Response to Roy & van der Weide, 2022

Mahesh Vyas

June 8, 2022

Centre for Monitoring Indian Economy Pvt. Ltd.

Contents

1 Introduction

Sutirtha Sinha Roy and Roy van der Weide, both researchers at the World Bank have produced a working paper using CMIE's Consumer Pyramids Household Survey (CPHS) dataset to estimate poverty in India. This is Policy Research Working Paper 9994 titled "Poverty in India has Declined over the Last Decade but Not As Much As Previously Thought" dated April 2022 – (Roy & van der Weide, April 2022). This note comments on the particular use of the CPHS data in this paper.

The Roy & van der Weide paper assumes that the NSSO Consumption Expenditure Survey (CES) of 2011 is superior to CPHS of 2015-19. It highlights differences in outcomes seen in the CPHS 2015-19 data compared to the NSSO 2011 data and other datasets such as, NFHS or PLFS. It claims that these differences in outcomes are because of a shortcoming of the CPHS sample. It draws largely upon Anmol Somanchi's paper, "Missing the poor, big time, a critical assessment of the Consumer Pyramids Household Survey", August 2021 (Somanchi August 2021) to show differences in outcomes. It then proceeds to make adjustments to the CPHS data to first make it comparable to the NSSO's Consumption Expenditure Survey of 2011-12 and then to transform the resultant data to overcome its perceived shortcomings in sampling.

We see three shortcomings in the Roy & van der Weide paper on poverty.

1. First, we suggest that the authors' attempt to make the CPHS data comparable to the NSSO's CES of 2011-12 significantly dilutes the richness of the CPHS dataset.
2. Next, we suggest that the transformations are based on an unsubstantiated claim that the CPHS sample is biased against the selection of poor households. We comment only on the basis or reason for the transformation (which is the alleged bias in the CPHS sample) and not on the transformation itself or its results.
3. Finally, we also point out some other inaccuracies in the paper.

In the sections below, we explain each of these three shortcomings of the Roy & van der Weide paper on poverty in India. Then, we discuss briefly some of the outcomes discussed in the Roy & van der Weide paper.

2 Attempts at making comparable datasets

An attempt to make the CPHS data set comparable to the NSSO data set is justifiable if the two were measuring different subjects using a completely different sampling frame or if the differences in methods is known to lead to different results. Or, there could be other reasons to believe that the two data sets were not comparable and therefore it was necessary to make the two data sets comparable.

The NSSO and CPHS sampling methodologies are not starkly different. Both deploy multi-strata stratification sampling systems over geographical boundaries. There are differences beyond this, but it is not clear why these differences need to be removed. The authors do not argue that one approach would lead to different results compared to the other.

CPHS does not deploy the sampling-probability-proportional-to-size method but uses a disproportionately larger urban sample which is adjusted by using appropriate weights. Ceteris paribus, the two methods should yield similar results.

CPHS deploys a much larger sample of over 170,000 households compared to NSSO's about 110,000; it has more observations per household in a year (3 compared to NSSO's 1), deploys a more concise but up-to-date instrument, uses CAPI and deploys better controls than the NSSO could have done in 2011-12. Yet, the authors decide that NSSO would be the reference point. Effectively, the authors have decided to dilute a richer and more contemporary database to make it comparable to an old database.

The paper discusses the differences in sampling methodology and the instrument but it does not demonstrate that the differences would render the results non-comparable. It does not justify or make a case that it was imperative to remove the differences.

We discuss two specific problems in the paper's attempts to make the two datasets comparable. First we discuss the problems in making the sample observations comparable and then we discuss the problems in making the items of expenses per observation comparable.

2.1 Selection of the sample observations

The authors conduct several acrobatics with the data to create the ultimate dataset they use. We call these as acrobatics because the justification for many of these is unclear save for the fact that the authors wanted CPHS to mimic NSSO's CES.

The authors create a dataset of Waves to create a database of annual consumption expenditure data. A Wave is a four-month period in which CPHS conducts interviews. But,

Consumer Pyramids Data Extraction (CPdx) service that delivers the CPHS data provides a monthly series of consumption expenditure. CPHS collects monthly expenditure data from households during its interviews. The authors decided to ignore this rich granular data and use a far more truncated data set they create from the rich CPHS database. The only justification for this truncation is their desire to make CPHS mimic CES.

