data problems such as small sample sizes, data quality, or lack of good empirical strategies.

The authors observed that consanguineous marriage was reported to have a contradictory impact even within the same disease. For example, when looking at the effect of consanguinity on breast cancer of adulthood, Denic and Bener (2001) show that women aged between forty to fifty years old conceived by consanguineous parents have a 3% lower risk of contracting cancer compared to women conceived by non-consanguineous parents. However, Liede et al. (2002) find recessively inherited genes may significantly contribute to increasing the risks of breast and ovarian cancers. In the case of heart disease, Ismail et al. (2004) reported that consanguinity is associated with a premature myocardial infraction and suggested further studies, while Jaber et al. (1997) found no association and concluded that even within the population with a high rate of consanguinity there is no increase in the prevalence of heart disease. However, Bittles and Black (2010a) added that studies rely on comparison of consanguineous versus non-consanguineous offspring without accounting for non-genetic factors. This motivates our current exploration of consanguineous marriages' adverse impact through working with a large representative sample implementing the ordinary least squares model.

Standard of living and individuals' level of education are important non-genetic factors that contribute in determining the relation between consanguineous marriage and health impact. In wealthy countries such as Qatar, these factors are less concern when determining the exact relation, though data problems exists. Berner et al. (2007) investigated the impact on common adult health problems in Qatar. Their findings showed a significant increase in the prevalence of several diseases, including vision loss, diabetes mellitus, mental disorders, heart problems, and congenital abnormalities. However, the data problems such as selection bias in sampling might be a concern such that over fifty percent of the observations are in low economic status. The authors suggested further in-depth work to determine the relation in terms of morbidity and mortality problems.

Higher rates of observable data and timing of critical illness can be obtained by analyzing birth data, which may indicate as to why many studies are motivated in investigating the relation between consanguinity and offspring health outcomes at birth. Some of theses outcomes include anthropometric measurements (Kulkarni et al., 1990; Obeidat et al., 2010; Sibert et al., 1979).

Congenital anomalies (Naibkhil and Chitkara, 2016; Rosana, 2000; Agha et al., 2006). Reproductive outcome (Al-Awadi et al., 2008; Husain and Bunyan, 1997; Jaber et al., 1997). Overall, existing studies provide inconclusive evidence around the type of relation between offspring health problems and consanguineous marriage. This explains why we use a large dataset of Indian women to investigate the impact of consanguineous marriage on perinatal and child death after birth.

India is one of numerous governments around the world that provides public health services to their population. According to IHDS-II, 8% of all marriages in India are consanguineous marriages. Consanguineous marriage may have a significant impact on public health services as a consequence of offspring health problems. The government of India adopts Universal Health Coverage to ensure the residents obtain health services with no financial burden. The 2013 statistics show that the public health expenditure is around 1% of its GDP and constitutes 29% of the total health expenditure (World Health Organization, 2000; Nandraj and Dilip, 2017). While the government struggles to fund the public health sector, birth defects of consanguineous marriages would increase the demand of using the public health facilities. Hence, because of fund limitation, this might negatively affect the quality of public health services provided.

According to Verma and Bijarnia (2002), consanguineous marriage may well contribute towards the unusually statistically high number of Indian children born annually with health problems which sits at around one million. These include congenital malformations, down syndrome, Glucose-6-phosphate dehydrogenase deficiency (G6PD), beta-thalassemia, sickle cell disease, and amino acid disorders. Consanguineous marriage accounts for around 9.9% of all marriages in India (Sharma et al., 2021). Despite these significant figures, the relation between consanguineous marriage and health problems has been under-represented in research. Paddaiah and Reddy (1980) shed some light on the adverse impact of parental consanguinity on anthropometric measurements in newborn babies.[1] Their findings confirm adverse impacts only in families from lower socioeconomic backgrounds. The study covers only one small city from the subcontinent of India, depends on a simple comparison between families, and fails to consider a wide range of non-genetic factors. Ramegowda and Ramachandra (2006) studied the impact of consanguineous marriage on congenital heart diseases using a sample of 344 families in Mysore, a city in southern India. Their findings show that first cousin marriages increase the risk of congenital heart diseases by 44.68%, and uncle-nice marriages increase the risk by 46.81%. Devi et al. (1987) investigate the adverse effect of consanguinity on genetic diseases using a sample of 407 infants and children. Their findings show that consanguinity leads to an increase in the risk of single gene defects, polygenic disorders, and Down's syndrome by 10%, 3% and 1.7%, respectively. However, these studies present with small sample sizes in addition to lacking robust empirical approaches. Using small sample size or weak empirical method are potential issues of bias and low statistical power. The impact of consanguinity on health is under represented in literature which provides a space for greater research around the potential connections with a wide range of health problems. According to Patki and Chauhan (2016), recurrent spontaneous miscarriages rate among Indian

[1] The anthropometric measurements that the study covers include weight, height, chest girth, calf girth, head girth and head length.

women aged between eighteen and forty-five years old is around 32%. Such high rate validates the purpose of our research.

3.3 Data

We use the India Human Development Survey 2011/2012 (IDHS-II), a representative sample of 42,142 households spreading over 1,420 villages and 1,042 urban neighbourhoods across India. Based on two one-hour interviews for each household covering 384 districts of India, the survey provides a rich source of health and behaviors information, demographics, and socioeconomic variables. It is an outcome of collaborative work between the National Council of Applied Economic Research (NCAER), New Delhi, and the University of Maryland.

In our study, we examine eligible married women who have given birth aged fifteen to forty-nine years old, representing 25,405 individuals as per the IHDS-II. Table 3.1 shows summary statistics of the key variables in the context of this study. The dependent variable is based on the pregnancy outcome, more specifically it takes a value of 1 if an eligible woman experienced stillbirth, miscarriage, or child death after birth, and 0 otherwise.[2] When analyzing our sample, we find that an average of 31% of offsprings presented with birth problems. The main independent variable of interest is a dummy variable that takes a value of 1 if a woman is related to her partner by blood, and 0 if she isn't.[3] On average, around 8% of marriages in our sample are intra-family

[2]According to Plagge et al. (2007), perinatal loss includes both miscarriage and stillbirth. Miscarriage occurs when pregnancy loss occurs before the 20th week of gestation. If pregnancy loss occurs after the 20th week, it is defined stillbirth.

