ANALYSIS RIGHTS

Recent Poverty Estimates Are Little More Than a Shot in the Dark

If you are impressed with recent "evidence" of rapid poverty decline in India, think again. The best estimates so far, from a World Bank study based on the Consumer Pyramid Household Surveys, turn out to stand no scrutiny.

Recent Poverty Estimates Are Little More Than a Shot in the Dark

Representative image. Photo: Dave G/Flickr CC BY ND 2.0

Jean Drèze and Anmol Somanchi

ECONOMY GOVERNMENT RIGHTS 12/APR/2023

Poverty estimation in India has traditionally relied on consumption expenditure surveys (CES) to estimate the share of the population whose per capita expenditure lies below the 'poverty line'. Until recently, CES rounds were conducted at regular intervals by the National Sample Survey Organisation (now National Statistical Office). With the suppression of the 2017-18 round by the Union government, however, the last available CES round (2011-12) is now more than a decade old. This has thrown a shroud of darkness on poverty levels and trends.

Meanwhile, the Centre for Monitoring Indian Economy (CMIE) launched a series of large-scale Consumer Pyramid Household Surveys (CPHS) that include consumer expenditure data. In principle, CPHS data could be used for poverty estimation. Unfortunately, the CPHS surveys fail basic tests of national representativeness.

In particular, poor households seem to be underrepresented in these surveys. The bias is far from trivial: according to CPHS, for instance, 100% of

households in Bihar (India's poorest state) had water within the premises, 98% had a toilet within the house and 95% had a television in 2019 – this is poetry.

The corresponding figures from the fifth National Family Health Survey (NFHS), a fairly reliable source, are much lower – as one would expect: 89%, 62% and 35% respectively.

There are major biases at the national level, too: for instance, the share of adults with no formal education in late 2018 was just 2%, according to CPHS, compared with 17% according to the Periodic Labour Force Survey (PLFS).

Acknowledging these biases, two World Bank economists – Sutirtha Sinha Roy and Roy van der Weide, hereafter RW – recently proposed an interesting method to correct them. Their study covers a lot of ground, but we focus on the correction method for now. This method consists of "re-weighting" the CPHS observations in a way that brings the means of basic socioeconomic variables in line with independent, credible estimates of these statistics – for instance, education levels. The re-weighting technique, discussed below, builds on the notion of "maximum entropy" (max-entropy for short). RW's poverty estimates based on re-weighted CPHS data are now the official World Bank poverty estimates for India in recent years.

RW's method is certainly a step outward from the black hole. But how well does it work? We try to shed some light on this using Monte Carlo simulations. Briefly, we create an artificial under-representation of poor households in PLFS data by dropping various sets of observations, and then use the max-entropy re-weighting technique to 'correct' the bias. Then we compare the poverty estimates that emerge from this method with the 'true' poverty rates associated with the full PLFS dataset.

This simulation exercise, and a similar one based on CPHS data, suggest that RW's method falls significantly short of bridging the full gap between biased and unbiased poverty estimates.

Maximum entropy re-weighting

RW's correction method works as follows: first, they identify a set of socioeconomic variables (education, occupation, asset ownership, etc.) that are found in CPHS as well as other reliable surveys, in particular, PLFS and NFHS. They then adjust the household weights in the CPHS sample using the maximum entropy approach to make the weighted means of these variables match the corresponding means in NFHS or PLFS. Finally, the adjusted weights are used to calculate poverty statistics.

More precisely, monthly per capita consumption expenditure (MPCE) is our 'focus variable', and we wish to estimate two 'focus statistics' derived from its distribution: mean MPCE and the poverty head-count ratio. We observe that various correlates of MPCE (education, occupation, etc., hereafter the 'control variables') have very different means in CPHS data and independent surveys. The idea then is to correct the observed bias in the control-variable means in the hope that this will also correct the unobserved bias in the focus statistics. The correction is based on adjusting the sample weights using a weight calibration technique that minimises the 'distance' between the original sample weights and the adjusted weights subject to meeting the specified control means. Maximum entropy re-weighting is a version of this technique associated with one possible way of measuring that distance.

