The Legacy of the Taliban on Female Education in Pakistan

IDS 701: Final Project Report

Clarissa Ache, Preet Khowaja, John Owusu Duah, & Rashaad Ratliff-Brown

April 5, 2022

Summary

With a clear objective of overthrowing the federal government of Pakistan and establishing a strict shariah law in the country, the Taliban has made significant strides in capturing control of large areas of Pakistan since its founding in the early 1990's. Nested within this objective is the rejection of western ideologies, of which education of girls is a key component. This study seeks to investigate the impact that the Taliban has had on girls' education in rural areas.

To achieve this objective, we compare girls who are in 20 Taliban controlled villages to girls who are in villages which are not controlled by the Taliban. In order to measure the effect of the Taliban influence in women's access to education in Pakistan, we first define which areas of the country have been under Taliban control. After the terrorist movement was launched in 2007, leaders of the TTP and sympathising militants settled in what was known as the North-West Frontier Province (NWFP) of Pakistan, controlling 20 tribal regions and districts: South Waziristan, North Waziristan, Orakzai, Kurram, Khyber, Mohmand, Bajaur, and Darra Adamkhel, and districts of Swat, Upper Dir, Lower Dir, Bannu, Lakki Marwat, Tank, Peshawar, Dera Ismail Khan, Mardan, Charsadda, and Kohat.

The data set has been sourced from Pakistan Living Standards Measurement (PSLM) survey. The survey covers rural and urban areas of the four main provinces of Pakistan and collects information on a wide range of topics such as education, diarrhoea, immunization, reproductive health, pregnancy, history, family planning, pre and post-natal care, and access to basic services. Using a difference in difference approach, we estimate the causal treatment effect on the treated group (ATT) by comparing the girl's enrollment rate in the two groups both before and after assuming they would trend similarly over time if Taliban intervention had not occurred.

A thorough discussion of the challenges faced using the PSLM Survey has been outlined in the report. Some of these limitation include; lack of survey data for 2009-10, 2016-17 and 2017-18, collection of districts surveyed is not the same each year, random subsets of the population are surveyed each year. In conclusion, we did not find a significant average effect in the enrollment rate in school for girls in cities that were controlled by the Taliban.

Introduction

The Taliban, a Sunni Islamist nationalist group that was founded in the early 1990's, was able to gain control parts of Afghanistan and Pakistan over the years as part of their religious-political movement. The Tehrik-e-Taliban (TTP) is the political party of Pakistan that has alliances to the militant factions of the Taliban that stated its objective as the overthrow of the federal government in Pakistan and implementation of strict shariah law in the country. Part of their orthodox agenda is preventing girls from attending educational institutions, particularly schools that follow Western models of education. In 2012, they attempted to assassinate education rights activist, Malala Yousufzai, who went on to become a global champion for female education and youngest Nobel Peace Prize laureate. Today, Malala's Fund states the importance of girls education as:

> Without education, girls are more likely to marry young, suffer from preventable diseases and experience poverty throughout their lives. Without educated girls, communities, countries and our world suffer from slower economic growth, more conflict, poor public health and increased risk from the effects of climate change. [5]