[3]We do not differentiate whether the couples are first or second cousin because of data limitation.

marriages.

Our data includes controls to account for demographic, socioeconomic and geographic factors that might influence health outcome and the occurrence of intra-family marriage. The data shows that Hinduism, Islam, Christianity, Sikhism, and Buddhism are the five major religions in the country and constitute 83.7, 11.5, 1.9, 1.3, and 0.8% of the population, respectively. The proportion of consanguineous marriages for each religion from the total cases of intra-family marriages are 75.6, 20, 2.26, 0.19 and 1.6%, respectively, for Hinduism, Islam, Christianity, Sikh, and Buddhism. The average household monthly income is 11,368 Indian rupees.[4] This amount is relatively low, though income is a financial factor that affects offspring health and might influence the likelihood of intra-family marriage. A dummy variable represents whether a household owns a land or cultivate agricultural land. On average, around 47% of the households own and cultivate agricultural land. Woman's education variable represents the number of completed years of schooling, where the average education level of women is five completed years of schooling in total. A dummy variable represents a woman's health status, on average, 77% of women in our sample reported having good health status. Also, a woman's height and weight are indicators of her general health. The average women's height and weight are 152 centimeters and 51 kilograms, respectively. The average age of women when they marry is 18.02 years. The average household size is 5.4 persons. A dummy variable represents whether a woman lives in an urban area. On average, around 32% of eligible women live in urban areas.

Although this is a caveat in our variable, as the closer the blood relation between parents, the higher the risk of offspring health problems (Zlotogora, 1991), it still does not falsify our empirical strategy.

[4] 11,368 Indian rupees is equivalent to amount of $150.

Dummy variables represent the state of residence.

Our instrumental variable (IV) is the yearly ambient level of violence towards women by district computed from the Indian National Crime Record Bureau. To obtain the ambient level of violence, we sum the registered incidences of kidnapping against women to the incidences of rape and divide by the population size and finally multiply by 100,000.[5] Figure 3.1 and 3.2 show a histogram of the rate of gender-based crimes including rape and kidnapping.[6] Both figures look similar, which indicate that the rate of gender based crime such as rape and kidnapping occurs in similar proportion through out all districts. Furthermore, the correlation between both crimes is high at 0.66.

3.4 Methodology

In this section, we present the ordinary least squares model (OLS) as the primary method to estimate the impact of consanguineous marriage on pregnancy outcomes. To support our work, we implement a standard instrumental variable approach.

[5]On average, the population by district is around 900,000

[6]The IV works as a proxy for the ambient level of violence against young women in a district such that it covers the years 2011 and 2012.

3.4.1 Ordinary Least Squares (OLS)

We utilize the OLS to investigate the relation between a young female that marries a family member and the experience of birth problems.

$$Y_{is} = \beta_0 + \beta_1 BloodRelative + \sum_{k=2}^{K} \beta_k X_{k,is} + \gamma_s + e_{is} \tag{1}$$

Y_{is} is a binary dependent variable indicating whether a woman i from a state s experiencing mortality at birth or during gestation including stillbirth, miscarriage and child death after birth. Since the dependent variable is binary, the OLS method constitutes a linear probability model. $BloodRelative$ is the binary explanatory variable of interest representing whether a woman is related to her husband by blood, and β_1 is the coefficient of interest. Because of data limitation, we do not differentiate whether the couples are first or second cousin. Although it is believed that the closer the blood relation between parents, the higher the risk of offspring health problems (Zlotogora, 1991), it still does not falsify our empirical strategy. $X_{k,is}$ represents a set of socioeconomic and demographic factors discussed in the data section, γ_s is a time invariant state fixed effect and e_{is} is the error term.

Our model is an ordinary least squares model with a dummy outcome variable. Thus, it constitutes a linear probability model (LPM), which is easier to interpret as well as does not suffer from the convergence failures; however, this model is associated with some challenges, including heteroscedastic errors that might lead to inconsistent standard estimates, inaccurate P-values as a consequence of non-normality of the outcome variable, and nonsensical predicted probabilities that do not range between zero and one. In our case, the OLS model is still valid and not affected by these aforementioned disadvantages. To avoid any statistical power issue, we work with a large sample size, and to fix any heteroscedasticity issue, we use robust standard errors and clustered standard error at the town/ or village level.

The exogeneity assumption is necessary to hold here to have unbiased OLS coefficient estimate. The assumption requires all regressors in the model to be uncorrelated with the error term when controlling the other factors, and thus the parameters, $\beta's$, represent the marginal effects. The standard conditional mean independence assumption is crucial to have the OLS unbiased and constant estimator for the coefficient of main variable of interest, β_1.[7] We assume that the conditional expectation of e_{is} does not depend on $BloodRelative$ if control for $X_{k,is}$, where k=2, 3, 4, ..., K. Indeed, our main explanatory variable, a woman related to her husband by blood, is affected by financial and family controls. Those factors impact upon the health outcome variable as well. Therefore, due to the purpose of our study it is of a less concern than for other studies as the setting of our data allows us to potentially control for these factors effectively. However, thinking of the potential existence of some unobservable factors that may still affect our main explanatory variable and the outcome may constitute a threat to the internal validity of our results. For example, malnutrition variable is not available in our data, but the dataset allows us to use a proxy for that such as including a continuous variable represents a young woman's height or a dummy variable represents a young woman's health status. Indeed, around 77% of women in our sample

[7] $E(e_{is}|BloodRelative, X_{k,is}) = E((e_{is}|X_{k,is})$

reported being in good health. To support the OLS results, we implement the standard 2SLS approach as it is known for the ability to deal with the endogeneity issue that might come from potential omitted variable bias.

3.4.2 Two-Stage least squares (2SLS)

Since the intra-family marriage variable (*BloodRelative*) by itself is not exogenous, we implement the 2SLS model to support our results from the OLS model. We extract the part of variation that is not correlated with the error term by utilizing an exogenous shock coming from the yearly crime rate against women per 100,000 population in a district.[8] By assumption, this instrumental variable, the ambient level of violence against women in the district that an observed young woman lives, is uncorrelated with the error term. Medical literature sheds some light on the link between prenatal stress and adverse birth outcomes (Coussons-Read, 2013). Previous studies show that prenatal stress impact is significant when it comes from big trauma such as earthquakes and catastrophic accidents; however, in our study it may be less of a concern. We control for a young woman health status such that around 77% of our sample reported being in good health. As well, the IV represents the ambient level of violence rather than crimes at personal level. Previous studies show that violence rates against women in neighbourhood area increase the likelihood of young female marriages, which will be discussed in details in the sub-section of instrument relevance. Our two-stage least squares model is shown in the following system of two linear regression equations.

