

IN CONVERSATION WITH:

**MS. ADITI ROY
BHOWMICK**



Aditi is an accomplished first-year PhD student in Public Policy at Harvard Kennedy School, specializing in Economics. Previously, she served as the India Director for Development Data Lab, where she co-authored several research studies and spearheaded the development of a gender research agenda.

Aditi's interest in development economics research was piqued during her work on education and childhood development programs in India.

Editor: In India, we observe a trend of increasing centralisation of social sector schemes. This is parallel to the increasing emphasis on collecting disaggregated data and going down to the district level. What implications does this have for policy makers given that data provides a more effective source for evidence backed and targeted policy making, however the schemes are being increasingly formulated with the centre in mind?

Aditi Roy Bhowmick: I think the response to this question would be amiss without talking about the political economy first. We talk a lot about evidence-based policy but that's actually not how policy is formulated. Policy is not made based on a neat assessment, or some study that happens based on a census or a sample survey. That's not how it happens, irrespective of which government party is in power, it is broadly aligned on development priorities. There should be a minimum standard of health and education services so there is no kind of disagreement of facts, but the disagreement comes in terms of how much budget is going to go to specific priorities and ultimately it comes down to how much finance or revenue is available to individual departments to implement their objectives.

No one disagrees that all of the things they want to do are important but then there's a limited budget in the economy. And those decisions get made less in terms of data and more in terms of relationships between secretaries of specific departments and the minister associated with a particular department say, rural development, that's how money gets allotted, that decision will invariably happen at the central level, and it has

always been so and I don't think it happens the other way around anywhere in the world. Where data does come in is after the political economy decision has been made that in a particular year, say, we are going to focus on say increasing foundational literacy to some level across the country, or we want to make sure that every kid in the 5th grade is able to read a 4th grade level copy, or say the Government goes into mission mode about that or say the Government goes into mission mode about vaccine delivery, etc then in the implementation phase is where the expanding data collection, that's where it actually comes in.

A lot also goes wrong in implementation. Oftentimes, there is granular data; say you have a poverty targeted programme and you give an unconditional amount of money based on your income, if it is below a particular threshold and that money you get a sense of how much to give each district based on your understanding of poverty distribution at the district level. But if the data is in poverty and consumption measures even to the government department can only be readily available at the state level you are not able to do targeting at the district level. That's where implementation can go away, you don't know if your policy was actually successful because you wanted the resources to be targeted to certain areas but you didn't have the data to direct you even when sending out the resources and you also don't have the data that directs you to see whether it's reached those places and what the long run outcome seems. So to your question about schemes being decided at central level, that's always the case and that will remain the case. I think why data has a lot of potential in changing things is in the ability to continue to do the things that the Government has been doing much better and to enable them to stop missing the mark so ridiculously.

But I personally think data is a necessary, not sufficient condition to have these things work more efficiently because what is the incentive of the district level officer? If I can tie whether those outcomes are changing, how are incentives currently, he will not get into trouble if he has shown that some form has been uploaded and I have shown that the money has reached the account, he is not going to get into trouble based on some very cleverly designed reporting showing that targeting was actually on point showing that the outcome you wanted to move, actually moved. So, the incentive system is also not there so I think it's great that there is a lot of geographically disaggregated data available but there is a lot of room for innovation in the sphere of how policy is then implemented and monitored across levels of government setting and that's not just for the responsibility of the Government, that's also the responsibility of policy researchers and think tanks of society. If the Minister of Health comes to you and says we are trying to implement a program and I want you to do outcome-based accountability taxing of how well this programme is working, what are your suggestions as an expert? What program design should I incorporate to make that happen? Then as the civil society expert you need to be informed about what will work well and you need to be ready with renovation based on the resources available. That's how I think about the nexus of data. I think data has a much larger role to play in making sure that once the policies are decided and implemented, they are on mark, as to what policies are getting decided because that will always be a political economy question. Anyone who thinks the other way that policy gets made on the basis of research and evidence is wrong because that's not how power works and that is the fact of life.

Editor: Thank you so much ma'am for the answer, it was really insightful.

Editor: This question will sound a little convoluted but so far in discussions about open data for development, it is often pointed out that Government should fear public opinion being swayed by data because in essence, for the general public facts don't change their mind, figures, numbers on those screens won't change their mind. Therefore, do you think that digital literacy for the masses (to the point where these published figures and data available online) does change their mind? Can this be used to push for more progressive policy initiatives, once all of us collectively understand what is going on with data?

