An Alternative Way to Capture Income in Household Surveys: a validation exercise for Young Lives – Peru

Javier Escobal[1]
March, 2009

OBJECTIVE

The objective of this exercise is to identify whether or not we can capture "Total Monetary Income" and how it is distributed between sources, using a simplified version of the questionnaire.[2]

The pilot aims to evaluate a shorter method to collect the necessary indicators regarding household income but with a procedure that if validated should take considerably less time. In addition if some sources of income are more difficult to be revealed, asking for the importance of that activity and recover the value afterward may allow us to go around the potential underreport of this income source.

Still, there is no "free lunch" in order to make this alternative way of capturing monetary income a workable alternative; we need to assure that the respondent is able to convey reasonably well how the monetary income of the household is divided between alternative sources.

We ask the interviewee to list all its income sources and to assess the relative importance of each source using 20 pebbles. After identifying the largest source of monetary income we proceed to ask in detail the amount obtained from that source. Having the distribution (percentage accrued from each source) and the amount for the most important income source, the total income and the income obtained from the other sources can be easily deducted. For example, if the respondent tells us that it has two income sources and most of the household monetary income comes from a non-agriculture wage activity, approximately 60% of their monetary income we can continue asking the detail of that activity without looking to the second income source, In our example the amount perceived through the first income source was 1,200 nuevos soles. If one divides this amount by the share of that source (1,200 divided by 0.6), the result (2,000) represents the total household monetary income. The second income source can be easily obtained (800 soles in our example).

VALIDATION PROCESS

To validate this alternative method, the survey used in the pilot collected the assessment of the relative importance of each source asking the respondent to allocate 20 pebbles between the different income sources that were declared. Once the

[1] With the assistance of Eva Flores
[2] We decide to aim only at total monetary income leaving non-monetary income as an extra module that will be captured in a separate module. From previous field experience we thought this was more sensible as valuing non-monetary income is much more complex.

respondent had pointed out the distribution of the 20 pebbles, the estimated percentages can be calculated. Next, information for the all income sources was gathered. In this way the instrument aimed also to identify the correct relative importance of each monetary income source, which can be compared against the estimated distribution obtained from the pebble exercise.

To validate this proposed methodology several statistical tests are proposed to evaluate similarities among the income distribution of the "true" (true distribution) collected and the distribution estimated based on the new methodology (estimated distribution).

1. Sample mean test (T-test). T-Test were used to check for significant differences in the average Total Monetary Income for the full sample and for the sample divided by urban/rural.

Table 1. Total Monetary Income: Sample differences between "True" and "Estimated" Values

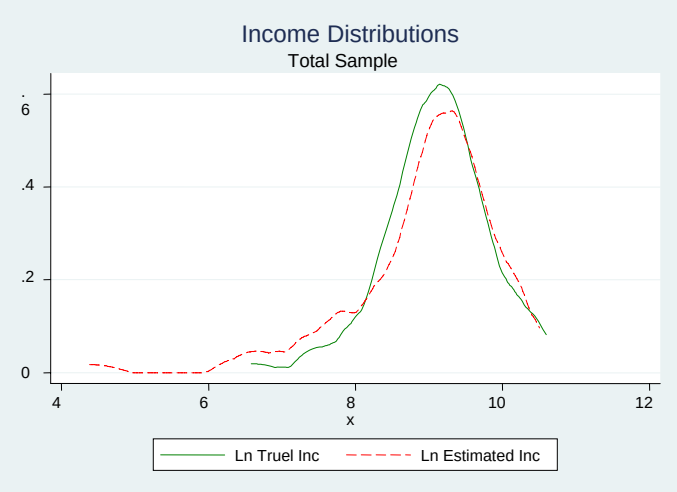
	True Income	Estimated Income	Difference (%)	P-value
Total Sample	11303	11267	-0.31%	0.98
By Area				
Urban	14504	14703	1.35%	0.93
Rural	8711	8486	-2.65%	0.87

Source: Peru YL Round 3 Pilot

As we can see, there is no significant difference between the mean of the true and the estimated total monetary income. Similarly, there is no difference when we split the sample by area of residence.