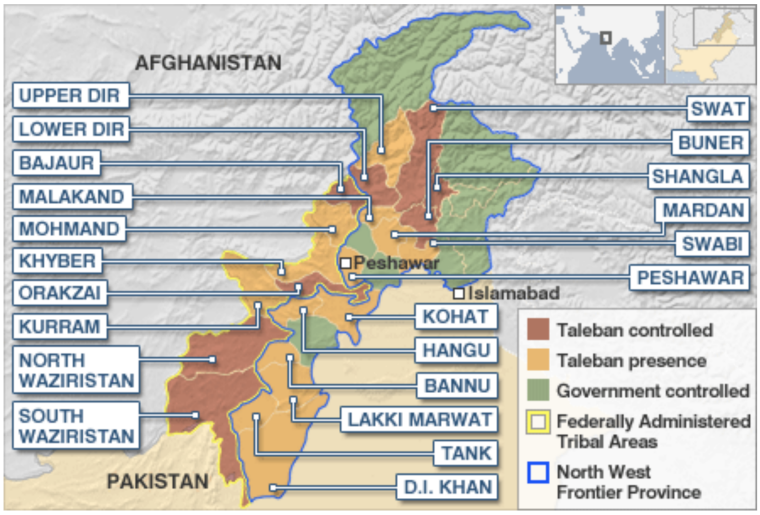

Figure 1: Taliban Occupation in Pakistan by 2009

Aligned with the origin of Malala's activism story, we propose the hypothesis that women's rights are weakened in areas of close proximity to the religious-extremist terrorist groups. For this reason we are interested in studying the legacy the Taliban groups have left for Pakistani girls. More specifically, we are interesting in finding the **causal effect TTP terrorist incursions that started in 2007 have had on girls' access to education in rural areas.**

Model

In order to measure the effect of the Taliban influence in women's access to education in Pakistan, we first define which areas of the country have been under Taliban control. After the terrorist movement was launched in 2007, leaders of the TTP and sympathising militants settled in what was known as the North-West Frontier Province (NWFP) of Pakistan, controlling 20 tribal regions and districts: South Waziristan, North Waziristan, Orakzai, Kurram, Khyber, Mohmand, Bajaur, and Darra Adamkhel, and districts of Swat, Upper Dir, Lower Dir, Bannu, Lakki Marwat, Tank, Peshawar, Dera Ismail Khan, Mardan, Charsadda, and Kohat.[1]

We define the beginning of the TTP incursion from 2007 to 2009 as the intervention we want to measure the effect of. Our approach is to compare the difference across time in girls' school enrollment rates in the 20 cities that were controlled by Taliban after 2007 (our treated group), to the enrollment rate in cities that were not controlled by Taliban (our control group). We propose this *difference in difference* design to approximate an experimental model because there exist permanent baseline differences between the populations that were under Taliban influence and those that weren't. This exercise allows us to mitigate the bias that were to arise from those permanent differences if we looked simply at enrollment trends after the Taliban took over these regions. Using the difference in difference model, we can estimate the causal treatment effect on the treated group (ATT) by comparing the girl's enrollment rate in the two groups both before and after assuming they would trend similarly over time if Taliban intervention had not occurred. This model, however, assumes that Taliban occupation (our treatment variable) must not be correlated to enrollment rates of girls. In the next section it is demonstrated that there is no significant difference between the enrollment rate in both groups.

The model proposed also assumes measured enrollment rates in both groups of cities (the Taliban-occupied and the rest) will follow parallel trends in the period of time before the Taliban intervention. Given the 20 cities that were occupied by Taliban militia were all considered rural, we decide to only assign rural areas to our control group. Rural areas in Pakistan have more similarities in socioeconomic and demographic indicators, for this reason we would expect education indicators to trend in similar rates across years if none had suffered Taliban occupations. The validity of this assumption is also explored in the next section.

Finally, in order to infer causal effects from our model we must assume Stable Unit Treatment Value Assumption (SUTVA) holds. SUTVA indicates there are no different treatments applied to our treated

group, which could be violated given some Taliban occupations are stronger, and thus, more influential than others. SUTVA also holds if there are no 'spillover effects', which is also potentially violated given the rural areas in the control group are geographically connected to rural areas in the treatment group. We keep these potential violations in consideration as we generate conclusions of our study.

Data

The Pakistan Bureau of Statistics (PBS) conducts the Pakistan Living Standards Measurement (PSLM) survey as part of a program started in 2004 with the objective of monitoring of key social indicators that would help design of poverty reduction strategies. The PSLM survey is geographically representative, and the data collected is publicly available in the PBS website for every year since 2004 to 2020 except for years 2009, 2016, and 2017. In this study, we collected all the survey responses for each of the years made available.