Aditi Roy Bhowmick: There is some nice literature in the Political Economy space. I think there is a paper by Abhijeet Banerjee where prior to Delhi Municipal level elections they put out the report cards of current incoming candidates, what their performance record was and from what I recall of the paper, that did influence how people ended up voting. I remember from my time in undergrad in Economics and Political Science, in the Political Science classes we used to read a lot about under what circumstances do people vote in terms of Identity vs Policy and one of the papers we used to study in the Indian context, had a general takeaway that in the absence of clarity, identity was kind of a heuristic. There's a lot of confusion about what your manifesto is and it's not even a part of your campaign that strongly, so in the absence of all of this information, I think identity is a pretty fair bet. If we align on both things, I think in that sense there's a lot of room for us to get very sophisticated around elections in India.

For instance, I don't know if you've looked at '538', it's a platform led by Nate Silver that looks at this in the US context. If you look at election campaign videos in the US whether at primary elections or at all levels of governance, they talk about the specific

policy issue they are going to look into, but that is not to say that people still use heuristics to vote in the US. So that being said, there is credit given to the fact that people do have a sense that X senator is going to focus on A, B, and C and I guess the current government did say that they were going to do something about Section 370, and they did that, so in some sense they stuck to their manifesto, and people who voted for them probably kept that in mind. Talking about data and how it can sway people's political opinions and preferences, I think there is a lot more to be done at lower levels of elections. I don't even know how aware people are. I am personally ignorant about the statistics about participation in local elections, urban local body elections, for instance. I don't know how politically savvy we are; maybe we don't see the full feedback.

I do see data playing some pretty interesting role, in just one example over the past three-four years. I don't know if you have noticed, but a lot of leading news houses in India now have a data team, and they might still be figuring out what exactly to do with the data team. It is a little bit weary, but it wasn't a thing, and now it is, and around state-level elections, you do now see reports and little data corners in the news section with constituency-level demographics shown, and people are getting more savvy. To what extent does that influence how people vote? I'd be very surprised if that connection is already being made, but I think they are just getting started. Like I mentioned in the municipal elections study from Delhi, I think there are instances and pretty controlled experiments where it has been shown that when you arm people with information on policy, they are likely to incorporate that.

I don't know if any of you want to adopt research as your career, but there's so much to be thought about. I've just raised a bunch of questions that I don't have answers to. I'm

sure a lot of research has been targeting it, but this question is really interesting to me: If women's voter turnout has increased over the past two general elections, what are the issues they are voting on? You know, prohibition is the big thing that they have, and both are hugely interesting questions and how can data lead to it? It's not a question of whether politicians will fear it or not; it's something that's going to happen. I think of data as an alternative language that's already showing up in how we communicate with each other.

Editor: Thank you for that answer, ma'am. One of your previous answers which took into account the power play which goes into policy making. That point brings us back to a root problem which I believe a lot of economists have faced, that statistics in any country is not free from corruption or vested interests of the organisation that's actually collecting them. So, my question is, what are the aspects of India's official statistics machinery that constrain the formulation of policies for well-being and development, and how can these pitfalls be amended?

Aditi Roy Bhowmick: The co-founders of Development Data Lab, my colleagues and I, we put together a paper called Big Open Data Vision for Development in India where we were assessing India's performance with respect to its data and statistical capacity over time, when India actually used to be pretty ahead of the curve in the times of PC Mahalanobis, when he was leading the charge at ISI and setting up the NSS. It was really ahead of the curve after accounting for its level of development, but over time I think we have fallen way behind. We don't have a sense of what has happened to consumption. NSS used to be a pretty reliable source overtime to track consumption, but with the 2017 one being reformed, it resulted in a lot of confusion about what happened to poverty.

And the World Bank and the IMF cannot agree on whether poverty has declined or not since 2012. This confusion breeds ill-informed debates, and it gets all personal and political and the reason I'm pretty attached to data is that it's neutral in a way that's refreshing. I often go in with my personal priors when I'm doing some analysis in the data, and I know that I can be surprised and then you just have to take it as it is. The thing that has been really frustrating in trying to work with publicly available data in India over the past decade, is poor temporal resolution in that we don't have granular information about what happened to consumption poverty since 2012 and that's a big gap. But also, to take stock, India is still better off than a lot of countries.

