The Key To Success*

Education Trends From The 1990/1991 Demographic and Health Survey In Pakistan

Mohid Sharif

11 April 2022

Abstract

We often take our right to free education in Canada for granted while people in third world countries have to pay for education and often times realize they can do without one. In this data study I obtained a National Family Health Survey from Pakistan, hosted by the Demographic and Health Surveys program in the US. I converted the education research table into a usable dataset which I analysed and visualized to observe changes in literacy rate over generations and provinces. In doing so I concluded that literacy rate in Pakistan has been on a steady incline but is rather slow, therefore there should be action taken to promote education for the development of the country.

Contents

*Code and data are available at: https://github.com/MohidSharif/Issues-With-Education-In-Pakistan

1 Introduction

I grew up in Pakistan, I know first hand how difficult it is to get an education there. Low earning family opt to have their children work labor jobs over giving them an education. Education is not free in Pakistan, schools are privately owned and so having to pay for school when families are already earning pennies is very difficult. When I saw the Pakistan Demographic and Health survey I wanted to see how the literacy rate has changed over the years and if the country is doing something about their low literacy rate.

I obtained the Demographic and Health survey and found a table about house hold education in Pakistan. I then had to understand how the table was recorded and organized. The table organized data by background characteristic which was separated into subgroups such as age groups, residence, and province. The table organized education for each background characteristic by education level; primary, middle, and secondary. The table also had records of the sample recorded for each characteristic. After I had an understanding of what data I had available to me I knew what I needed to do. I decided to create line graphs that would compare the generational literacy rate and calculate how fast literacy rate is increasing in Pakistan.

Using my data to analyze how the literacy rate has changed over the last 55+ years I noticed that literacy rate is on a steady incline in Pakistan. The percent of people falling in the "no education" category have decreased by ~20% when comparing males of ages 65+ and 10-14. I also decided to see how each province fairs on its own in terms of literacy rate, my models showed me that Balochistan had the lowest literacy rate over a ~20% difference from the other provinces. Looking at these stats its very important that Pakistan put more effort into increasing the literacy rate in Balochistan and make education more accessible for future generations.

MALE

Background characteristic	No education	Primary	Middle	Secondary+	Missing	Total	Number of persons	Median years	Mean years
Age group									
5-9	44.4	54.8	0.4	--	0.4	100.0	4102	0.7	0.6
10-14	23.8	51.4	20.0	4.5	0.3	100.0	3274	3.6	3.2
15-19	28.4	17.8	20.8	32.6	0.3	100.0	2448	6.3	5.3
20-24	33.9	19.1	13.2	33.5	0.3	100.0	1883	5.7	5.4
25-29	39.7	20.7	12.4	27.1	0.1	100.0	1561	5.2	4.8
30-34	45.5	16.7	11.3	26.0	0.4	100.0	1269	4.4	4.4
35-39	44.7	17.1	10.2	27.7	0.3	100.0	1083	4.5	4.5
40-44	50.6	20.2	8.1	20.9	0.2	100.0	951	0.0	3.8
45-49	54.3	14.3	8.7	22.3	0.4	100.0	766	0.0	3.7
50-54	61.6	15.9	6.1	16.0	0.4	100.0	678	0.0	2.9
55-59	64.6	17.2	6.1	11.4	0.8	100.0	505	0.0	2.4
60-64	73.8	11.3	5.6	9.3	0.1	100.0	708	0.0	1.8
65+	80.1	9.3	5.0	5.2	0.4	100.0	1398	0.0	1.3
Residence									
Total urban	26.9	29.8	13.7	29.4	0.3	100.0	6535	5.0	5.0
Major city	27.0	27.0	13.4	32.3	0.2	100.0	3772	5.2	5.2
Other urban	26.7	33.5	14.0	25.6	0.3	100.0	2763	4.7	4.6
Rural	50.7	29.8	9.3	9.8	0.4	100.0	14106	0.0	2.4
Province									
Punjab	40.8	30.2	12.0	16.9	0.1	100.0	12330	1.8	3.4
Sindh	44.0	31.5	7.8	16.0	0.7	100.0	4962	1.0	3.2
NWFP	46.7	27.6	11.0	14.4	0.2	100.0	2597	1.0	3.0
Balochistan	63.4	20.8	6.6	7.1	2.0	100.0	752	0.0	1.7
Total	43.1	29.8	10.7	16.0	0.3	100.0	20641	1.3	3.2