CPHS sample households are interviewed thrice every year. The NSSO's CES survey involved a single interview done during the survey year – July 2011 through June 2012. The number of household observations offered by CPHS was of the order of 430,000 in a year. In comparison, the NSSO had about 110,000 observations. Ceteris paribus a sample of 430,000 observations would yield far more reliable estimates than a smaller sample. This would be true even if we account for diminishing returns over increasing sample size. Yet, the authors decided to truncate the CPHS sample size to about a third just to make it comparable to the NSSO's method of interviewing a household only once in a year.

There is no justification for this truncation save for the desire to make CPHS mimic CES. There is no explanation of the impact of this truncation on the estimations either.

The authors also "exclude districts that are covered by the NSS consumption survey but not by the CPHS to obtain geographical consistency in our analysis". It is not clear how observations can be excluded so arbitrarily from a sample selected through a stratification sampling process. It is also unclear how this exclusion was done because the NSSO sample is based on the 2001 Census and the CPHS sample is based on the 2011 Census and a district of 2001 is often not comparable to a district of 2011. For example, Hazaribagh of 2001 was bifurcated and the resultant population of Hazaribagh of 2011 was three-quarters the population of Hazaribagh of 2001. The paper shows its inability to do the mapping across district boundaries of different vintages quite eloquently and therefore this arbitrary exclusion of observations raises worries about its impact.

For example, the paper says that CPHS covers 514 districts out of 718 districts. This is incorrect because the authors have used districts of two different vintages in the comparison. The 514 districts are of the 2011 Census. In the 2011 Census there were only 640 districts, not 718. The authors refer to a sub-division of the 640 districts into 718 districts and they failed to map (or seek a mapping) of the 514 districts of 2011 to the updated 718 districts. It could have been better for them to just stick to the 2011 comparison as that was the basis of the sampling.

The authors have tried to approximate the CPHS data to NSS's 30-day uniform recall period. In doing so, they have discarded all observations available in CPHS with 2-month,

3-month and 4-month recall. This emulation of the 30-day uniform recall period is done without any evaluation of the advantages of a monthly time-series that is built from a series of 1-month, 2-month, 3-month and 4-month recall period. More importantly, the authors have also discarded the 7-day recall data on consumption expenditure available from CPHS. This exclusion is not explicitly stated in the paper but it is also not stated that the 7-day recall data on consumption expenditure have been used.

The treatment of fast-frequency expenditure items in the paper is not clear. Section 5.2 of the NSSO's questionnaire for consumption expenditure in 2011-12 consists of several expense heads for which the answers are sought only for a 7-day period. In CPHS expenses on similar expense heads are collected for a 7-day recall, a 1-month recall, a 2-month recall, a 3-month recall and a 4-month recall. It is not clear that the paper exploits the faster-frequency better-recall expense heads available in both CES and CPHS.

Having decided to use only the 30-day recall data, the paper then goes on to select only one of the three observations available for a household in a year. The authors had the choice to use all the three observations available for a household or they could have used an average of the three observations. Both these choices would have retained at least some of the richness of the CPHS data. Instead, the authors chose to use only one observation picked by random. This discards the opportunity to incorporate seasonality effects and also overcome the other limitations of a one-point estimate. The only reason why the authors seem to have made the choice they made was once again to merely immitate NSSO's CES. The objective of immitating the CES seems to override the objective of obtaining reliable estimates of consumption expenditure of households. Mimicking CES therefore seems to be the bigger objective than estimating poverty.

In an effort to make the CPHS dataset comparable to the CES dataset, the authors have probably created a new sample set of households for a year. The sample is probably larger than the sample of any singular Wave of the year. Such a new sample would presumably demand a new set of weights. Documentation on this would be useful. These weights should not be called CPHS weights (because they are not CPHS weights) but something else - like pooled sample weights.

The paper states that they use the weights provided by CPHS. They also use the non-response rates provided by CPHS. CPHS provides weights at the Waves level and the months level. But, the authors use neither frequency. They have mapped three monthly observations to Waves, pooled them and then randomly chosen one observation per household for a year. So, they have created a sample of households for a year. Since CPdx does not provide any year-level weights, it is not clear which weights the authors chose in their estimations.