I forget when the last census in Pakistan was, it was definitely not 2011 and in many West African nations where friends of mine work in the policy space, the data is even worse. But that's not to pat ourselves on our backs and say, 'Oh look, there are still countries and systems where that's how it works' because that's the race to the bottom, especially because that country (India) was pretty ahead of the curve for a long period of time. So I think the need of the hour is definitely plugging that gap, the consumption data gap that is widening. The new census may still happen as soon as possible. But that is also tied to politics in all sorts of interesting ways. And to your broader point about if the production of data is invariably and maybe increasingly political, how do you get your information from? As researchers, the beauty for us is that there are multiple levels and sources of data. You have the 2011-12 census data and then you have separate researchers and sample surveys, such as the IIT's and so on.

They may be representative at the district level, but the way to check whether things are in order is to take the census data, aggregate it at the district level and then see if things speak to each other. And then you also have other researcher generated surveys,

sample surveys, as well as satellite data. So, we are in a good place where we are able to triangulate. If you're trying to uncover some kernel of truth about a question you care about, suppose you care about the relationship between availability of childcare in the form of more Anganwadi centres in urban areas and female labour force participation. You have multiple sources to track that from, and that should ideally allow you to give some sort of a convergent answer. If not, then that is also pretty informative, so there's enough to still be thorough. We can always do better.

The thing that's frustrating is typically publicly available data is not disaggregated beyond the district level, there's that time component and things probably don't speak to each other, but that is some of the work we were trying to do while developing Data Lab and Niti Aayog came out with a platform called the National Data Analytics Platform, which I think is a great resource for students in the economic department. You have a government data platform where you have multiple data sources that can actually be merged on the platform itself. So, you can take, say, the annual survey of industries, collapse it at the district level, and merge it at the district level to something else that you're interested in, say School enrolment data, I don't have the thing populated in my head. I don't know why exactly you need that, but you can actually do it, now that you can visualise and create output on the platform itself and integrate it into whatever term paper or blog piece that you're writing.

So, for the curious mind, I think there's plenty of resources. The non-government sector is also stepping in a big way, so there's a Development Data Lab, but there's also a lot of organisations in the same space. You have PRS Legislative research which produces a lot of very important statistics about the parliament, what goes on in the parliament, how long does it take to pass each bill, what is the attendance like during

elections, how does it break down education levels of parliamentarians? All those kinds of analyses are possible. But then they also do a lot on breaking down bills and legislation that are passed in parliament. Separately, there's Civic Data Lab which produces something called Open budget Portal or something similar, with the budget documents typically in the form of PDFs and it takes a lot of work to actually pass them and turn them into something that you can readily pull in, analyse and create the charts from. So, they've done that work into all those PDFs, making them into readily usable data sets. The beauty of being in India is that we have so many problems and the scale is so large, but there's also a lot of information being generated, and things are changing rapidly. So, for the curious mind and for people who are interested in working with data, it's a pretty exciting time and place to be.

Editor: Thank you, Ms.Bhowmick. You said that India has a lot of challenges as a country in itself. So, could you tell us what were the unique challenges that you faced as a researcher while you were handling data sets unique to our country?

Aditi Roy Bhowmick: I have only gotten started in my research career. So just a backdrop, after college, I worked at J-PAL South Asia and I was doing primary data collection and that was a pretty unique experience for me because most of us have grown up in urban India or at best, a Tier 2 town, but I have lived a pretty sheltered life and then I went abroad for college and as soon as college was done I was just helicoptered to Madhya Pradesh. Suddenly, I was working in districts such as Jhabua and Alirajpur that have the lowest adult illiteracy rates in the country and that was such an eye-opening experience. There is so much inequality of experience in India and our life, like a kind of silo to such a great degree, depending on whether you're born in urban or rural India, what caste you were born into and what the income and education

background of your parents are, that your lives just don't intersect. If not for that experience, I would have never been in any of those places. So that was a pretty eye-opening experience for me, being at the helm of data collection.

I've actually implemented a survey where I've been the surveyor sitting in the household, piloting the surveys that I've then trained the enumerators to go and do. So, I've seen the end of the data selection, and it is really hard. At times to get to school, in the monsoon season, there's a brook that is overflowing and you have to wait for a boat to come and get you to the school to conduct the survey. And by the time you reach, it may be too late, and the school may have shut and there are many other things as well. You know, going through the jungle, half of your enumerators are getting scared of wildlife and things like that. So, I have seen that, I have seen the millions of ways in which surveys go wrong so I understand what a massive exercise primary data collection is and that was hugely educational for me. So, I have worked with primary data collection there, and then I also interacted with government officials at the block and district levels, and I saw them actually filling up the little forms that get uploaded and aggregated, which we then do our analysis from. They are typically working in one-room offices and really far from any thriving urban centre. I am just trying to help you visualise what it looks like when the data is being generated and when it reaches where it does. So, it is a pretty crazy feat to pull off and I see them do it.