Figure 1: The Table Contained in the 1990/1991 Demographic and Health Survey

FEMALE									
Background characteristic	No education	Primary	Middle	Secondary+	Missing	Total	Number of persons	Median years	Mean years
Age group									
5-9	58.7	40.7	0.1	--	0.4	100.0	3840	0.0	0.4
10-14	48.5	35.5	12.4	3.2	0.5	100.0	2998	1.0	2.1
15-19	54.9	15.5	10.5	18.9	0.2	100.0	2219	0.0	3.2
20-24	63.9	13.6	6.4	16.1	0.1	100.0	1798	0.0	2.8
25-29	72.0	10.0	5.0	12.9	--	100.0	1669	0.0	2.2
30-34	75.3	10.4	4.0	9.5	0.7	100.0	1207	0.0	1.8
35-39	79.0	9.1	4.8	6.9	0.2	100.0	996	0.0	1.5
40-44	83.4	7.0	3.3	6.0	0.3	100.0	871	0.0	1.2
45-49	86.3	6.5	2.8	3.4	1.0	100.0	602	0.0	0.9
50-54	93.0	3.0	1.9	1.7	0.4	100.0	805	0.0	0.5
55-59	92.8	3.9	1.5	1.0	0.7	100.0	597	0.0	0.4
60-64	94.5	1.8	1.3	1.6	0.9	100.0	528	0.0	0.3
65+	95.3	2.2	0.3	0.9	1.4	100.0	839	0.0	0.2
Residence									
Total urban	42.8	28.4	10.5	18.0	0.3	100.0	6126	1.5	3.4
Major city	37.6	28.9	11.6	21.6	0.3	100.0	3511	2.8	4.0
Other urban	49.9	27.6	9.0	13.2	0.2	100.0	2615	0.9	2.7
Rural	79.4	15.8	2.6	1.6	0.5	100.0	12855	0.0	0.7
Province									
Punjab	63.7	22.2	6.1	7.7	0.2	100.0	11389	0.0	1.8
Sindh	66.2	20.2	4.6	8.2	0.8	100.0	4345	0.0	1.7
NWFP	81.6	12.2	3.0	2.9	0.3	100.0	2570	0.0	0.8
Balochistan	88.5	7.1	1.4	1.2	1.8	100.0	676	0.0	0.4
Total	67.6	19.9	5.2	6.9	0.4	100.0	18981	0.0	1.6

Figure 2: The Table Contained in the 1990/1991 Demographic and Health Survey

2 Data

I obtained my data from the **DHSProgram** (IIPS 2009) final report database, using the 'pdftools' package (Ooms 2022) and the statistical programming language **R** (R Core Team 2020). I used the **tidyverse** package for data cleaning and manipulation (Wickham et al. 2019) and **kableExtra** for table formatting (Zhu 2021). The header includes two lines of code **"usepackage{float}"** which allows the use of float in our R markdown and the line **"floatplacement{figure}{H}"** (user9112767 2018) which keeps the tables and figures locked in the specific place where they are written in R markdown.

This data set contains the highest level of education achieved for every age group, residence and province. Education level is classified into four categories; no education, primary, middle, and secondary+. Age groups are organized from 5-9 to 65+ years of age, each age group being a group of 5 years. Residence is organized as either total urban, major city, other urban, or rural. Finally provinces are classified as Punjab, Sindh, NWFP, and Balochistan.

The data we are interested in is average level of education for each age group and average level of education for each province. I will look for any patterns that show progress in education in each age group. We also want to see if any province has lower literacy rate than expected and how this concerns the countries literacy rate.

(Tables 1 and 2) show the cleaned data organized by gender.