Clearly, this is an approximate (partial) correction for at least two reasons. First, it relies on a restricted list of observable control variables that are available in CPHS as

well as in independent, credible surveys. The list used by RW is reasonably comprehensive, but it could still miss important *unobserved* predictors of MPCE.

Illness and crop failures are two examples. Second, the focus statistics may depend on the *distribution* of control variables (indeed, their joint distribution) and not just on their mean. Correcting the means of control variables is not the same as correcting their joint distribution.

How well does the approximation work? Based on limited validation checks, RW seem to take the view that it works quite well for poverty estimation purposes, and that their adjusted weights even "… transform the CPHS into a nationally representative dataset". The simulation exercise below, however, suggests that this view may need re-examination.

Simulation exercise using PLFS data

We begin by testing the efficacy of RW's max-entropy re-weighting method on PLFS data for 2017-18. We start with the full PLFS sample, which may be treated as the 'universe' for our purposes, and calculate the focus statistics – mean MPCE and the poverty headcount ratio.

We then create artificially biased (or contaminated) samples by randomly dropping poor households from the full PLFS sample in four different ways. Using RW's method (with similar, though not identical, control variables), we adjust the household weights in these contaminated samples and then re-estimate the focus statistics. If the correction method works well, these adjusted focus statistics from the contaminated samples should be quite close to the full-sample values.

The contaminated samples are generated as follows:

i) Baseline contamination: Randomly drop 50% of households (HHs) in the lowest four MPCE deciles.

ii) Gradient contamination: Randomly drop 70%, 50% and 30% of households in the poorest, second poorest and third poorest MPCE decile respectively.

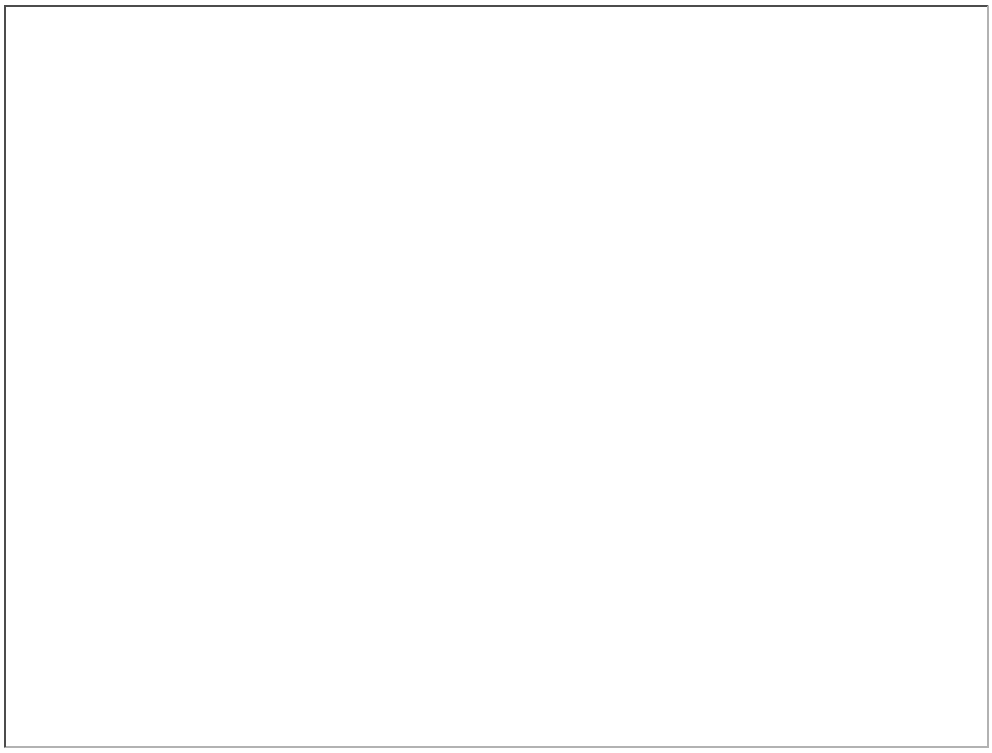
iii) Censored contamination: Drop all households in the poorest MPCE decile.