I saw primary data collection both by doing it for our survey and seeing other government officials collecting it first-hand. And I have seen cheating and funding of educational data, one of my principal investigators actually has a paper on this- on "Pratibha Parv", which is an annual exam that happens in Madhya Pradesh. You should go and look it up. Then, during my master's I was working with a lot of secondary data

and the Data Lab and so I have had exposure to the whole of what is available. My honest take is that there is a lot more available. There are many many issues with administrative data production, yet things come together- best example, during covid, daily infection bulletins, in some places someone has a table that has all the data in it and putting it all together is a nightmare and we saw that.

During Covid-19, India had got a whole lot of people who were really overworked, had their day jobs and were running it all at the end, trying to make sense of all the mess but somehow it worked, maybe it was coordinated at the back end but it wasn't coordinated on the front end. But when they wanted to, they got it done. They got it done with the vaccination platform, right? So, I am not saying that it's an issue. Thankfully in my experience as a researcher, I have felt more overwhelmed by taking what is there, and I constantly wonder "What should I do with this?". You know there is so much and there are so many interesting questions I want to ask and discuss.

Editor: Thank you so much for sharing that with us, Aditi. It was an eye-opening experience for all of us as well. This question deals with the nature of data- whether it is collected in India or elsewhere, data sometimes tends to tell an incomplete story. An index such as participation of women in the labour force does not show the entire truth of women's economic empowerment because it doesn't consider factors such as the autonomy that women have over the income that they earn, so my question to you was how can these aspects of ground reality be considered when you are analysing ground level issues and coming up with policy recommendations for them?

Aditi Roy Bhowmick: I am trained as a researcher, so my job is to produce research which is unbiased from the data and is very clear on what it is coming to. I am not concerned with other aspects such as how do I, sitting here, look at the whole set of

research with respect to a particular policy area and then how do I condense it into action icons for policy makers. I think that part is really difficult and is definitely not the training I have had as much as from a research aspect. So that's one thing. I haven't seen much research translating into policy aspects.

When you are doing an analysis, you need to be very clear about the parameter that you are measuring and what question you are looking to answer. I think where one ends up making a mistake is using one outcome and one metric to be a catch-all. There are so many that you are not actually considering. This is something that I am just beginning to learn, so research is all about precision. This is the data set that I have and out of all these interesting things that are going on that are related to the question I am asking, this is the specific channel I am being able to identify and answer. So female labour force participation is one channel, okay? And what does female labour force participation mean?

Some countries will define it as whether you are working in your own field as female labour force participation while some surveys might not do that. And that is telling you different things about female labour force participation. You define it broadly as getting paid in terms of cash. Do you classify only that as female labour force participation? Then maybe not cultivation in your own farm because here you are not getting cash payment. But then what do you do? Do you then count sitting in a shop that is attached to your house for half of the day as labour force participation? So, it is kind of tricky and there is measurement error. I am saying that even within calculating female labour force participation there are so many nuances and you have to be so clear about, read all of the survey documentation if you are working with the ICS data. You need to go through the documentation and see exactly what information it is that you are

providing and that your analysis is only saying something about that. So, when someone is taking that data point and is carelessly comparing it to the labour force participation measure from some other survey, that's wrong, right? So, there is precision and in itself, these things are very important. Therefore, even if I am only counting labour force participation where I am getting paid, how that has moved over time is important.

My mother just stopped working after ten years of working as a preschool teacher and she wasn't making a lot of money, it was a paid job but just stepping out of the house, going to the marketplace, having a trade associated with her name is already a huge thing and that changes her identity and how she engages with the society and the economy and now that she is retired out of it like, you know, we kind of have conversations about that. So labour force participation alone is telling a lot but again as you rightly pointed out, if you use that metric to then say all sorts of things like, "Oh women are better off" and all that- we don't know if they are getting enough economic autonomy by bringing in more money, We don't know if they're actually having to do a lot more work to compromise for the fact that their husbands allowed them to go and work. So, as good researchers we should aim to be precise and know the limit of what we should be able to say to be as informed as possible about every little outcome that you are putting out into the words. We don't have measures to collect the nuanced things you are alluding to and for some questions the only way to get there is by asking people using very cleverly designed surveys; surveys that depend not only on the concerns of economics but also of psychology and anthropology and so on.