Table 1: Literacy Rate Data For Males

Background Characteristic	No Education	Primary	Middle	Secondary+
5-9	44.4	54.8	0.4	NA
10-14	23.8	51.4	20.0	4.5
15-19	28.4	17.8	20.8	32.6
20-24	33.9	19.1	13.2	33.5
25-29	39.7	20.7	12.4	27.1
30-34	45.5	16.7	11.3	26.0
35-39	44.7	17.1	10.2	27.7
40-44	50.6	20.2	8.1	20.9
45-49	54.3	14.3	8.7	22.3
50-54	61.6	15.9	6.1	16.0

Table 2: Literacy Rate Data For Females

Background Characteristic	No Education	Primary	Middle	Secondary+
5-9	58.7	40.7	0.1	NA
10-14	48.5	35.5	12.4	3.2
15-19	54.9	15.5	10.5	18.9
20-24	63.9	13.6	6.4	16.1
25-29	72.0	10.0	5.0	12.9
30-34	10.4	4.0	9.5	0.7
35-39	9.1	4.8	6.9	0.2
40-44	83.4	7.0	3.3	6.0
45-49	86.3	6.5	2.8	3.4
50-54	3.0	1.9	1.7	0.4

We can observe that our values for each education level are organized in decimals, this represents the percentage of the population that completed this level of education. We want to ignore the first age group since ages of 5-9 would have no data available for education level above middle school.

We want to see how the literacy rate in Pakistan has changed over the years. To do this I will visualize my data on a line graph comparing the literacy rate for each age group. This way I can compare how much the literacy rate has changed for each age group and continue doing this for each education level.

3 Results

I first filtered the unwanted rows, this included any rows that did not include age groups. From here I needed to graph one line for each education level, I graphed each line and associated it with a color, this way I could create a legend for all my lines. Now that I had my graph with lines for each education level, I labeled my X and Y axis and created my legend.

(Figure 3) shows the generational literacy rate comparison for each education level for men.