PSLM Survey Structure

The PSLM is collected in a two-stage stratified sample design. Districts (also called "Stratum") are classified as urban and rural areas. Each urban or rural district is divided into enumeration blocks consisting of 200-250 households. Districts are taken as Primary Sampling Units (PSU) or as primary hierarchy. Households are taken as Secondary Sampling Units (SSU) or as secondary hierarchy. At the household level, the each observation in the survey collects information on a wide range of topics using an integrated questionnaire which comprises of multiple sections, each of which looks at a particular aspect of household behaviour or welfare. These sections are on education, diarrhoea, immunization, reproductive health, pregnancy, history, family planning, pre and post-natal care, and access to basic services.

The expanse of the PSLM survey consists of all urban and rural areas of the four main provinces in Pakistan excluding military restricted areas. The sample size of the PSLM survey is calculated based on other surveys the PBS conducts in parallel. Keeping in view the variability that exists within the population for the characteristics for which estimates are to be prepared, population distribution, level of estimates, and field resources available; the sample size varied marginally across the years.

Data Aggregation and Manipulations

To meet the objectives of the study, we obtained data from the education and demographic sections of the PSLM survey directly from its website, keeping only individuals with ages between 4 and 15. This age group has been selected to match the United Nations Millennium Development Goals that seek to ensure children of this age group have access to primary and secondary education. After subsetting for these individuals, we have 1,761,674 observations.

In order to obtain comparable units of observation across years, we calculated the school enrollment rate for each district sampled, for boys and girls, separately.

$$\text{Enrollment Rate }_{girls} = \frac{\text{Total enrolled girls in the district}}{\text{Total number of girls surveyed in the district}} \tag{1}$$

$$\text{Enrollment Rate }_{boys} = \frac{\text{Total enrolled boys in the district}}{\text{Total number of boys surveyed in the district}} \tag{2}$$

As mentioned, each district has a designation of whether it is considered a rural or urban area. However, some districts correspond to subdivisions of the same city (they are given the same name in the survey encoding). This happened for locations that have both rural and urban areas. Since many locations that had the two designations and were sampled (and encoded) twice in the PSLM, we defined them as a separate observations. That determines our unit of observation as district ("survey stratum") by year, by rural/urban designation, and by sex.

It is important to note that individual level data on other social and economic factors such as religion and ethnicity would have been very informative to our study, that is, if they had been collected in the survey. The regions of Taliban-stronghold as ethnically Pashtun, and considerably distinct from the ethnic composition of other rural areas in our survey. This variable could serve as a stand-in for culture and provide a more robust causal analysis.

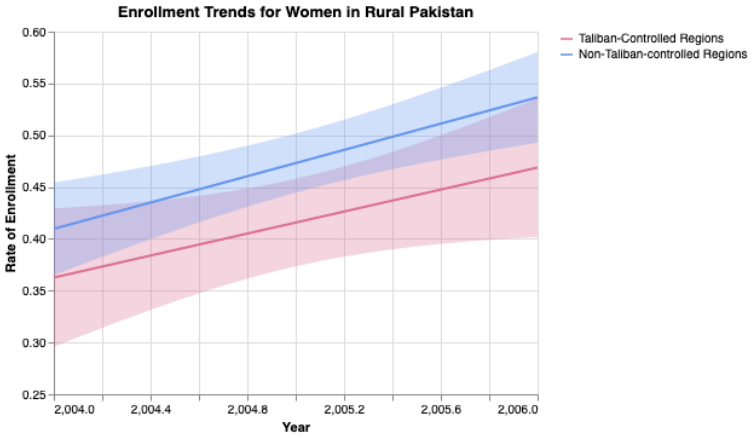

Figure 2: Our proposed model assumes parallel trends in absence of Taliban intervention. We will assume the difference in the trends is negligible.

Education Data in the PSLM Survey

To measure access to education, which is the response variable, we collected data on current enrollment for each individual in a household. Initially, We find that overall, out of those enrolled in school as at the time of the survey, 58% were male while 42% were female. However, the trend flips when we look those who are not enrolled in school. 54% of those individuals are females while 46% are males. Exploratory Data Analysis (EDA) reveals a consistent distribution of males and females for both groups of enrolled and not enrolled in keeping the overall averages. Also, data on whether an individual has ever been admitted to school was collected to gain a better understanding of the response variable. We find that out of individuals who have never been admitted to school 54% are females while 46% are males. Of those who have been admitted to school, we see that females represent 43% and males represent 57%. At a macro-level, 58% of individuals are currently enrolled and 42% are not.