[8]This covers the yearly total incidences of rape and kidnapping for 2011 and 2012 as our observations spread over these two years.

$$BloodRelative = \beta_1 Z_d + \sum_{k=2}^{K} \beta_k X_{k,is} + \gamma_s + u_{is} \qquad (2)$$

$$Y_{is} = \alpha_1 + \alpha_2 \widehat{BloodRelative} + \sum_{k=2}^{K} \alpha_k X_{k,is} + \gamma_s + \epsilon_{is} \qquad (3)$$

Equation 2 represents the first stage of the 2SLS. The endogenous variable, $BloodRelative$, is regressed on the instrument, Z_d, the set of covariates, $X_{k,is}$, state fixed effect, γ_s, and the error term, u_{is}. Equation 3 represents the second stage of the 2SLS. Y_{is} is regressed on the predicted value $\widehat{BloodRelative}$ and the same set of covariates from equation 2. ϵ_{is} represents error term. Y_{is} represents our binary outcome variable, that is, whether a woman experienced perinatal loss or child death after birth. $\widehat{BloodRelative}$ is the predicted value of our endogenous variable generated from running the first stage. $X_{k,is}$ is the same set of covariates including individual and household controls the same as equation 1. To correct for heteroskedasticity, robust standard errors and cluster standard errors at the town/or village level are used. Despite having the treatment and outcome as binary variables, the linear IV model still provides a consistent estimate even under non-normality (Angrist and Pischke, 2008). The underlying assumption for the validity of our instrument is that birth problems are not affected directly by ambient level of violence against women in the region, but only through its impact on consanguineous marriage.

Instrument validity

In this sub-section, we discuss the relevance and exclusion restriction of the instrumental variable.

Instrument relevance

The relevance assumption means that the endogenous variable is significantly correlated with the instrumental variable. In our case, $Cov(Z_d, BloodRelative_{is}) \neq 0$. Many studies have demonstrated that the ambient level of violence against women in the area where they reside affects young women and their families behavior (Sarkar, 2021; Mourtada et al., 2017; Chakraborty et al., 2018; Kohno et al., 2020). For example, living in such an environment may pressure women into marrying faster or marrying a family member, or reducing total number of completed schooling years.

Concerning marriage, parents live in areas with high level of violence experience added pressure in regards to marrying off their daughters because of threats to traditional culture high values placed on a woman's virginity. Within these beliefs, loss of virginity and pre-marriage pregnancy would be considered as having a significant negative impact on a young female future value in the marriage market.[9] In this case, intra-family marriage becomes a better alternative as the ambient level of violence rises. As such, it must be noted that consanguineous marriages are usually easier to set up and cheaper

[9]Marriage in India is complicated with matching process that consider social aspects such as caste, tradition, religion, customs, et cetera. In addition, it is associated with financial costs paid by both bride and groom's families. The cost is usually divided between the two families based on negotiation. As well, the bride's family provides a dowry to groom's family such that the amount is based on negotiation built on quality of the bride (Chowdhury, 2010)

with less concern of dowry transfer (Do et al., 2013).

Instrument exogeneity

The exclusion restriction assumption is that our instrumental variable, the ambient level of violence, affects perinatal and child death only through a woman's propensity to marry within a family. This second assumption requires that the IV be uncorrelated with the error term; however, it is mathematically unverifiable. In practice, there are few potential challenges can give rise to a nonzero correlation between the IV and the error term. These challenges are that the ambient level of violence may affect health outcomes through young woman behaviors including early marriage, or reducing education level. Therefore, to verify the validity of this assumption, we will discuss these challenges with some explanations.

First, in terms of the early marriage, studies show that elevated level of crimes against women increases the probability of early marriage, which could affect offspring health negatively. Therefore, to role out this concern of potential channel adverse impact of child marriage on offspring health we control for age at time of marriage. In another two exercises, we restrict a sub-sample to include only young women married at eighteen years old and above to eliminate any concern in terms of child marriage. Then we restrict a sub-sample to exclude the states that are known to display the highest rates of child marriages, and this will add a justification that the impact is not driven by these states.[10] Second, in terms of education, woman health behaviour such

[10]Published statistics shows that the majority of bride child marriages are concentrated in five states: Uttar Pradesh, Bihar, West Bengal, Maharashtra and Madhya Pradesh with concentration of 36, 22, 22, 20 and 16 million bride child marriages, respectively (UNICEF, New York, 2019). For more information the report can be found in the following link https://www.unicef.org/india/media/1176/file/Ending-Child-Marriage.pdf

as accessing health services and attending regular check ups can be improved by increasing the education level, which eventually positively influences the child's health. However, since violence against women in a region may reduce a woman's total number of completed schooling years, we control for a woman's level of education. Then as a sub-sample, we restrict our sample to include only illiterate women with no education to role out the potential channel through reducing number of completed schooling years.

Our instrumental variable, the ambient level of violence, includes only the rape and kidnapping, and excludes any unwilling sexual contact between a woman and her intimate partner.[11] The instrumental variable does not violate the relevance and exclusion restriction assumptions, which will be confirmed in the results section below.

3.5 Results

3.5.1 OLS results

Table 3.2 represents the results from the OLS regression. It shows that on average consanguineous marriage increases the risk of women experiencing perinatal and child death significantly by 2.8 percentage points. The OLS estimate seems not to underestimate the impact of consanguinity on pregnancy outcomes. Our results are consistent with previous work that assures the adverse impact of consanguineous marriage on offspring health.

[11] Based on section 375 of the India Panel Code, a woman is presumed to have a sex with her husband when she enters the marital relation. Thus, India is one of the countries that does not criminalizes unwilling sex contact between couples.

The effect sign of the other controls in the model validates our research findings. For example, household income and a woman's education level have a negative sign, which means a higher household income and education level would decrease the adverse effect on offspring health. Higher income level means more available food, better nutrition and greater ability to obtain access to higher quality health care. Also, it must be noted that completing a higher level of education enables women to increase their awareness around general healthy practices and behaviours, which may reduce the risk of adverse effect on offspring health.