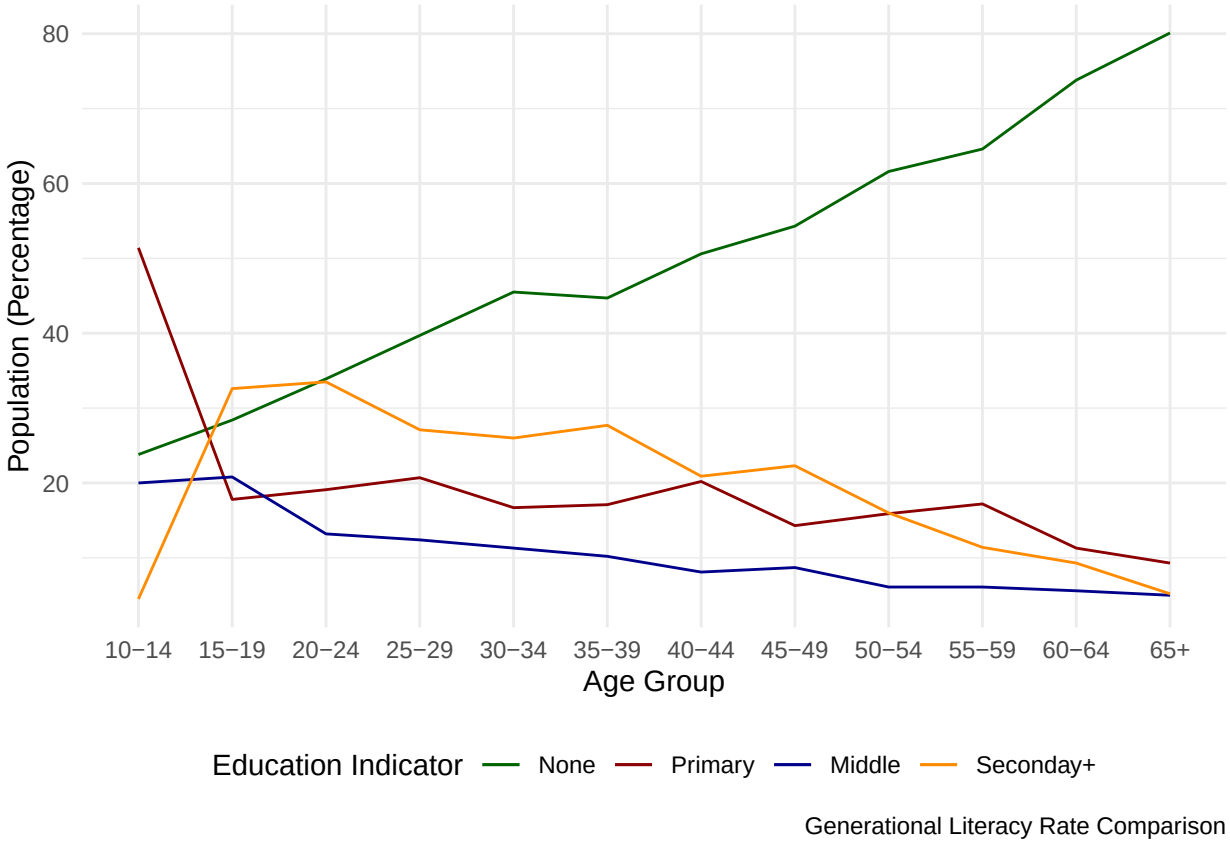

Figure 3: Male Generational Literacy Rate

Our graph shows a steady decrease in percentage of people in the "no education" category, if we calculate this into rate of increase of literacy from the ages 10-14 to 65+ we can see how the rate of literacy in Pakistan is increasing every year. We can do this by first obtaining the values at our end points, our current "no education" rate for 10-14 years of age is 23.8% and for ages 65+ is 80.1%. From here we can take the difference and divide by the years passed. This would give us a 80.1-23.8 = 56.3, and we can divide this by the years passed which is 65-(10to15) = 50to55, dividing 56.3 by 50 and 55 we get an approximate 1.1% increase in education per year for the past 50-55 years in Pakistan.

I now want to look at the difference in literacy rate in each province. For this I can just create a graph to visualize the percentage of the population in each province that falls under "no education". To do this I can first filter background characteristic so that I am only working with the provinces, from here I can plot the points for their values corresponding to each province.

(Figure 4) shows the statistics for each province vs no education rate.