The question is relevant because the pooled sample for a year will likely be bigger than the sample in any of the individual Waves or months of a year. Each household therefore would need an appropriately smaller weight than it is assigned in any of the Waves or months of the year. But, this contradicts what the authors say – that they use household level weights and non-response rates from CPdx. There is significant variation in non-response rates of Waves of a year. This is evident in the data for 2019-20 when the non-response rate varied the most from 64.4 per cent in the January-April 2020 Wave to 84.8 per cent in the May-August 2019 Wave. In 2018-19, it varied between 83.9 and 86.5 per cent; in 2017-18 it varied between 80.6 per cent and 84.6 per cent. The variation in the preceding two years was a bit smaller. It is not correct to apply the non-response rate of a household interviewed in one Wave to a different sample.

While the authors use the household level weights and the non-response rates from CPdx, they reject the weights provided in CPdx for weights for individual members of households. They do so because they believe that the population projections based on weights provided by CPdx have become imperfect over time. CMIE has stated so in its documentation as well. The authors have therefore used the household size of the CPHS sample to re-assign individual weights. As a result, the weight assigned to an individual thus derived is a product of household weights and share of individuals within a household. Households are known to grow at a faster rate compared to the population and CPHS shows a much sharper fall in fertility than even the SRS. It is unclear what the authors have achieved by this mix of data to derive the new weights for individuals in the CPHS dataset.

2.2 Selection of items of expenses

The difference in instruments deployed by the CES of 2011-12 and that of CPHS deployed during 2015-19 reflects the changing pattern of household consumption. For example, CPHS has specific expense heads on vehicle parking and tolls; on insurance and fitness services which were not covered in the NSSO's CES of 2011-12. Roy and van der Weide exclude these in their computations because they were not covered by CES of 2011-12.

The authors have incorrectly stated that CPHS does not capture household appliances, personal transport equipment or other durables. CPHS captures expenses on kitchen appliances and household appliances separately. Expenses on personal transport equipment and consumer durables are captured through EMIs (equated monthly installments to service borrowings during their purchase of these items). CPHS explicitly provides data on EMIs for vehicles and durables besides houses and others. The authors have dropped EMIs while using the CPHS data.

By not including these and similar other expenses, the paper under-estimates household expenses from CPHS and possibly overestimates poverty possibly even after the data transformations they perform.

In fact, what is the justification to match the instruments at the item-level? Fundamentally, both surveys capture household consumption expenditure. It cannot be denied that EMIs, parking fees, tolls and insurance for example are household consumption expenses in today's world. Excluding them is akin to saying that when we compare say, NSSO 2011-12 with an NSSO survey of a pre-internet and pre-mobile phones era we should exclude internet and mobile phone expenses to make them comparable. This defies logic.

By sticking to the consumption basket that is common to both NSSO and CPHS we get a basket that represents neither. It does not represent the consumption basket of 2011-12 and not of 2016-20. This strategy discards the advantage of CPHS reflecting a change in the consumption basket of households since 2011-12.

3 Claims of anti-poor bias in the CPHS sample

The authors suggest, like Anmol Somanchi (August 2021), that the differences in outcomes derived from CPHS and other datasets is because of a "bias" in the CPHS sample. Specifically, they claim that the CPHS sample is biased against the inclusion of poor households.

The claim that the CPHS sample is biased against the inclusion of poor households is the principal motivation for much of the work of Roy & van der Weide April 2022. But, the paper does not demonstrate a sampling bias against the poor in CPHS. It merely conjectures that the differences in outcomes observed are the result of a sampling bias. This is not entirely convincing.

The authors rely on a claim made by Jean Dreze and Anmol Somanchi that the CPHS method of selecting the sample households from villages is biased against households that live on the outskirts of villages and, since poor people, it is believed, live largely on the outskirts of villages CPHS systematically misses them from its sample. But, they do not demonstrate that CPHS indeed misses selection of households from the outskirts or provide any evidence that the poor indeed live on the outskirts of villages. There is therefore no evidence that the CPHS sample systematically misses poorer households.

The critical claim that CMIE misses the poor because the poor live on the outskirts of villages and because CMIE does not reach the outskirts of villages, is a two-level

conjecture. First, it is a conjecture that the poor indeed live on the outskirts. Second, it is a conjecture that the CPHS sample does not reach the outskirts of the villages.