Heterogeneity by religion

Diversity in religions can be seen in India. The country is a place of numerous religions including Hinduism, Islam, Christianity, Sikhism, and Buddhism. These five major religions in the country constitute 83.7, 11.5, 1.9, 1.3, and 0.8% of the population, respectively. This diversity can lead to different religion practices including restrictions on food, early childbearing and wanting more children. Indeed, some previous studies shed some light on the mixed impact of parental religiosity on fertility and child development (Morgan and Hayford, 2008; Glazier et al., 2018); however, in this sub-section of our study, we assume type of religion might drive heterogeneous impact on offspring health. Thus, in Table 3.3, we look at the impact by religion. The coefficient estimates show no significant statistical difference among the impacts by religion type.

3.5.2 2SLS results

Table 3.4 shows the results we obtain in the first stage. Following the impact of the ambient level of violence against women on consanguineous marriage shows that an increase in crime rates per 100,000 population in a district significantly increases the probability of a young woman marrying a blood relative by 2 percentage points. The first stage F-statistics value is 18.5, which is greater than 10 as a threshold suggested by researchers (e.g., Stock and Yogo, 2005; Staiger and Stock, 1997). This indicates the absence of a weak instrument and the maximum bias in the IV estimator is less than 10%.

Table 3.5 shows the results from the second stage estimation of our 2SLS estimator. It shows that on average consanguineous marriage increases the risk of women experiencing perinatal and child death significantly by 2.2 percentage points. This result is not statistically different from the coefficient estimate of the OLS. This assures the reliability of the OLS estimate and eliminate the concern of omitted variable bias and measurement error as threat to internal validity.

Table 3.6 shows the results for different sub-samples to discuss the robustness of the analysis as explained in the methodology. In column 2 and 3, we show the results of eliminating the child marriage concern. In column 2, when restricting the sample to all eligible women married at age eighteen years and above, we find a significant impact of 2.1% percentage points. The impact in this sub-sample of married women brings no statistical difference from the impact obtained by using the full sample. This exercise shows that the impact of ambient violence against women is affecting through our claimed channel of consanguinity. In column 3, when restricting the sample to all eligible women living outside the five well known states with the highest proportion of child marriages, we find a significant impact of 2.3% percentage points. Our results from these sub-samples in column 2 and 3 show early marriage does not matter much. As such, what matters the most is inheriting recessive genes due to the consanguineous marriage issue. In column 4, when restricting the sample to uneducated women, we find a significant impact of 1.8% percentage points. This result shows that even for uneducated women, where ambient level of violence against women does not affect their level of education, ambient level of violence still adversely affect pregnancy outcomes. This result confirms the significance of consanguineous marriage as the main channel that ambient level of violence affects offspring health through. The ambient level of violence affects young women's behaviors by increasing their chances of marrying a family member, which adversely affect pregnancy outcomes.

3.6 Conclusion

Our study focuses on the impact of consanguineous marriages on pregnancy outcomes. In order to identify such impact, we use the IHDS-II which provides data about young women in India, one of countries with the highest rates of consanguineous marriages.

Previous studies have proved inconclusive when evaluating the impact of consanguineous marriage on offspring health. This study fills a gap in existing literature by making three key contributions: (1) it utilizes a large, good quality and representative sample of young Indian women. (2) it uses a woman experiencing stillbirth, miscarriage, and child death after birth as an outcome; this might be easier to investigate than to assess the impact on health once adulthood has been reached, where more non-genetic confounding factors should be considered. (3) it utilizes the ordinary least squares model controlling for a wide range of financial and family factors. The model is supported by applying the instrumental variable method, where the instrument represents the ambient level of violence against women in the region where they reside.

The results of the ordinary least squares model show that being a young woman related to her husband by blood leads to an increase in the probability of experiencing perinatal and child death significantly by 2.8 percentage points. These results confirm the adverse impact of consanguineous marriage on offspring health. This conclusion is supported by the 2SLS approach that leads to 2.2 percentage points. Finally, although some previous studies show the importance of parental religiosity on fertility and child health, we find no heterogeneous impact across religion.

Our analysis and findings provide key evidence that could support policy-makers in implementing effective schemes to reduce the negative impact of consanguineous marriages on health outcomes. However, a cost and benefit analysis should be undertaken to estimate the value of consanguineous marriage. In order to ensure more rigorous data in further studies, we suggest expanding the health outcome variable to include a wide range of pregnancy and non-pregnancy health outcomes. Such data, in combination with using the inbreeding coefficient to determine the level of consanguinity of married couples, constitute an interesting avenue for further research.

References

Agha, M. M., Williams, J. I., Marrett, L., To, T., & Dodds, L. (2006). Determinants of survival in children with congenital abnormalities: a long-term population-based cohort study. Birth defects research. Part A, *Clinical and molecular teratology*, 76(1), 46–54. https://doi.org/10.1002/bdra.20218

Akrami, S. M., & Osati, Z. (2007). Is consanguineous marriage religiously encouraged? Islamic and Iranian considerations. *Journal of Biosocial Science*, 39(2), 313–316. https://doi.org/10.1017/S0021932006001684

Akyol, Ş. P., & Mocan, N. (2020). Education and consanguineous marriage (No. w28212; p. w28212). *National Bureau of Economic Research*. https://doi.org/10.3386/w28212

Al-Awadi, S. A., Naguib, K. K., Moussa, M. A., Farag, T. I., Teebi, A. S., & El-Khalifa, M. Y. (2008). The effect of consanguineous marriages on reproductive wastage. *Clinical Genetics*, 29(5), 384–388. https://doi.org/10.1111/j.1399-0004.1986.tb00509.x

Al-Gazali, L. (1998). A genetic aetiological survey of severe childhood deafness in the United Arab Emirates. *Journal of Tropical Pediatrics*, 44(3), 157–160. https://doi.org/10.1093/tropej/44.3.157

Al-Rahal, N. K. (2018). Inherited bleeding disorders in Iraq and consanguineous marriage. *International Journal of Hematology-Oncology and Stem Cell Research*,12(4), 273-281. www.ncbi.nlm.nih.gov/pmc/articles/PMC6375379/

Angrist J., & Pischke, J. (2009). Mostly harmless econometrics: an empiricist's companion. *Princeton University Press, Princeton*