At this stage of the analysis, we notice signs that there may be other factors that could influence the trends of enrollment between female and male over years of our study. These baseline differences observed between males and females are informative for any regression based causal inference on this data set and justifies the use of difference in difference analysis, given that other assumptions hold. Although, we admit that males and females could be compared with regression if we were to use matching in future studies, this will come at a cost of losing needed information in this panel data and, consequently, statistical power. Since pre-Taliban intervention trends for girls and boys were not similar, we compare girls' enrollment rates in districts controlled by the Taliban to girls' enrollment rate in other rural districts, where the pre-2007 enrollment trends as similar as observed in Figure 2. This eliminates confounding information on whether girls live in rural or urban areas and also fulfills a key assumption of our difference-in-difference analysis.

In Figure 2, we can observe trends are almost parallel for enrollment rates for girls prior the interventions studied. In our model, we will assume that in absence of Taliban intervention, girls enrollment in education would still move in parallel across years after 2007. Considering the limitations of the data used to produce this trends, we will assume the difference in the trends is negligible.

This assumption is important for our causal inference because we want to ensure that the selection of sub-provinces in our treatment or control groups was not related to our response variable (rate of enrollment for girls in rural areas). A literature review convinces us that the incursion of the Taliban into the given cities is mainly due to geo-political reasons that are not directly related to rate of enrollment [1].

Demographic Data in the PSLM Survey

Demographic data collected was on age, marital status, gender, province and sub-province where the household of the individual was located. We decided to select these variables because we estimate that the enrollment status of an individual can be influenced by the above-mentioned variables. We see that the mean age of males and females in the survey is 9.223 and 9.275 respectively. After carrying out a t-test to find out if our data set was balanced across males and females for the age variable for each year

the survey was conducted, we got mixed results. For 2004, 2011, 2013, we see that difference observed in the ages of males and females is not statistically significant. Thus, for these years we have balance in age across males and females. However for the remainder of the years, there is imbalance in ages between males and females.

We also notice that about 57% of individuals in the data set live in rural areas while 43% live in urban areas. Individuals, yes children between the ages of 4 and 15, in rural areas are married (621) three times more than individuals in urban areas(256).

For a difference in difference analysis to be valid, the demographic composition of treatment and control groups must be stable for repeated cross-sectional design. A contingency table has been built with each year on each row and treatment and control groups as the columns. A chi-squared test on the above-mentioned contingency table shows that the composition of treatment and control groups across all years is not statistically different. Hence the composition of both groups is stable for the repeated panels (years) of the data.

HERE we need to add table that shows that other variables are stable for grils in taliban-controlled group vs. all other rural places

Challenges Using the PSLM Survey Data

Over the course of our data cleaning, exploration and aggregation we flagged some issues with the dataset that could affect the outcome of our analysis. Primarily, there is missing data in various important ways:

- We do not have survey results from the surveys of years 2009-10, 2016-17 and 2017-18, and we rely on a regression fit in our final analysis estimate the trend in enrollment rate for these years.
- The collection of districts surveyed is not the same each year. This is especially challenging when it comes to further dividing our data into treatment and control groups by districts. To mitigate this problem, we first ensure that we have enough districts with data across the 13 survey years. We also use fixed effects for the districts in estimating the difference in difference coefficient using Panel regression.
- The PSLM surveys a different random subset of the population each year. To ensure that the units observed were consistent across the years we aggregated enrollment rate by district.
- For the year 2007, the rates of enrollment for both male and females significantly drop to 0.3, which is inconsistent with previous trends that we observe. 2007-08 was a time of significant political turmoil in Pakistan, with the assassination of former Prime Minister, Benazir Bhutto, in late 2007, and rampant public protests and strikes, which can explain this steep dip in enrollment rates. However, since 2007-08 is also the year of major Taliban incursions into cities, we use that year as our intervention point, excluding the year's information from our difference-in-difference analysis.
- In the survey from 2018-19, we also see information for only 32 rural districts, since our analysis is limited to rural regions, whereas the count ranges between 90 and 110 for other year surveys. This is an issue for modeling since it means, the composition of districts in 2018 looks significantly different from that in other years. In order to control for this problem, we ensure that the mean enrollment rate in 2018 is consistent with the trends we see before and after 2018. The mean of the enrollment in rural areas for 2018 did not deviate too much from the trend observed.