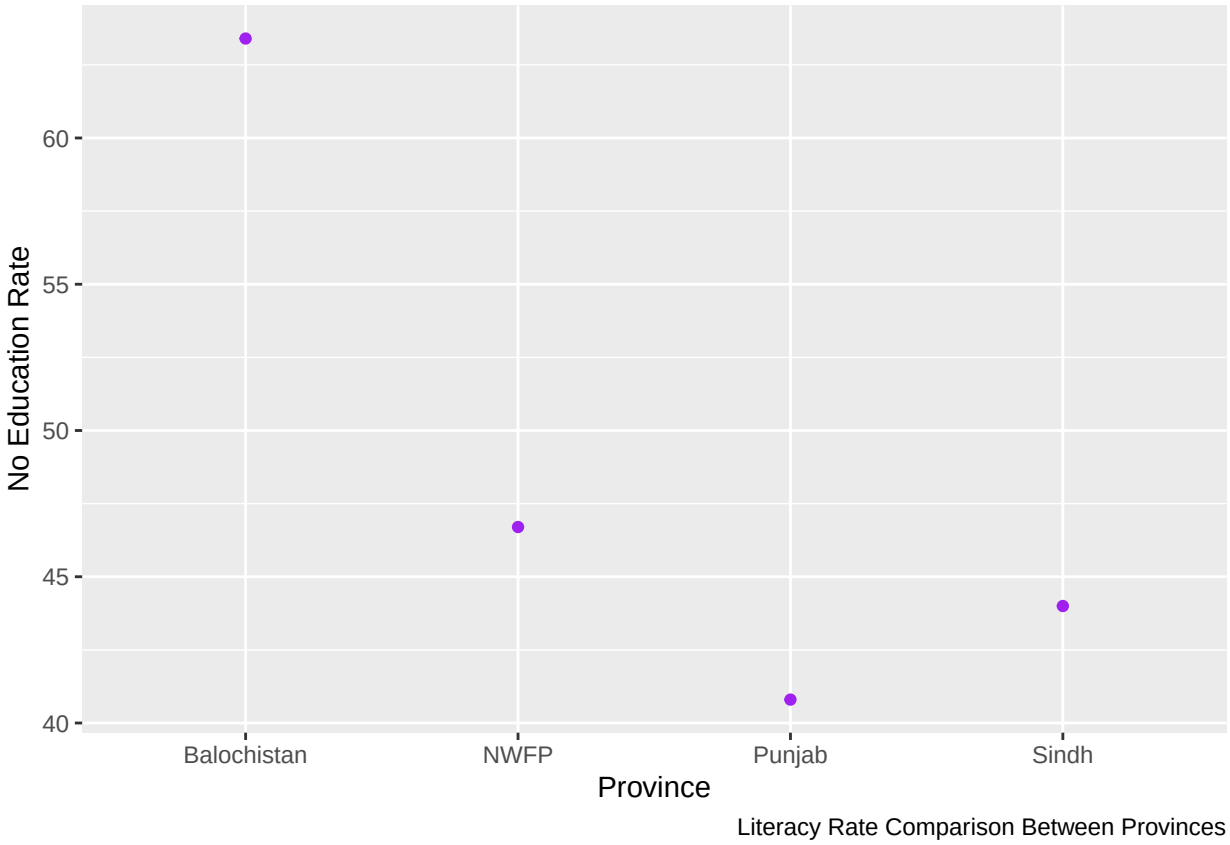

Figure 4: Male Regional Literacy Rate

From (Figure 4) we see a drastic difference in literacy rate between Balochistan and the other provinces. Balochistan seems to have the lowest literacy rate in Pakistan while the other provinces are in ~7% range of each other.

Balochistan has been known as the most neglected province in Pakistan despite being the largest and a resource-rich area. Blochistan has been going through a number of issues including terrorism, poverty, and illiteracy. Education is one of the major issues that plays an important role in the social and economic spheres of a country and Balochistan has been neglected for years (NADEEM 2018). Balochistan also has one of the lowest number of educational institutes for women in the world.

4 Discussion

Using our findings from above we can conclude that the key areas for improvement in Pakistan's literacy rate are Balochistan and overall increase in access to education. Though the overall literacy rate in Pakistan is on a steady incline it is still far too slow at a 1.1% per year. As the literacy rate reaches higher percentages the increase will become smaller and progress will be slowed again. Pakistan needs to make education more accessible, whether that be providing free education or making affordable education options available. Education is necessary for a country's progress and educating the general public about general knowledge is the key to eliminating corruption that has overrun the country for decades.

Rate of primary school education for ages 5-9 has peaked, doubling from prior age groups. This is a great sign for future education, since children attending primary school are more likely to continue their studies. Pakistan surprisingly has a higher population that is attending secondary+ education than middle school, this could be due to an interest in studies later in life, or maybe students are returning from abroad to study for their higher studies. I have personally seen this case from relatives that live abroad as its much easier to get into a medical school in Pakistan than in western countries. So families will come abroad to study up until the end of high school and then return to study in Pakistan past high school.

Balochistan is a big area of worry with its extremely low literacy rate. Balochistan has one of the lowest rate of literacy for women in the world, sitting at over 88% illiteracy rate. This is one of the biggest problems that need to be resolved, but truth be told, Pakistan has been run by corrupt officials for decades. The officials in Pakistan care more about making money and their reputation than making a difference in their country. Progress is slow, people are uneducated and uneducated people are less likely to understand the issues regarding politics and the needs for social and economic change in the country.

There isn't all bad news though. There are many well educated individuals in and outside the country that are trying to make a difference. People are studying abroad and taking their knowledge back to Pakistan and doing their best to educate the public. The public has become very hostile towards corruption and stand up to fight against it when they see it and this has been bringing the country on the incline over the years which we can see from our findings. In conclusion I think the future generation in Pakistan will be well educated and progress in the development in the country's social and economic sphere. Pakistan needs to make education affordable and provide education facilities in areas that are depraved of them. Balochistan is one of the main areas of concern that must be taken care of immediately, building small schools and providing education there would be a great way to start. I think in a 5-10 more years Pakistan can reach and even surpass the average literacy rate in the world.

Appendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to strengthen the capabilities of the population research centres in Pakistan.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by the National Institute of Population Studies alongside IRD/Macro International Inc.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The creation was funded by the Government of Pakistan.
4. *Any other comments?*
   - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - The instances represent the background characteristic of the individials that were sampled, which were furthere classified into age groups, gender, area of residence, and province.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 22 instances
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset does contain all possible instances.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consists of 9 values which are as the percentage of the instance that falls under each category, the categories include: no education, primary school, middle school, secondary+ school, missing, total, number of persons, median years, mean years.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - The first column of each instance corresponds to the state it represents, which is the label associated with it.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - There is only 1 instance of missing information. In which case, the missing value is due to unavailable data. Namely, missing data due to ages 5-9 unable to attend middle school+
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - There are no relationships between individual instances.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - There are no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- There are no errors, sources of noise, or redundancies in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There is no confidential data, and the dataset is publicly available.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - Columns that might cause anxiety include: lack of education available to women, this is due to the fact that Pakistan is still a developing country and thus gender role stereotypes exist.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - The dataset identifies sub-populations by age groups and gender. Male and female is data is completely seperated as the data is significantly different. Data for age groups is due to the availability of education for certain age groups.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It is not possible to identify individuals in any way.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - None.
16. *Any other comments?*
    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data associated was collected with the help of the Institute of Population studies in Islamabad, Pakistan. The data was collected through field work from all provinces.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - Manual human curation.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?* -Geographic stratification subdivided each state into regions, from which villages were further stratified
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - The operations were supervised by the senior field staff of each region. Compensation data is not available.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of*

the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
   - The data was collected over December 1990 and May 1991.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - Ethical review processes were not conducted.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - Data was obtained from Demographic and Health Surveys website: dhsprogram.com
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - The individuals voluntarily interviewed with data collectors. The notice of data collection is not available.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - The individuals consented to the collection and use of their data. The exact language in which consent was granted is not available.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
   - A mechanism to revoke consent was not provided.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - An analysis of the potential impact of the dataset and its use on data subjects was not conducted.
12. *Any other comments?*
   - None

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - The data was originally obtained in PDF format. The table from the survey PDF was converted to a usable data frame in R using the pdftools library in R.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - The raw data obtained from the PDF is saved in inputs/data/raw_data.csv
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - R Software is avalaible at https://www.R-project.org/
4. *Any other comments?*
   - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - No it has not.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - https://github.com/MohidSharif/Issues-With-Education-In-Pakistan
3. *What (other) tasks could the dataset be used for?*
   - The dataset can be used for examining the state of households, women and children in Pakistan in

1990/1991.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - The cleaning process is very specific to the way this table was formatted in the original PDF and may not work on other tables.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - No.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset will be distributed using Github.
3. *When will the dataset be distributed?*
    - The dataset will be distributed in April 2022.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - The dataset will be released under the MIT license
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - There are no restrictions
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No such controls or restrictions are applicable.
7. *Any other comments?*

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Mohid Sharif
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Can be contacted via github
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There is no erratum available currently.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Currently there is no plan of updating the dataset.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - The dataset was made via survey findings conducted in Pakistan. There are no applicable limits as the people took part voluntarily.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe*

how. If not, please describe how its obsolescence will be communicated to dataset consumers.

- The older versions would not be hosted. Data may be updated which can be checked through the commit history on github.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - There is no way currently for people to contribute on the dataset

B References

Wickham (2016), Ooms (2022), Wickham et al. (2019), Xie (2022), Bilal Haq (2022)

Bilal Haq, Ritvik Puri. 2022. "Now That's Fresh Water," April. https://github.com/haqbilal/India-1992-DHS-State-Findings.

IIPS. 2009. "National Family Health Survey." *NFHS, India*. IIPS. http://rchiips.org/nfhs/.

NADEEM, FAJAR. 2018. "Illiteracy in Balochistan," February. https://voiceofbalochistan.pk/opinions-and-articles/social-development/illiteracy-in-balochistan/.

Ooms, Jeroen. 2022. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

user9112767. 2018. "How to Hold Figure Position with Figure Caption in PDF Output of Knitr?" *Stack Overflow*. https://stackoverflow.com/questions/29696172/how-to-hold-figure-position-with-figure-caption-in-pdf-output-of-knitr.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.