Bachir, S., & Aouar, A. (2019). Study of the impact of consanguinity on abortion and mortality in the population of Beni Abbes (southwestern Algeria). *Egyptian Journal of Medical Human Genetics*, 20(1), 1. https://doi.org/10.1186/s43042-019-0004-7

Bener, A., Hussain, R., & Teebi, A. S. (2007). Consanguineous marriages and their effects on common adult diseases: studies from an endogamous population. Medical principles and practice : *International Journal of The Kuwait University, Health Science Centre*, 16(4), 262–267. https://doi.org/10.1159/000102147

Bittles A. (2001). Consanguinity and its relevance to clinical genetics. *Clinical Genetics*, 60(2), 89–98. https://doi.org/10.1034/j.1399-0004.2001.600201.x

Bittles, A. H., & Black, M. L. (2010a). Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences*, 107(suppl_1), 1779–1786. https://doi.org/10.1073/pnas.0906079106

Bittles, A. H., & Black, M. L. (2010b). Consanguineous marriage and human evolution. *Annual Review of Anthropology*, 39(1), 193–207. https://doi.org/10.1146/annurev.anthro.012809.105051

Bittles, A. H., & Black, M. L. (2010c). The impact of consanguinity on neonatal and infant health. *Early Human Development*, 86(11), 737–741. https://doi.org/10.1016/j.earlhumdev.2010.08.003

Chakraborty, T., Mukherjee, A., Rachapalli, S. R., & Saha, S. (2018). Stigma of sexual violence and women's decision to work. *World Development*, 103: 226–38. https://doi.org/10.1016/j.worlddev.2017.10.031

Chowdhury, A. R. (2010). Money and marriage: the practice of dowry and brideprice in rural India. https://paa2010.princeton.edu/papers/100225

Coussons-Read, M. E. (2013). Effects of prenatal stress on pregnancy and human development: mechanisms and pathways. Obstetric medicine, 6(2), 52–57. https://doi.org/10.1177/1753495X12473751

Denic, S., & Bener, A. (2001). Consanguinity decreases risk of breast cancer – cervical cancer unaffected. *British Journal of Cancer*, 85(11), 1675–1679. https://doi.org/10.1054/bjoc.2001.2131

Devi, A. R., Rao, N. A., & Bittles, A. H. (1987). Inbreeding and the incidence of childhood genetic disorders in Karnataka, South India. *Journal of Medical Genetics*, 24(6), 362–365. https://doi.org/10.1136/jmg.24.6.362

Do, Q. T., Iyer, S., & Joshi, S. (2013). The economics of consanguineous marriages. *Review of Economics and Statistics*, 95(3), 904-918. https://doi.org/10.1162/REST_a_00279

Glazier, J. D., Hayes, D. J., Hussain, S., D'Souza, S. W., Whitcombe, J., Heazell, A. E., & Ashton, N. (2018). The effect of Ramadan fasting during pregnancy on perinatal outcomes: a systematic review and meta-analysis. *BMC Pregnancy and Childbirth*, 18(1), 1-11. https://doi.org/10.1186/s12884-018-2048-y

Hamamy, H. (2012). Consanguineous marriages: preconception consultation in primary health care settings. *Journal of Community Genetics*, 3(3), 185–192. https://doi.org/10.1007/s12687-011-0072-y

Hamamy, H., Antonarakis, S. E., Cavalli-Sforza, L. L., Temtamy, S., Romeo, G., Kate, L. P. T., Bennett, R. L., Shaw, A., Megarbane, A., van Duijn, C., Bathija, H., Fokstuen, S., Engel, E., Zlotogora, J., Dermitzakis, E., Bottani, A., Dahoun, S., Morris, M. A., Arsenault, S., … Bittles, A. H. (2011). Consanguineous marriages, pearls and perils: *Geneva International Consanguinity Workshop Report: Genetics in Medicine*, 13(9), 841–847. https://doi.org/10.1097/GIM.0b013e318217477f

Hayford, S. R., & Morgan, S. P. (2008). Religiosity and fertility in the United States: The role of fertility intentions. *Social Forces*, 86(3), 1163–1188. https://doi.org/10.1353/sof.0.0000

Husain, M. A., & Bunyan, M. A. (1997). Consanguineous marriages in a Saudi population and the effect of inbreeding on prenatal and postnatal mortality. *Annals of Tropical Paediatrics*, 17(2), 155–160. https://doi.org/10.1080/02724936.1997.11747879

Ismail, J., Jafar, T. H., Jafary, F. H., White, F., Faruqui, A. M., & Chaturvedi, N. (2004). Risk factors for non-fatal myocardial infarction in young South Asian adults. *Heart*, 90(3), 259–263.

https://doi.org/10.1136/hrt.2003.013631

Jaber, L., Merlob, P., Gabriel, R., & Shohat, M. (1997). Effects of consanguineous marriage on reproductive outcome in an Arab community in Israel. *Journal of Medical Genetics*, 34(12), 1000–1002. https://doi.org/10.1136/jmg.34.12.1000

Jaber, L., Shohat, T., Rotter, J. I., & Shohat, M. (1997). Consanguinity and common adult diseases in Israeli Arab communities. *American Journal of Medical Genetics*, 70(4), 346–348. https://doi.org/10.1002/(SICI)1096-8628(19970627)70:4<346::AID-AJMG2>3.0.CO;2-R

Khlat, M. (1989). Inbreeding effects on fetal growth in Beirut, Lebanon. *American Journal of Physical Anthropology*, 80(4), 481–484. https://doi.org/10.1002/ajpa.1330800407

Kohno, A., Techasrivichien, T., Suguimoto, S. P., Dahlui, M., Nik Farid, N. D., & Nakayama, T. (2020). Investigation of the key factors that influence the girls to enter into child marriage: A meta-synthesis of qualitative evidence. *PLOS ONE*, 15(7), e0235959. https://doi.org/10.1371/journal.pone.0235959

Kulkarni, M. L., & Kurian, M. (1990). Consanguinity and its effect on fetal growth and development: A south Indian study. *Journal of Medical Genetics*, 27(6), 348–352. https://doi.org/10.1136/jmg.27.6.348

Liede, A., Malik, I. A., Aziz, Z., Rios, P. de los, Kwan, E., & Narod, S. A. (2002). Contribution of BRCA1 and BRCA2 mutations to breast and ovarian cancer in Pakistan. *The American Journal of Human Genetics*, 71(3), 595–606. https://doi.org/10.1086/342506

Mete, C., Bossavie, L., Giles, J., & Alderman, H. (2019). Is consanguinity an impediment to child development outcomes? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3468612

Mourtada, R., Schlecht, J., DeJong, J. (2017). A qualitative study exploring child marriage practices among Syrian conflict-affected populations in Lebanon. *Conflict and Health*, 11(S1), 27. https://doi.org/10.1186/s13031-017-0131-z

Naibkhil, N., & Chitkara, E. (2016). Consanguineous marriages increase risk of congenital anomalies-studies in four generation of an Afghan family. *Biomed Res*, 27(1), 6.