2. Graphic Analysis (Using Kernel distribution /K-Density). We expect to obtain similar shapes between both distributions. The logarithmic of each distribution has been used for this analysis in order to standardize the shapes and make them a bit more symmetrical.

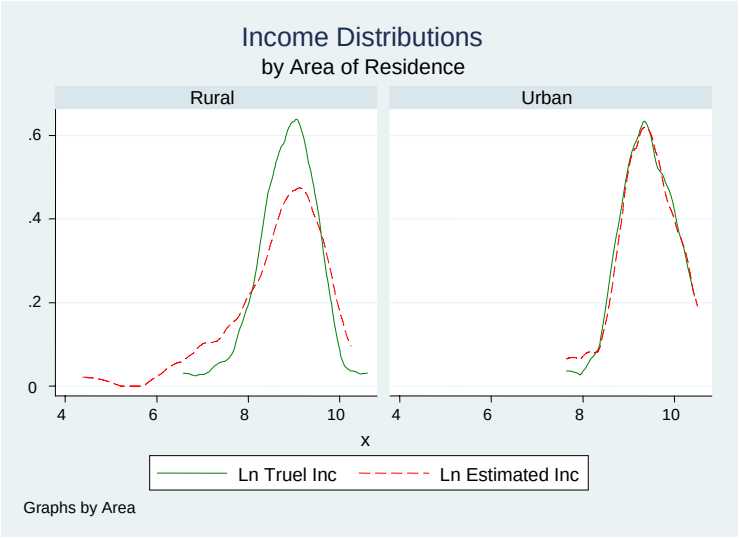
Figure 1. K-Density comparison between Income Distribution



Source: Peru YL Round 3 Pilot

As Figure 1 can show, both income distributions are quite similar. However it is evident that the lower tail of the distribution is quite different. If we test for equality of distributions using the standard Kolmogorov-Smirnov test we find that both distributions were not different at the 90% confidence level.

Figure 2. K-Density comparison between Income Distributions by area of residence



Source: Peru YL Round 3 Pilot

Figure 2 shows that for urban areas, there are no significant differences between both income distributions. However for rural areas there is a not negligible difference between the true distribution's peak, which is higher, and the estimated one. This tells us that the values of the true income distribution are more concentrate around the

average that the values around the estimated income distribution. We will comment on this at the end of this validation report. If we test for equality of distributions using the standard Kolmogorov-Smirnov test we find that both distributions were not different at the 90% confidence level for the urban subsample, but the distributions are different for the rural area.

3. Quintile Comparison. We divide both income measures into quintiles. Thus, we have two quintiles series: one for the True income distribution and other for the estimated one.

Table 2. Common Income Quintiles between households

	Estimated Income					
True Income	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Total**
Q1	**14**	1	1	0	0	**16**
Q2	1	**14**	0	0	0	**15**
Q3	0	0	**12**	3	0	**15**
Q4	1	0	0	**11**	3	**15**
Q5	0	0	2	1	**12**	**15**
Total	**16**	**15**	**15**	**15**	**15**	**76**

Source: Peru YL Round 3 Pilot

Table 2 shows how households are assigned into different quintiles according to both measures. In most of the cases, each household is allocated in the same quintile independent of the measure used. If one formally test the consistency of both distributions using Pearson's chi-squared test and Fisher's exact test we find that both distribution are not distinguishable at the 1% significance level.