That the poor live on the outskirts needs to be validated. It could well be true. But, it needs better validation than the two references provided by Anmol Somanchi. The first is "States and Minorities" by Dr.B R Ambedkar in 1947. The second is "Caste-ing Space: Mapping the Dynamics of Untouchability in Rural Bihar, India" by Indulata Prasad in 1970. The latter refers to the dilution of caste based control over land following the Bodhgaya Land Movement in Bihar in the 1970s. It in fact, talks about a reduction in the distance between Dalit and non-Dalit dwellings in Bodhgaya in Bihar.

Both references refer to caste and not specifically to income or consumption expenditure. However, as Somanchi himself states, there is no caste bias/discrepancy in the CPHS sample - "CPHS seems to be broadly consistent with the Socio-Economic Caste Census 2011 and NFHS-4 (2015-16) in terms of the share of scheduled caste and scheduled tribe households" (Somanchi, August 2020).

India has changed dramatically since the times referred to by Anmol Somanchi. His claim that the poor live on the outskirts of villages remains a conjecture till demonstrated better than the two sources he refers to in making his case.

We do not claim that the poor do not live on the outskirts or even that the poor do not essentially live on the outskirts. We merely assert that this is not a given. It is a conjecture for today's India.

The second conjecture is that CPHS does not reach the outskirts of villages. While this claim needs a thorough investigation because the CPHS methodology does leave scope for such an outcome, the following needs to be borne in mind.

CPHS sampling begins at one end of the main street of a village. Often, the starting of the main street is on the outskirts. It is not easy to avoid the outskirts in the CMIE sampling system. The average village in India has 300 households. The systematic random sampling exercise of CPHS requires the selection of every nth household in the village, where n is a random number between 5 and 15 and the sample size required is 16. If the random number is 5, then CMIE would exhaust the selection of 16 households on the main street only if the main street contained at least 80 households. 80 households cannot be found easily on just one street of a village. Therefore, there is a high probability that the CPHS sample will include households from the outskirts. If the random number is 10, then we need 160 households on the main street, and if the number is 15, we need 240 households on the main street for CMIE to exhaust the sample selection on the main

street itself. Evidently, the CPHS sample cannot easily escape including households from the outskirts.

We do not know the distribution of households by the inner / main streets or outskirts of villages. The CPHS sampling methodology does provide a higher probability of selection to households on the main streets. But, whether this translates into an under-representation of the outskirts depends upon the size of the village and the random number used to administer the systematic random sampling selection. It cannot be merely assumed that the outskirts are not represented in the CPHS sample.

The CPHS sampling methodology involves multiple stratifications and then a simple random selection process of selecting villages. The sample consists of over 3,900 villages across India selected from over 98 per cent of the rural population. There are rich and poor villages within this set of sample villages. There are essentially-Dalit villages and non-Dalit villages. Roy and van der Weide do not complain about this selection. There was no "bias" in this selection.

The claim that the CPHS sample is biased against the selection of poor households is therefore based on the assumption that the poor cannot be found even in the Dalit villages except on the outskirts of the villages.

The differences between outcomes seen in CPHS against outcomes of similar indicators in other databases are worthy of investigation. Given the objective of the Roy & van der Weide paper of estimating poverty in India, it was important to check if the data do indeed under-represent the poor rather than take it as a given.

We admit that there are limitations in the sample selection processes of CPHS. These are well documented and available in the How We Do It section of the CPHS website consumerpyramidsdx.cmie.com. However, to call the CPHS sample as biased is incorrect. A bias connotes a deliberate attempt to be selective in the selection of households – in this case, to keep the poor out. This is certainly not the case with CPHS.

If the CPHS sampling methodology is consciously or deliberately biased against the poor then by definition it should not contain any poor households. But, this is not true. Somanchi (August 2021) does not state that there are no poor households in CPHS. Roy and van der Weide (April 2022) do find poor households.

It may be more correct to say that both these papers find fewer poor households or fewer households with characteristics of poor households than their a-priori beliefs based on other surveys. This is very different from saying that the CPHS sample is biased against poor households. The latter implies a deliberate attempt to keep poor households out of

the survey. This is not true in reality and it would be an incorrect inference drawn from their observations as well.