Analysis

This section subsets our analysis into two parts. In the first part, we evaluate the difference in trends of girl's enrollment rates in the two groups. In the second section, we estimate the difference in difference coefficient through a regression.

Observed Difference-in-Difference

In this part of our analysis, we look at women's rates of enrollment in school across the years the PSLM survey was conducted except for the period between 2007 and 2008. In Figure 3, we compare women in Taliban-dominated regions to women in other rural areas. Note that trends of enrollment in school are negative after 2007 for women living in regions of Taliban domination. This is in contrast to their trends of enrollment before 2007 and also to the trends for women in other rural regions after 2007. Importantly, the enrollment trends for women in other regions continues to increase (at a slower rate) whereas for women in Taliban-controlled regions it declines as observable on Figure 3. From the data, there seems to be a negative effect on enrollment rates of strong Taliban-influence in the region.

Note that just after the discontinuity between 2007 and 2008, values of enrollment are flipped compared to the trend prior 2007. That means that enrollment rate in Taliban-controlled cities is higher compared to the control cities in 2008. This inconsistency with the trends in years before the terrorist campaign started (2004-2006) may indicate a violation of model assumptions regarding parallelism. Therefore, we may not have a strong enough conclusion about negative impact to enrollment rate of girls caused by Taliban presence.

Quantitative Difference-in-Difference

An average difference in difference coefficient is calculated from modeling our outcome variable as in the following equation:

$$Y_i = \alpha + \beta T_i + \gamma t_i + \delta(Titi) + \epsilon_i \tag{3}$$

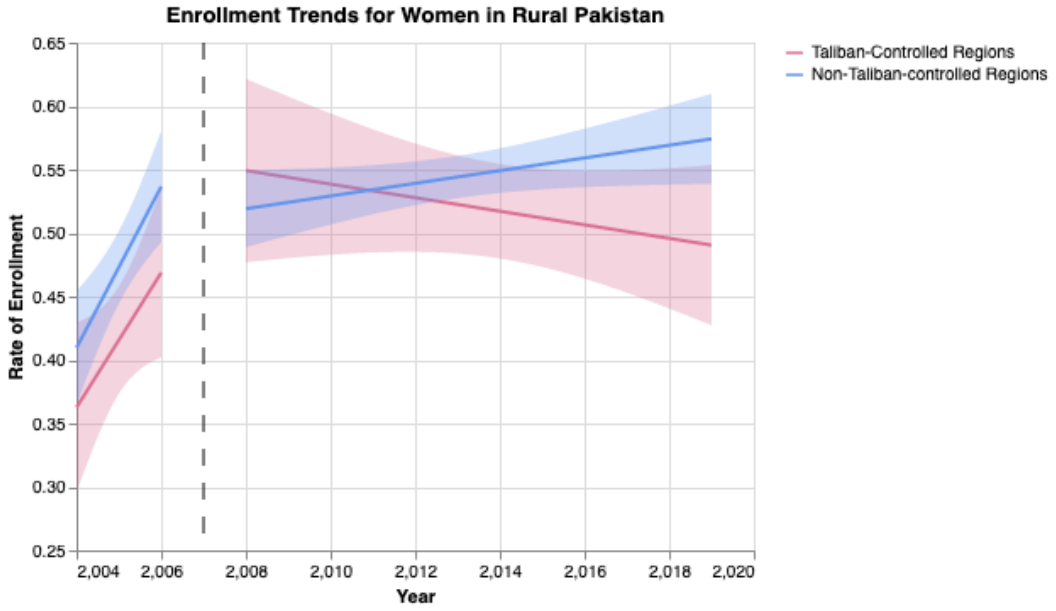

Figure 3: Difference in difference plot of women's enrollment rates in rural Pakistan

where:
α = intercept term
β = effect of Taliban occupation compared to no occupation, holding time effect at baseline
γ = effect of time (post 2007) compared to years pre-occupation, holding treatment at baseline
δ = average treatment effect of the treated (Difference in Difference coefficient)
ϵ_i = the random, unobserved "error" term which contains all determinants of Y_i