Despite the fact the distribution matches very well, it is interesting to note that the average differences within each quintile are non significant in all cases except for the first quintile in which estimated income seems to underestimate the true mean monetary income. (See Table 3)

Table 3. Differences among Income distribution quintiles

Quintiles	True Inc.	Estimated Inc.	Difference	P-value
Q1	3660	2666	-37.3%	0.07 *
Q2	6873	6867	-0.1%	0.99
Q3	9442	9733	3.0%	0.36
Q4	12926	13269	2.6%	0.52
Q5	24122	24374	1.0%	0.92

Source: Peru YL Round 3 Pilot

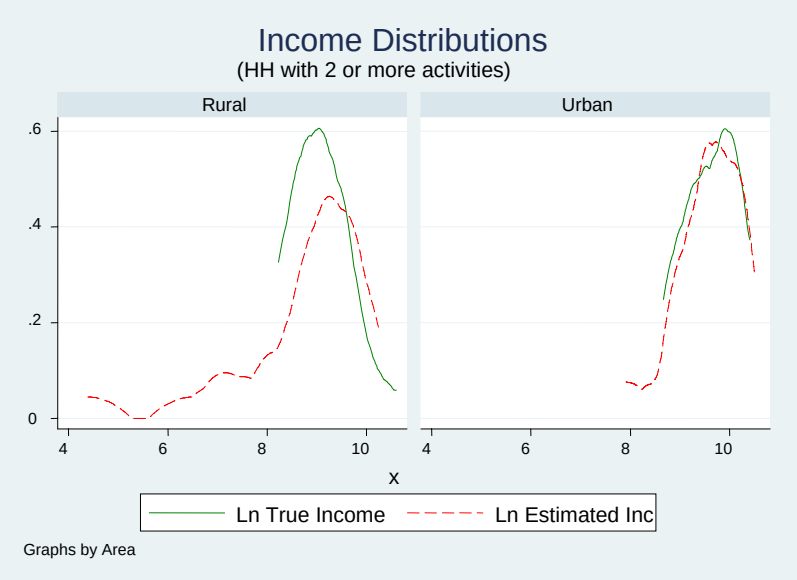
FINAL OBSERVATIONS

There is evidence that the shorter version of the questionnaire is able to capture reasonably well the true monetary income measure. On average both measures are indistinguishable and both measures rank equally well the households by income level.

These findings could be validating this new methodology. However there are some concerns regarding the use of this instrument to consider before applying this new instrument to the YL data collection.

1. The results are partially driven by the fact that a number of respondent have only one income source. If we consider only those that have 2 or more income sources the results are still reasonable, but a bit less robust. While the average income are not significantly different, the shape of both distribution differ to greater degree they did for the whole sample.

Figure 3. K-Density comparison between Income Distributions Households with 2 or more activities



Source: Peru YL Round 3 Pilot

Table 4. Sample Differences for Households with 2 or more activities

	True Inc.	Estimated Inc.	Difference	P-value
HH 2>= Act.	13547	13300	-1.86%	0.91
By Area				
Urban	17402	17494	0.53%	0.98
Rural	10707	10210	-4.87%	0.85

Source: Peru YL Round 3 Pilot

2. Is interesting to highlight that the new methodology seems to be a bit confusing to the mother/interviewee where he/she is not sure whether to report or not the income of those family members that do not contribute the household budget.

In some cases, the activity was reported but not the total amount perceived by it. In other ones, the activity was reported but not well represented in the marvel distribution despite representing higher income than the one pointed out by the interviewee.

CONCLUSIONS

The new methodology seems to work reasonably well when collecting income information especially in urban areas where household members do not have more than 2 activities monetary income generating.

Before applying it to the YL data collection instruments an extensive training is recommended to the fieldworkers in order to properly guide the interviewee during the marvel distribution to capture correctly the relative importance of the main activity.

The comparison in the timing of both alternatives is pending. This should allow us to have a better understanding around how much we are saving in terms of time. However a first calculation shows that we may save on average no more than 15 minutes. This may sound a bit too small, but it might be the case that substantive amount of time may be saved in those households with multiple income sources.

On the methodological side it may be useful to improve the validation exercise, for purposes that may go beyond YL[3], increasing the sample size, which may be generated if we embedded this type of instrument in a large survey and randomize who receives the short or long version of the module.

[3] For example, LSMS type of survey may benefit from this procedure, shortening their income module and been able to expand the questionnaire in other directions.