*iv) Scrambled contamination***:** First, randomly drop 50% of households in the lowest four MPCE deciles. Of what remains, randomly drop 20% of Muslim, Scheduled Caste (SC) and Scheduled Tribe (ST) households; then 20% of households with casual labour as primary income; and then 30% of households within lowest quartile of education levels.

Each variant begins with random pruning of households in the lower end of the MPCE distribution. In the last variant (Scrambled), we additionally drop households in other disadvantaged categories. These contamination scenarios may look a little radical, but we see no reason why they would necessarily exaggerate CPHS's ability to miss poor households.

We randomly generate 100 contaminated samples for each variant, except for the Censored variant where only one sample is possible. Thus, we have a total of 301 contaminated samples. For each of these samples, we begin by estimating the control means (that is, the means of our control variables) and focus statistics using unadjusted weights. Next, we adjust the household weights in the contaminated samples, using the max-entropy method, to bring the control means in line with their 'true' values (that is, their original values in the full PLFS sample).

Finally, we use the adjusted weights to estimate adjusted focus statistics. The results, averaged over all contaminated samples for each variant, are presented in Table 1.



Based on a poverty line of Rs. 972/month at 2011-12 prices (Rangarajan Committee), adjusted to 2017-18 prices using RBI's New Consumer Price Index Combined (2012 base year).

Notes: (i) In this table, 'adjustment' refers to the use of adjusted weights based on the max-entropy method. Adjusted control means are not shown because they are the same, by construction, as the 'true' control means. (ii) All control variables are household-level variables. The focus statistics take individuals as the unit and assign the same adjusted household-level weight to all members within a household. (iii) The standard errors of the adjusted focus statistics (not shown) are very small owing to the large sample sizes (more than 70,000 households in all contaminated samples).

We tried to remain as close as possible to RW's implementation of the method, but two key differences remain. First, RW include household assets among the

control variables, while we are unable to do so since PLFS does not collect asset data. Second, RW apply the max-entropy method at the state level (separately for rural and urban areas), whereas we apply it at the all-India level. In both respects, their adjustment method is likely to be more precise than ours. On the other hand, the artificial biases created in our contaminated samples may (or may not) be easier to repair than the CPHS biases, because they follow a simple pattern.

Simulation results

Looking at the first panel in Table 1, we see that contamination creates biases in the sample in expected directions: with fewer poor households, the contaminated samples also have a lower share of SC and ST households, a lower share of households doing casual labour, better-educated households on average, and so on. Interestingly, however, the deviation of control means from their 'true' means are very small in most cases, except in the Scrambled variant. This is perhaps a little surprising, since contamination involves dropping a large number of poor households and major deviations of focus statistics from their true values. In the Baseline variant, the headcount ratio drops by nearly 16 percentage points, but the control means barely change in most cases.

This observation helps to explain the fact that rectifying the control means (using the max-entropy method) does not make much difference to the focus statistics: after adjustment, mean MPCE and the headcount ratio are still quite close to their unadjusted values – except, here again, under the Scrambled variant. This suggests that the max-entropy method does not work very well, at least not with these sorts of contamination patterns and control variables.

The Scrambled variant presents a somewhat different picture. In this variant, the unadjusted control means