Editor: You have worked largely with extensive datasets and one such important data set in India is the National Family Health Survey. It is widely referenced and is

considered reliable, especially in the development sector. But, if we look at the total number of households according to the census, we have 300 million households. However, the sample accounts for only 3 lakh households which results in a sampling variation of less than 0.5 percent. Does this make the findings of the survey less reliable? And if so, then how do we ascertain for a sampling ratio that might be ideal?

Aditi Roy Bhowmick: When we work with data, we start with documentation as it ensures that the samples drawn, and different sample surveys represent different levels. What I mean by representatives is, say you have the ISPF survey. If we assume the survey says that it is represented at the district level, what that means is that the set of output was drawn into the sample within a particular district. If you compute the average consumption per capita for that set of output at the district level, it won't be far off from the average household consumption if you had truly considered all of the households in that particular district.

I am asked if it is true that the X dataset is representative at the district level. It is very interesting that based on the sample of X thousand households you are saying these average numbers when mapped on the full distribution at X levels should hold true and that is something to keep in mind. The NSS, perhaps, represents the state level and that needs to be checked. So, when you make policy recommendations based on something, say education outcomes at district levels have a different trajectory compared to district B. If your sample is not representative at the district level, then your analysis is true for the sample, but the sample might not be true for the underlying distribution entirely. That is very important to realise. It is particularly important when comparing conclusions based on different surveys and data. So, if you produce data, document it. These things are complicated, no matter how many times you get trained in the

collection, sampling and representativeness, it always takes a minute to recall these things- every time you are on the job and working with data, just refresh because it is not intuitive.

Editor: Thank you for your answer. We have come to the last question now. From what we have heard about the work that you have done and as a Master's Student you have learned so much about working in this field. Are there any tips that you would give to an undergraduate student who wants to work on their data skills and build their foundation in this field? What advice would you give to them?

Aditi Roy Bhowmick: I can only give advice that worked for me, and the blueprint is different for everyone, so keep that in mind. What I have always enjoyed is writing and once I started to move into research work as part of academically large research-based projects, I was constantly taking pre-notes and then incorporating those notes into little thin pieces that I would like. One principle that I have always followed is that I would always have side projects on the burner irrespective of what I was doing.

I wrote on a bunch of issues I was always interested in, but I was getting a lot more context on during my time in the field. I would also read associated papers on that and do a lot of writing. During my Master's programme specifically I had side research papers with the purpose of "If I'm learning something in class, I want to apply it." During my undergrad I was slightly confused regarding what I wanted to do. I had no idea that there is an actual career in research. I wanted to do Consulting and Finance as everyone else was doing but that wasn't working out and I wasn't feeling it. I didn't know what else one would do; I was very curious about Development Economics but how does that career happen and that was very confusing because there was no clear blueprint like a consulting job.

But if you realise that is not what you want then it is a little scary. And then data happened to me and all of that worked out, but I came a little late to the research game because I didn't know what I wanted to do but it ended up well. Like I said I always enjoyed writing and found that I like numbers to an extent that I like grounding my writing in quantitative work. So, when all these boys were doing coding and data analysis, I also wanted to do that. It was never a conscious thing but that is how people talk to each other and I just wanted to do that. I was not inherently interested in data analysis. But don't do things in life just because the boys are doing it, I just got interested in that and I got pretty good at it.

Now I genuinely enjoy it. If I go an entire week without doing data analysis I kind of get really empty. Sometimes I just need some blog posts, like outside my research work. I do some descriptive analysis and I enjoy that and just to keep the mindset that do not let things go rut but one thing that I have noticed is that none of these side adventures have ever gone waste. So, if you are going down a rabbit hole, you have an interesting question and you want to write something on it, sometimes you may reach a point thinking why are we doing this and you realise that actually what I learnt from that fed into me.

A lot of these side adventures, what I thought were nonsense adventures, I see coming together now. I knew certain things about myself, and I enjoyed writing, data analysis came along the way. I had a sense of general issues that I cared about and like thinking about and I started putting them together and just kept practising. Oftentimes I do things on my own when it is not part of my job and I think that ends up making a lot of difference. And you just repeat this over and over again and forget about the outcome. People may read some of my work now but that is sort of fleeting, most people who

have engaged with my work in the past three years are on Twitter (now X) and if X shuts down tomorrow, all of my work is just gone and that's fine, do not attach too much worth to anything like that and remain true to your core work. Readership will happen either in the traditional way or other ways. So that's the principle and here we are.

Editor: Thank you so much. You have no idea how inspirational you were, especially to the upcoming graduates who want to go into research. Thank you for this interview, it was really lovely talking to you.

Aditi Roy Bhowmick: Thank you for having me.