The coefficient obtained from fitting the regression model to our data are shown in Appendix Figure 1. The estimated average effect of Taliban occupation (in the cities occupied) is 0.0322, which means that the difference, on average, in enrollment rate of Taliban-occupied cities compared to the average of all other rural cities is 0.032. This coefficient being small but positive means that the districts with Taliban presence, on average, did not see a decrease but a slight increase on average after the 2007 incursion. One important limitation of the average difference in difference is that it does not test for parallel trends because its a mean value. This coefficient is also not statistically significant, with a p-value of 0.35.

Given the data limitations and the time dimension of the treatment effect we are evaluating, we could say the Taliban occupations that solidify in 2009 may not have had immediate effect in school enrollment for girls in those cities. That means, it may have taken some time for the Taliban have social impact in the areas they settled in throughout their movement. Although, this may not be the most plausible explanation if we account for testimonies of women who lived in Taliban-controlled areas during 2009.

There also may be other factors that influence the rural regions of Pakistan that have opposite effect on reported enrollment rates for girls in schools. People in rural regions may have been reported in the PSLM survey that girls in their household are officially enrolled in schools but they were not being allowed to attend school by Taliban rules. If this was true for the years following the TTP incursion, our data could be biased and not a real representation of the education indicators.

Conclusion

Include a summary of problem statement and our answer to research question

We conclude we did not find a significant average effect in enrollment rate in school for girls in cities that were controlled by Taliban in 2007. The results we get can be attributed to the limitations in our data, its incompleteness, and we propose using a complementary linear probability model that can help us understand better the factors affecting the likelihood of girls being enrolled in school. We could also include other variables like income and religiosity in this regression for a more informative study.

Next Steps

Appendix

0.1 Difference in Difference Regression

Dep. Variable:	rate_enrollment	**R-squared:**	0.027
Model:	OLS	**Adj. R-squared:**	0.024
Method:	Least Squares	**F-statistic:**	9.993
Date:	Wed, 06 Apr 2022	**Prob (F-statistic):**	1.69e-06
Time:	20:36:29	**Log-Likelihood:**	30.367
No. Observations:	1089	**AIC:**	-52.73
Df Residuals:	1085	**BIC:**	-32.76
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.4742	0.015	31.737	0.000	0.445	0.503
C(post_2009)[T.1.0]	0.0823	0.017	4.744	0.000	0.048	0.116
C(Treated)[T.1]	-0.0584	0.046	-1.282	0.200	-0.148	0.031
C(post_2009)[T.1.0]:C(Treated)[T.1]	0.0322	0.052	0.614	0.539	-0.071	0.135

Omnibus:	230.948	**Durbin-Watson:**	1.650
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45.917
Skew:	-0.067	**Prob(JB):**	1.07e-10
Kurtosis:	2.003	**Cond. No.**	12.2

References

[1] Rahimullah Yusufzai (2009) A Who's Who of the Insurgency in Pakistan's North-West Frontier Province: Part Two – FATA excluding North and South Waziristan. Publication: Terrorism Monitor Volume: 7 Issue: 4

[2] Center for International Security and Cooperation (CISAC). Freeman Spogli Institute for International Studies: https://cisac.fsi.stanford.edu/mappingmilitants/profiles/tehrik-i-taliban-pakistan

[3] Pakistan Conflict Map, (BBC 2009): http://news.bbc.co.uk/2/hi/south$_a sia$/8046577.stm

[4] David Albouy. Berkeley Economics 131. Program Evaluation and the Difference in Difference Estimator https://eml.berkeley.edu/ webfac/saez/e131$_s$04/$diff.pdf$

[5] The Malala Fund. Malala's Story. https://malala.org/malalas-story