Nandraj, S., Dilip, T. (2017). National Health Accounts Estimates. The more things change, the more they remain the same. *Economic and*

Political Weekly, 52(23), 20-23.

Obeidat, B. R., Khader, Y. S., Amarin, Z. O., Kassawneh, M., & Al Omari, M. (2010). Consanguinity and adverse pregnancy outcomes: the North of Jordan experience. *Maternal and Child Health Journal*, 14(2), 283–289. https://doi.org/10.1007/s10995-008-0426-1

Paddaiah, G., & Reddy, G. G. (1980). Effect of consanguinity on anthropometric measurements in the newborn. *The Indian Journal of Pediatrics*, 47(2), 133–136. https://doi.org/10.1007/BF02822878

Patki, A., & Chauhan, N. (2016). An epidemiology study to determine the prevalence and risk factors associated with recurrent spontaneous miscarriage in India. *Journal of Obstetrics and Gynaecology of India*, 66(5), 310–315. https://doi.org/10.1007/s13224-015-0682-0

Paul, D. B., & Spencer, H. G. (2008). "It's ok, we're not cousins by blood": the cousin marriage controversy in historical perspective. *PLoS Biology*, 6(12), e320. https://doi.org/10.1371/journal.pbio.0060320

Plagge, J., Antick, J. R., & Armstrong, J. (2007). Perceptions of perinatal loss: miscarriage versus stillbirth: (644012007-001). *American Psychological Association*. https://doi.org/10.1037/e644012007-001

Rao, P. S. S., & Inbaraj, S. G. (1980). Inbreeding effects on fetal growth and development. *Journal of Medical Genetics*, 17(1), 27–33. https://doi.org/10.1136/jmg.17.1.27

Rosano, A. (2000). Infant mortality and congenital anomalies from 1950 to 1994: An international perspective. *Journal of Epidemiology & Community Health*, 54(9), 660–666. https://doi.org/10.1136/jech.54.9.660

Saedi-Wong, S., & Al-Frayh, A. R. (1989). Effects of consanguineous matings on anthropometric measurements of Saudi Newborn infants. *Family Practice*, 6(3), 217–220. https://doi.org/10.1093/fampra/6.3.217

Sarkar, S. (2021). Local crime and early marriage: evidence from India (WIDER Working Paper No. 2021; WIDER Working Paper, Vol. 2021). *UNU-WIDER*. https://doi.org/10.35188/UNU-WIDER/2021/975-4

Shami, S. A., Schmitt, L. H., & Bittles, A. H. (1990). Consanguinity, spousal age at marriage and fertility in seven Pakistani Punjab cities. *Annals of Human Biology*, 17(2), 97–105. https://doi.org/10.1080/03014469000000842

Shawky, R. M., El-Awady, M. Y., Elsayed, S. M., & Hamadan, G. E. (2011). Consanguineous matings among Egyptian population. *Egyptian Journal of Medical Human Genetics*, 12(2), 157–163. `https://doi.org/10.1016/j.ejmhg.2011.07.001`

Sonawat, R. (2001). Understanding families in India: a reflection of societal changes. *Psicologia: Teoria e Pesquisa*, 17(2), 177–186. `https://doi.org/10.1590/S0102-37722001000200010`

Stern, C. M. (1979). Intrapartum sensitisation of human infants to the Rhesus-D antigen. *Archives of Disease in Childhood*, 54(4), 315–317. `https://doi.org/10.1136/adc.54.4.315`

Tadmouri, G. O., Nair, P., Obeid, T., Al Ali, M. T., Al Khaja, N., & Hamamy, H. A. (2009). Consanguinity and reproductive health among Arabs. *Reproductive Health*, 6(1), 17. `https://doi.org/10.1186/1742-4755-6-17`

United Nations Children's Fund (2019). Ending child marriage: a profile of child marriage in India. UNICEF, New York.

Verma, I. C., & Bijarnia, S. (2002). The burden of genetic disorders in India and a framework for community control. *Community Genetics*, 5(3), 192–196. https://doi.org/10.1159/000066335

Zlotogora J. (1991). Problems in diagnosis and delineation of inherited disorders in highly inbred populations. *American Journal of Medical Genetics*, 41(4), 451–453. https://doi.org/10.1002/ajmg.1320410414