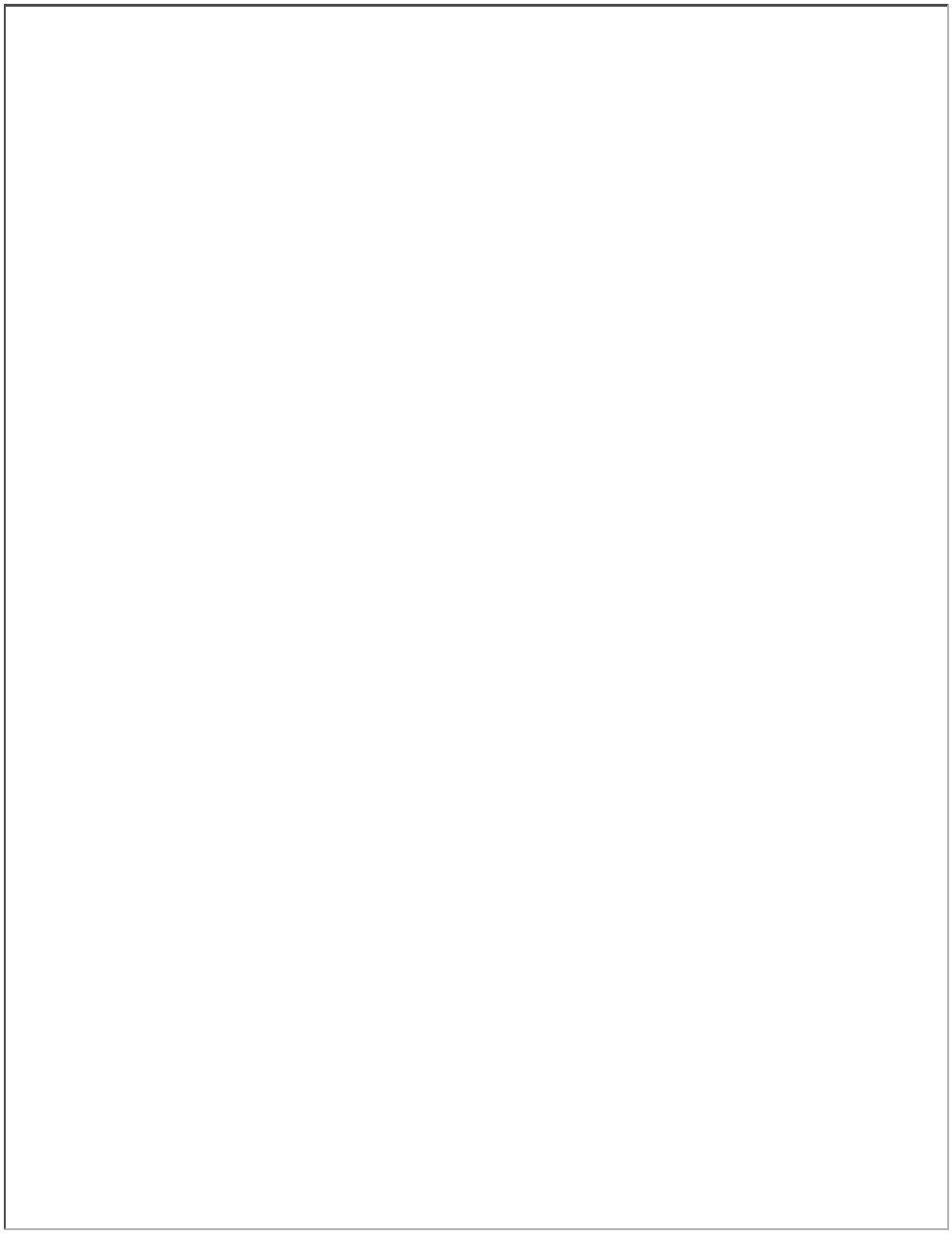
deviate quite sharply from their true values in many cases, and the adjustment method has more impact. In fact, it seems to work reasonably well for mean MPCE: the gap between the unadjusted and true values of mean MPCE reduces by 68% after adjustment. For the headcount ratio, however, the gap reduction is just 30%. Even in this variant, the max-entropy method is of little help for purposes of poverty estimation.

The reason why the adjustment method works better for the Scrambled variant is not difficult to understand. In this variant, contamination is partly based on dropping households at random among groups that are defined in terms of *control variables* rather than MPCE. The control variables, therefore, are well-placed to repair the bias. If we skip the first step in the Scrambled variant so that contamination is based *exclusively* on control variables, it turns out that the max-entropy method repairs almost 100% of the bias in focus statistics. This suggests that for some purposes, re-weighting may work quite well. For instance, if women are thought to be underrepresented in an opinion poll, *more or less at random*, then giving women more 'weight' (based on, say, Census estimates of the female-male ratio in the population) might correct the bias. Poverty estimation with CPHS data, however, is another matter.