CMIE announced (see "There are practical limitations in CMIE's CPHS sampling, but no bias", Economic Times, 23 June, 2021) that it would investigate the claim made by Jean Dreze and Anmol Somanchi that the CPHS sample under-represent poor households who, they claim, live on the outskirts of villages. The CMIE sampling methodology systematically provides households that live in the main street of villages a higher probability of being selected compared to households that live in the inner streets of the villages. It is not clear how much exclusion of the poor does this skewed probability of selection cause in the final sample.

CMIE's investigation involves a physical verification of the over 63,000 rural sample households regarding their location within their respective villages. The total villages involved are 3,965.

CMIE started this investigation in September 2021 but suspended it in October 2021 because of inadequacy of field staff for the work at that time. Work restarted in May 2022 and is currently underway. We expect the field investigation to be completed by end of August 2022. Detailed data and results of this investigation will be made public. Corrections required in the sample will be initiated thereafter.

Initial results from the field investigations show that the size and shape of many villages have undergone substantial changes since the time when the sample was selected, which was nine years ago, in 2013. This is not surprising. But, it poses a new challenge regarding our response to the changes.

4 Other inaccuracies

1. Listing exercise.

   The paper laments that CPHS does not conduct a listing exercise when drawing its sample. There are two reasons why CPHS did not conduct a listing of households in the PSUs. First, efforts made by CMIE to conduct listing exercises were met by resistance from locals and local law enforcement agencies to an extent that the exercise posed a serious danger to the enumerator teams and therefore the entire survey execution in the location.

   Secondly, a listing exercise is not a lasting solution given that CPHS is a panel survey that is administered continuously. It never stops. CPHS is a continuous survey. Population projections are arguably the best suited to provide continuously

updated weights for such a survey. A village / CEB listing at the time of sampling could have provided the situation only at the time of the initial sampling. That was in 2013. It is impractical to conduct a listing exercise in a continuous panel survey.

This lament is typical of the paper's perspective. It systematically fails to see CPHS for what it is, but necessarily sees what it is compared to the NSSO. If the NSSO does a listing, it is assumed that CPHS must also do a listing.

2. The paper seems to suggest that a second-stage purposive stratification – of the PSUs as is done in the NSSO's CES is a superior sampling strategy than the CPHS's strategy that avoids such purposive stratification. CPHS avoids such stratification because it is not a consumption expenditure survey and therefore it is not advisable for its sampling design to include a second stage stratification based on the distribution of past consumption expenses within the PSUs.

   A second stage stratification by the distribution of historical MPCE (monthly per capita consumption expenditure) as done by the NSSO creates a purposive sample suitable only for the purpose of estimating MPCE that assumes a certain stability of the past distributions of MPCE. CPHS relies on stratification based only on geography.

   This again, is an example of the paper seeing CPHS essentially in the light of what the NSSO does rather than for what it is.

3. The paper claims that an expansion of the CPHS sample to include sample households from districts that were not covered earlier are concentrated in the comparatively poor and rural areas of the country. These districts according to the authors had a mean household consumption per capita in 2011 that is 18 per cent lower when compared to the districts that were already in the sample. The authors seem to suggest that additions to the sample were of significantly poorer households. This is not true. There was no significant difference in the income levels of households added or deleted. The impact of the changes on the income of the sample households is negligible. In the tabulation below we compare the average monthly income of households that were deleted from the sample and those that were added.

   In 2014, income of the added households was 0.4% higher than the deleted households. In 2015 it was 6.2% lower. In 2016, it was 34.5% lower but the deletions were of a very small set of just 247 households. As a result impact on the overall income was negligible. Most importantly, in 2017 when the sample was expanded substantially to cover a larger number of districts, the added households had an average monthly income that was 16.9% higher than the ones that were deleted.

These households were not poorer as alleged in the paper but were richer. Further, and more importantly, the net impact of the additions and deletions on the overall average monthly income is negligible – the percent difference is significant only at the second place after the decimal.

Table 1: Impact of changes in CPHS sample

Year	Households (no.)				Avg. HH income per month (Rs.)			
	Deleted	Added	Original	New	Deleted	Added	Original	New
2014	2,170	3,636	137,809	137,942	11,360	11,409	16,483	16,479
2015	9,173	8,776	130,761	130,728	14,367	13,474	15,935	15,930
2016	247	1,012	132,981	133,045	19,626	12,848	16,247	16,245
2017	3,904	6,964	134,890	135,145	14,495	16,937	19,274	19,263
2018	30	81	147,117	147,121	28,936	27,438	23,293	23,293

Notes:
Original HHs = Surviving HHs after deletions + deleted HHs.
New HHs = Surviving HHs after deletions + added HHs.