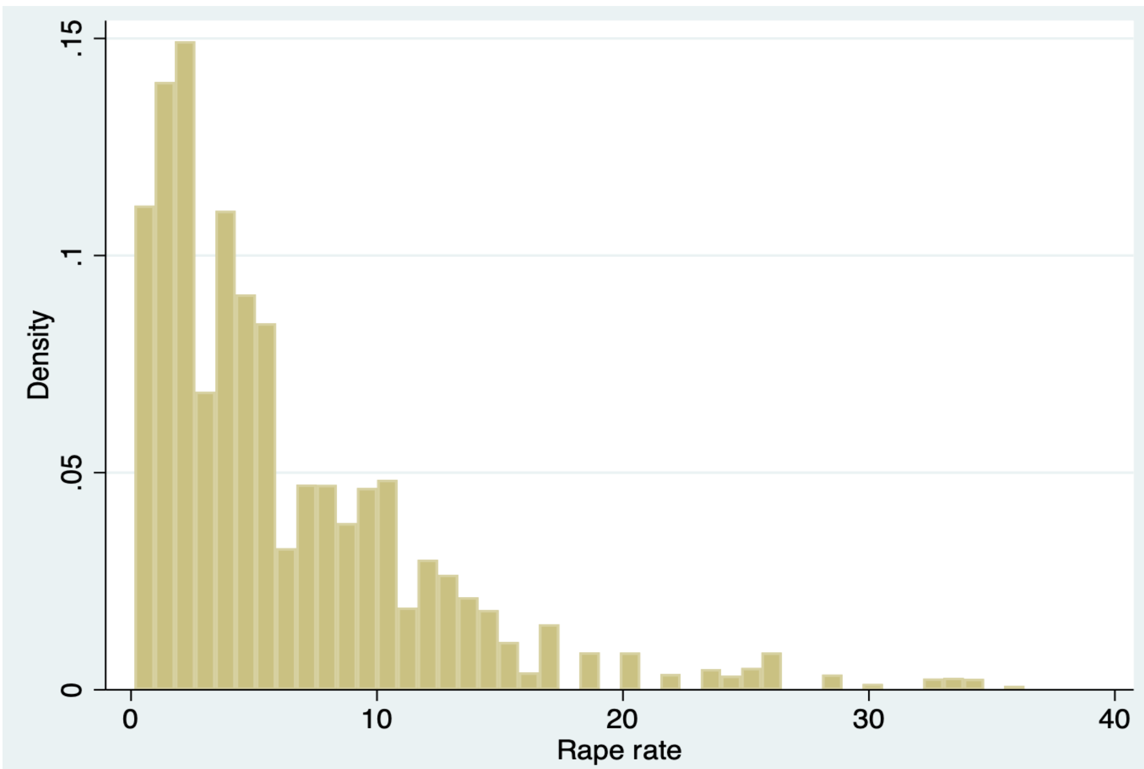

Figure 3.1: Distribution of rape rate against women (per 100K population in a year in a district)

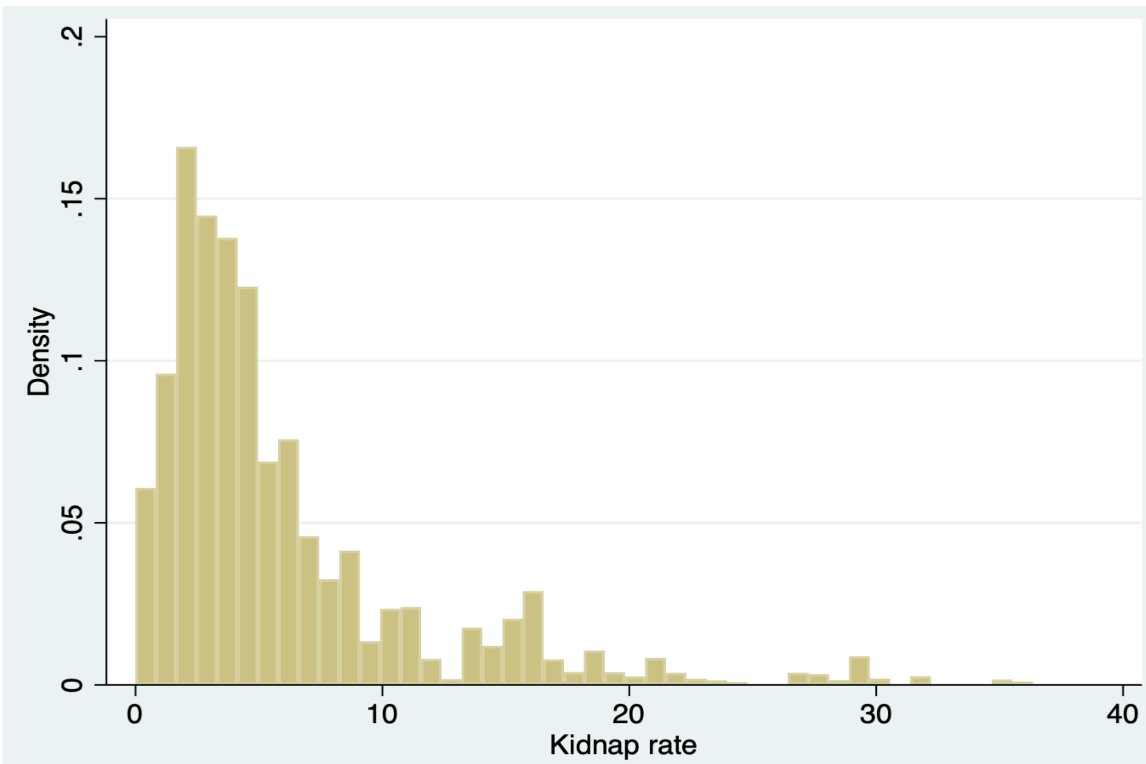

Figure 3.2: Distribution of kidnapping rate against women (per 100K population in a year in a district)