Simulations using CPHS data

In principle, similar simulations can be done using the CPHS dataset itself. Even if it is not representative, nothing prevents us from treating the full CPHS dataset as the 'universe' for the purpose of simulations, as we did with PLFS. The advantage of using CPHS is that we can enlarge the list of control variables, and in particular, include household assets. We did so, with the same extended list of control variables as that used by RW.

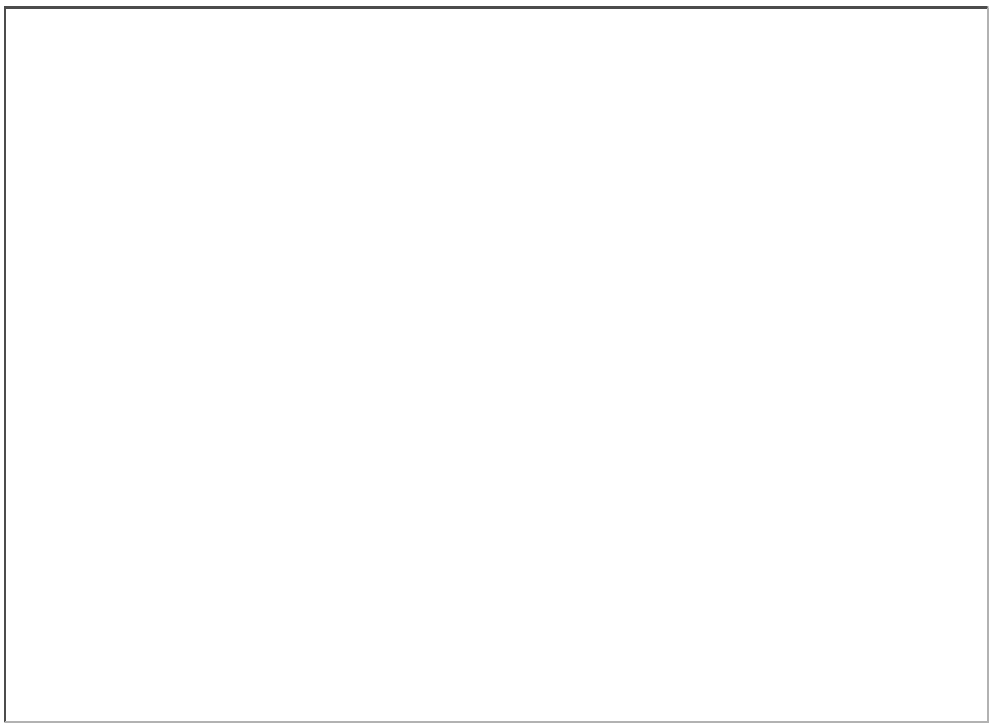
This may be regarded as a best-case scenario for the max-entropy method, not only because household assets are included among the control variables, but also because there is no need to fish out tentative 'target control means' from some independent dataset – we have appropriate targets within the full CPHS dataset itself. Even in this best-case scenario, however, the effectiveness of the max-entropy method is uncertain at best.



Based on a poverty line of Rs. 972/month at 2011-12 prices (Rangarajan Committee), adjusted to 2017-18 prices using RBI's New Consumer Price Index Combined (2012 base year).

Table 2 is similar to Table 1, with CPHS data for 2017

replacing data from PLFS 2017-18. Once again, the contaminated control means are strikingly close to the full-sample means, except in the Scrambled variant. In this best-case environment, the max-entropy method performs better than before. For mean MPCE, it is able to repair more than half of the bias away from the true mean in all variants, and as much as 69% of it in the Scrambled variant. For the headcount ratio, however, the repair effectiveness varies widely, from just 7% in the Censored variant to 62.5% in the Scrambled variant. Figure 1 conveys this variation, for both datasets.



Note: The figure presents the proportion (in %) of gap between true and estimated poverty head-count ratio that max-entropy re-weighting closes on average in each of the contamination variants.