4. The paper is right in pointing out that CPHS does not include the homeless. Since this is pointed out as a difference with NSSO, it would be useful to know how many homeless units are included in the NSSO sample. This will help us understand the significance of this difference. What if the NSSO also does not contain any data on the homeless?

5 Some comments on outcomes

Differences in outcomes as seen in CPHS compared to outcomes in other surveys require more investigation than attempted here. We are in the midst of conducting a detailed investigation into many of these differences. In the paragraphs below we seek to explain, very briefly, some of the differences.

Some of the differences in outcomes can be traced to definitional differences. These differences in outcomes cannot be considered as examples of any bias in the sample. Here are three such cases.

1. Literacy.

   CPHS definition of a person being literate is far more relaxed than that used by NFHS. In CPHS, a person is considered literate if she claims that she can read and comprehend in any language. NFHS requires that the person should have cleared

the sixth standard in school or should be able to demonstrate her ability to read a text presented by the enumerator.

2. Employment.

   CPHS definition of a person being employed is far more stringent than that used by PLFS. In CPHS a person is considered to be employed only if she is employed for a better part of the day of the interview or on the day preceding the day of the interview. PLFS allows a person to be classified as employed if she is employed for at least half a day in the last seven days and the status of being employed is assigned preference over all other statuses during the week. This is again not a sample bias problem.

3. Access to water and toilet.

   CPHS asks binary questions on access to water and toilet while NFHS has a more detailed set of questions on these facilities. It is possible that the differences in outcomes arise because of these.

The transformation overcomes these and other differences in definitions and execution. It overcomes differences caused by a different sampling frame (2011 Census), a different sample selection system (not probability proportionate to size), different level of stratification, different survey execution system, different measurement frequency, etc. The transformation is to make the CPHS dataset more comparable to other datasets. This transformation is independent of the sources or causes of the differences. The differences are true and the transformations are successful in overcoming these to an extent. But, the claim that the differences are because the CPHS sample is biased continues to remain a conjecture, or a hypothesis that must be tested.

The section on Expenditure (Section 3.2) reveals some interesting outcomes and we offer some comments relevant to our line of study regarding this paper.

The mean per capita expenditure (MPCE) derived by the authors presented in Table 1 of Roy & van der Weide 2022 are systematically higher than the MPCE derived by CMIE without any transformations. In 2016-17, estimates presented by the authors were 11.5% higher than the estimates of CMIE. The difference was relatively small, at 3.8% and 2.8% in 2017-18 and 2018-19. The difference increases to 9.9% in 2019-20.

Table 2: MPCE estimates

	MPCE		Roy&Weide/CMIE	MPCE Growth (%)		
Year	Roy&Weide	CMIE	(Ratio)	Roy&Weide	CMIE	NAS
2015-16	2,193	2,046	1.072			
2016-17	2,315	2,077	1.115	5.6%	1.5%	10.9%
2017-18	2,558	2,463	1.038	10.5%	18.6%	8.7%
2018-19	2,846	2,770	1.028	11.3%	12.4%	10.7%
2019-20	3,143	2,860	1.099	10.4%	3.3%	8.5%

It is surprising that the MPCE estimates made by the authors through various transformation turn out to be higher than the estimates made by CMIE without any transformations. It is surprising because the whole premise of the transformations was that the CMIE sample under-represents the poor. It follows, then, that the average MPCE in CPHS should have been higher. But, this is not true.

Given that (1) the authors have demonstrated the success of the transformations in Section 3.1 to make several outcomes other than MPCE to be more similar to those seen in other surveys and (2) it is assumed that the raw CPHS data under-represent the poor, then it follows that the transformed MPCE should have been lower than the MPCE derived from the un-transformed (raw) CPHS data. But, as the table above shows, this is not true. The MPCE estimates derived by the authors creating a dataset that mimics NSSO and then transforming this using the max-entropy approach leads to estimates of MPCE that are higher than the MPCE estimates using raw CPHS estimates without disturbing the dataset or performing any transformations.