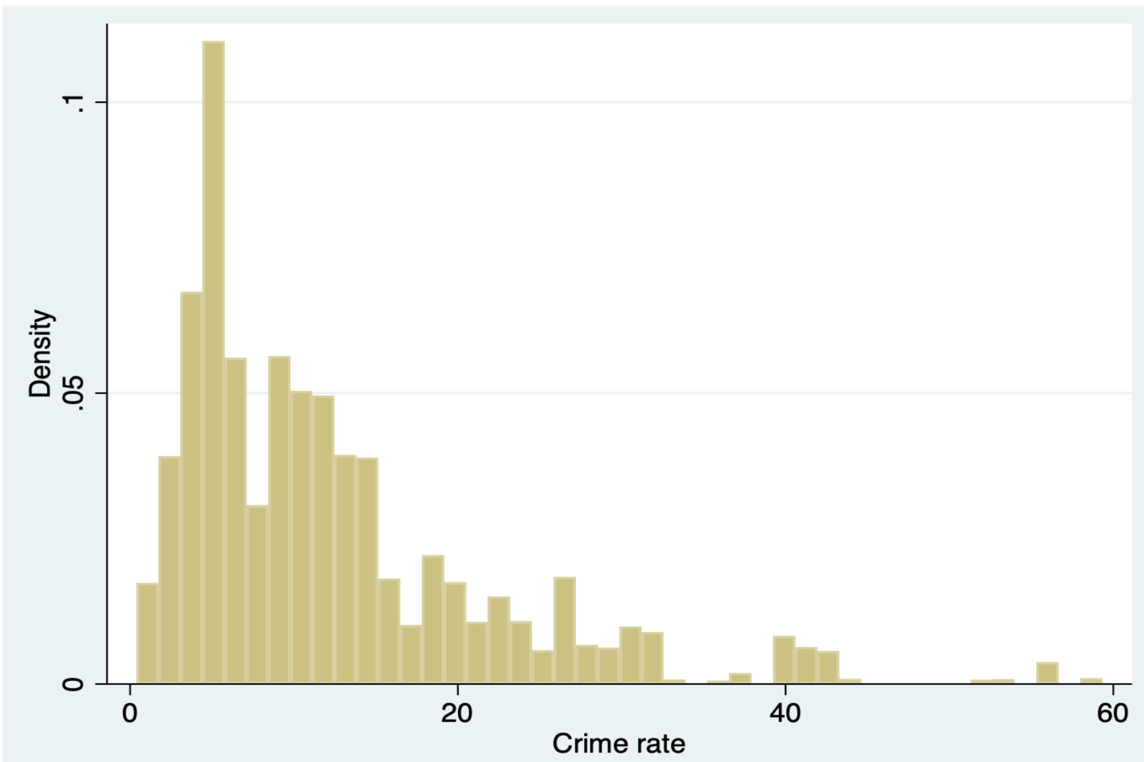

Figure 3.3: Distribution of the ambient level of violence against women (per 100K population in a year in a district)

Table 3.1: Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
A woman experienced birth problems	25,405	0.312	0.464	0	1
A woman is a blood relative with her husband	25,405	0.081	0.273	0	1
Kidnapping rate towards women	25,405	8.516	7.839	0	100
Rape rate towards women	25,405	7.569	7.99	0.14	67.97
Ambient level of violence towards women	25,405	15.08	15.76	0.40	129.7
Woman education (year)	25,405	5.054	4.925	0	16
Woman height (cm)	25,405	152.23	7.937	85	197
Age first started menarche (year)	25,405	13.90	1.469	9	28
Age at marriage (year)	25,405	18.02	3.476	10	45
Monthly income (rupees)	25,405	11368	18442	0	946,666
Health status	25,405	0.769	0.421	0	1
Highest male education in household (year)	25,405	8.007	4.919	0	16
Highest female education in household (year)	25,405	5.936	5.237	0	16
Father literate	25,405	0.425	0.494	0	1
Mother literate	25,405	0.228	0.419	0	1
Urban/ rural area	25,405	0.328	0.469	0	1

Note: A woman experienced birth problems is a dummy variable represents whether a woman experienced perinatal and child death after birth. The ambient level of violence towards women is the instrumental variable. Kidnapping rate represents the yearly rate per 100,000 population in a district. Rape rate represents the yearly rate per 100,000 population in a district. Woman's height is measured in centimeters. Self-reported health status is a dummy variable represents whether a woman in good health. Household average monthly income is measured in Indian rupees. Father literate is a dummy variable. Mother literate is a dummy variable. Urban/rural is a dummy variable represents whether a woman lives in urban area.

Table 3.2: Ordinary least squares (OLS)

	(1)	(2) Full model with robust S.E.s	(3) Full model with clustered S.E.s
Pregnancy outcomes			
Blood relative with husband	0.03***	0.028***	0.028***
	(0.00991)	(0.00989)	(0.01011)
Individual controls		√	√
Household controls		√	√
State FE	√	√	√
Observations	25,405	25,405	25,405

Note: In Column 2, we have the robust standard errors in parentheses. Column 3 has the standard errors clustered at the town/ or village level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 3.3: Heterogeneity by religion using the OLS

	(1) Full sample	(2) Hinduism	(3) Islam	(4) Christianity	(5) Sikh
Pregnancy outcomes					
Blood relative with husband	0.028***	0.0277**	0.0362	0.0382	0.00852
	(0.01011)	(0.0110)	(0.02043)	(0.02701)	(0.0134)
Individual controls	√	√	√	√	√
Household controls	√	√	√	√	√
State FE	√	√	√	√	√
Observations	25,405	21,058	2,884	563	601

Note: Column 1 represents the full model. Columns 2 to 5 represent results of regressions by religion type. Clustered standard errors at the town/ or village level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 3.4: Impact of violence on intra-family marriage (IV- first stage)

	(1)	(2) Full
Blood relative with husband		
Ambient level of violence towards women	0.022***	0.025***
	(0.00578)	(0.00583)
Individual controls		√
Household controls		√
State FE	√	√
Observations	25,405	25,405

Note: This table represents the first stage results of the 2SLS approach using linear probability model. F-statistic is 18.53.

Table 3.5: IV-Second stage. Linear probability model estimation

	(1) Robust S.E.s	(2) Clustered S.E.s
Pregnancy outcomes		
Blood relative with husband	0.0217***	0.0218**
	(0.007956)	(0.010488)
Individual controls	√	√
Household controls	√	√
State FE	√	√
Observations	25,405	25,405

Note: This table shows the second stage results, the impact of intra-family marriage on reproductive outcomes, using the IV-linear probability model. Column 1 has the robust standard errors in parentheses. Column 2 has the standard errors clustered at the town/ or village level in parentheses. * p<0.1, ** p<0.05, *** p<0.01. F-statistics is 18.53. Based on the 2SLS size distortion of associated 5% Wald test, the minimum eigenvalue statistic exceeds the critical value (16.38). The maximum bias in our instrument estimator is less than 10%. Thus, we can reject the null hypothesis that the instrument is weak (e.g., Stock and Yogo, 2005; Staiger and Stock, 1997).

Table 3.6: Robustness check

	Child marriage			
	(1)	(2)	(3)	(4)
	Full sample	Marriage at age 18+	Excluding 5 states	Uneducated women
Pregnancy outcomes				
Blood relative with husband	0.0218**	0.0215**	0.023**	0.0182**
	(0.010488)	(0.00876)	(0.00951)	(0.00902)
Individual controls	√	√	√	√
Household controls	√	√	√	√
State FE	√	√	√	√
Observations	25,405	9,391	17,303	10,317

Note: Column 1 represents the full model. Columns 2 and 3 represents sub-samples to validate the exclusion restriction from the child marriage concern. Column 2 represents a sub-sample of all women married at age 18 years and older. Column 3 represents a sub-sample of all women married from all states except the five well known states of high prevalence of child marriages. Column 4 represents a sub-sample of illiterate uneducated women. Clustered standard errors at the town/ or village level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.