We tried many other variants of the simulations presented here, without learning much more. The basic point remains that the effectiveness of the max-entropy method (in terms of percentage reduction in the difference between true and estimated focus statistics) is uncertain and varies a great deal between contamination variants. One pattern of interest is that the effectiveness of the method declines as the proportion of poor households being dropped increases. If that proportion is higher than

the true head-count ratio, so that the contaminated sample has no poor households at all, then the max-entropy method is virtually useless. As one colleague aptly put it, "you cannot re-weight yourself out of situations where there are no representatives of the group you are interested in".

Estimating poverty trends

Before concluding, we note that RW's main purpose was to estimate poverty *trends* beyond the 2011-12 CES round of the National Sample Survey (NSS). For this purpose, they attempted not only to correct biases in the CPHS sample (from 2015 onwards), but also to address the fact that CPHS expenditure data may not be comparable with NSS data for 2011-12.

They do this by imputing NSS-type consumer expenditure to CPHS households (instead of taking the CPHS expenditure figures at face value), based on two different approaches. In the first, NSS-type consumer expenditure is predicted using household characteristics for each CPHS household based on a model derived from NSS 2011-12 data.

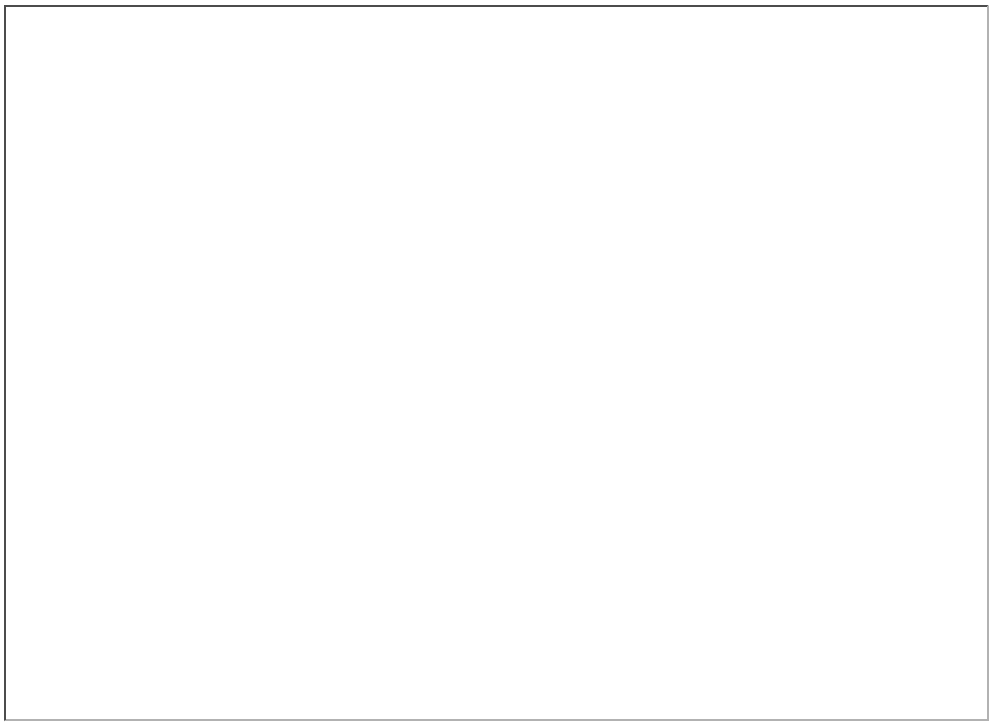
In the second, NSS-type expenditure is imputed directly from CPHS expenditure by imposing distributional assumptions and then using the method of moments. Both approaches add another layer of approximation to their estimation of poverty levels and trends. The results are interesting, but it is hard to guess how reliable they are. Their seeming precision – for example, in RW's summary statement that "extreme poverty is 12.3 percentage points lower in 2019 than in 2011" – is certainly difficult to reconcile with these multiple layers of approximation.