How must we interpret the higher MPCE the authors deliver after transformation?

Note the small difference between the transformed estimates and the un-transformed estimates particularly in 2017-18 and 2018-19. These expenditure estimates do not confirm that the CPHS under-represents the poorest and the richest households in the population as the paper claims in this section.

Table 1 in this section of Roy & van der Weide 2022 compares CPHS estimates with NAS. But, as the authors have noted the difference between NSSO and NAS is the same as the difference between CPHS and NAS. The contention is not with NAS but with transformed and un-transformed estimates. The difference in these is too small to proclaim that these confirm an under-estimation of the poorest and the richest.

A fall in the variance of log consumption per capita since 2011 as demonstrated by Roy & van der Weide 2022 in this section on Expenditure implies an expansion of the middle classes in India. By calling this fall a "gap" the authors imply that they expect the variance to remain constant. What is the justification for such an expectation?

It is noteworthy that the fall in this variance is high in urban and not much in rural. This is not what the analysis should have revealed because the suspicion all along was that CMIE does not adequately sample households on the outskirts of villages. There has been no such suspicion on CMIE's urban sampling. This again shows that the claim that the sample is biased is unsubstantiated.

Apparently, both the second and third moments around the mean suggest that the middle classes have expanded since 2011. The income distribution is more normally distributed than known earlier. And that it has expanded a lot more in urban India than in rural India. Is this non-intuitive? Should this be a reflection of reality or should this raise questions on sampling bias? It is not obvious that these depict any sampling bias. They seem to point towards a substantial expansion of the middle classes in India since 2011-12.

Questions on sampling bias are largely hypothetical and hitherto, unfounded whereas, the growth of the middle classes is evident in the growth of the consumer goods industries and personal credit industry in India.

In the four years between 2007-08 and 2011-12, personal loans from scheduled commercial banks grew at the rate of 10.7 per cent per annum. In the four years between 2011-12 and 2015-16, they grew at a faster rate of 15.5 per cent and in the next four years between 2015-16 and 2019-20 they grew even faster at 18.1 per cent per annum. Housing loans accelerated from 11.1 per cent to 17.1 per cent to 16.2 per cent per annum in the same period. Vehicle loans accelerated from 11 per cent to 14.5 per cent to 16.3 per cent per annum. Credit card outstanding amounts declined during the four years before 2011-12. In the following four-year period they accelerated from 16.5 per cent per annum to 33.6 per cent per annum. Credit card transactions shot up from a growth of 8.5 per cent per annum to 23.9 per cent per annum. Similar acceleration is seen in domestic air passenger traffic, mutual fund folios.

Such acceleration in growth, particularly of personal loans, cannot be driven only by the rich households. It is a reflection of the rapid growth of the Indian middle classes. CPHS seems to capture this phenomenon well.

6 The way forward

CMIE is grateful to all critics for the efforts they take to point out our shortcomings. It takes criticisms of the CPHS work seriously. There are three kinds of criticisms.

1. Misunderstandings.

   Many of the criticisms are misunderstandings. This is understandable. CPHS is a new data set and it is not a replica of other household surveys in terms of methodology, execution systems, frequency, delivery and also outcomes. We strive to help clear those misunderstandings. For example, it is a misunderstanding that CPHS is biased in favour of urban regions. In reality CPHS over-samples urban regions by design and adjusts this by appropriate weights.

   The way forward is for CMIE to explain in greater detail and in more ways than it has done so thus far to clear such misunderstandings.

2. Differences.

   Differences are mostly in the definitions. We disagree with the establishment in the definition of employment, for example. In doing so, CPHS throws a different light on the labour markets in India. We would like to persuade users to appreciate this difference rather than see the deviation from the norm as a shortcoming.

   The way forward is for CMIE to explain its stance clearly and justify its differences with the establishment.

3. Errors.

   It is possible that in spite of our best efforts CPHS erred in some areas. CMIE is committed to investigating potential shortcomings and fixing them.

   As mentioned earlier, CMIE has already initiated the process of investigating suspicion that CPHS under-samples households on the outskirts. It will report on the outcome of this investigation and it will make corrections to its sample as necessary as well.

We believe that the best way forward is to clear misunderstandings, appreciate or understand differences and overcome shortcomings by mending the sample and the sampling or data collection processes. CMIE is committed to move in this direction.