We can actually say a little more. RW's poverty estimates assume that the underrepresentation of poor households in

CPHS data is fully corrected by the max-entropy method. In fact, our simulation exercises suggest that the correction is only partial, and possibly far from a full correction. If so, RW's estimates are likely to understate poverty in 2015-19, and overstate poverty decline between 2011-12 and 2019.

There is another crucial problem with the estimation of poverty trends from CPHS data: the underrepresentation of poor households seems to have *grown over time* in recent years. Moreover, as noted earlier, repairing this gap tends to get harder as the gap gets larger – just like socks are harder to mend when the hole is wider. Thus, growing underrepresentation of poor households could easily create an illusion of poverty decline. Figure 2 presents one major hint of this. Quite likely, the precipitous decline of adult illiteracy between 2015 and 2019 (from 25% to 3% in just four years!) reflects the growing underrepresentation of underprivileged households in CPHS data.

Further, there is a striking co-movement between the trends in adult illiteracy and RW's adjusted poverty estimates. For all we know, the decline of adjusted poverty estimates in that period may well be an artefact of the growing bias in CPHS data.



Notes: i) Poverty estimates are taken from Roy and van der Weide (2022), Figures 15 and 17. ii) Adult illiteracy rates are estimated from unit-level CPHS data using original design weights (scaled by a non-response factor) for each of the three waves of each year and then averaged across waves.

RW's analysis ends in 2019, just before the Covid-19 crisis. What happened after that is unclear as things stand, but it almost certainly includes a sharp increase in poverty in 2020 (Azim Premji University 2021, World Bank 2022).

Concluding thoughts

A number of salutary lessons emerge from this exercise. First, it is possible for a sample to miss many poor households without this showing clearly in observable socioeconomic variables – including correlates of MPCE. This reinforces our earlier concerns about the credibility of CPHS data for poverty estimation: the observable biases may well be associated with a huge underrepresentation of poor households.

Second, the effectiveness of the max-entropy method depends critically on the nature of the biases it attempts to

correct. It may work reasonably well when the bias takes the form of households missing at random from observable groups. However, it is hard to know in advance (or even in hindsight) whether that is the case.

Third, it is one thing to correct mean MPCE, and quite another to correct a distribution-sensitive MPCE statistic like the headcount ratio. RW's implementation of the max-entropy method focuses on control means, but any aspect of the joint distribution of control and focus variables potentially matters.

Fourth, estimates of Indian poverty levels and trends based on the max-entropy method have an unknown and possibly wide margin of error. The method has been greeted with enthusiasm by the World Bank and others, but the proof of the pudding is still awaited.

Fifth, these estimates are likely to exaggerate the extent of poverty decline between 2011-12 and 2019. For one thing, the max-entropy method does not fully correct for the underrepresentation of poor households in CPHS data for 2015-19. For another, the underrepresentation of poor households was growing within that period, creating a spurious source of decline in poverty estimates. To paraphrase RW, poverty in India has probably declined over the last decade (up to the Covid-19 crisis), but how fast is anyone's guess.

Sixth, the concerns raised here potentially apply to a range of contexts, beyond just poverty estimation. Re-weighting on observables is often used to correct for selection bias. The assumptions required for this method to produce good results are reasonably clear from the technical literature, but they are sometimes overlooked in practical applications.

Finally, major doubts remain about the credibility of the

CPHS dataset as a nationally representative household survey. The nature of the CPHS biases, and the extent to which they can be corrected, are yet to be fully understood.

The silver lining is that another CES round is expected later this year. If it is released in good time without tinkering (a big if!), it may shed some light on poverty trends and also facilitate closer scrutiny of the max-entropy method. In the meantime, we remain largely in the dark.

The authors are grateful to Angus Deaton, Parikshit Ghosh, Thiago Scarelli, Sutirtha Sinha Roy and Roy van der Weide for insightful comments on an earlier draft; Sutirtha Sinha Roy also contributed helpful clarifications on max-entropy re-weighting.

Jean Drèze is an economist and Anmol Somanchi is an independent researcher.

This article was originally published on Ideas For India.