Rural Income Generating Activities Study:
Methodological note on the construction of income aggregates

Gero Carletto, Katia Covarrubias, Benjamin Davis,
Marika Krausova and Paul Winters[1]

I. Introduction

A major component of the Rural Income Generating Activities (RIGA) study was to construct comparable income measures from selected multi-purpose household surveys (see table in Annex I). The aim of the exercise was to provide annualized benchmark aggregates spanning four continents which, despite pervasive differences in the quality and level of information available in each survey, would be suitable for cross-country analysis. The objective of this document is to describe the methodology used in constructing the household income aggregates and their components included in the RIGA database.

Section II of this report discusses the general principles underlying the estimation of income aggregates. Section III describes the different components and breakdowns of our income measures and discusses some of the methodological issues behind their construction.[2] The final section of the document describes the data transformations and imputation procedures applied. More country-specific methodological issues and idiosyncratic adjustments are addressed in Annex IV.

II. General Principles for Estimating Income Aggregates

An income aggregate is a measure of household welfare that is based on the different sources of income – wage and non-wage, dependent and self-employed – that a given household can earn over a well-defined reference period. Although consumption-based money metric measures are more commonly used in welfare analysis since they are considered more accurate and easier to measure in a typical household survey, income-based measures and their components, particularly if properly measured, have an important function, as they allow for the meaningful analysis of a household's livelihood strategies and poverty. Set up as a monthly or annual indicator, the income aggregate is reported as an average net income figure. Our definition of income closely follows the definition given by the International Labour Organization (ILO) (Box 1).[3]

[1] Gero Carletto is a Senior Economist in the Development Research Group at the World Bank (gcarletto@worldbank.org); Katia Covarrubias and Benjamin Davis are Economists in the Agricultural Sector in Economic Development Service (ESAE) of FAO (Katia.Covarrubias@fao.org and Benjamin.Davis@fao.org); Marika Krausova is a JPA at the World Bank (mkrausova@worldbank.org); Paul Winters is an Associate Professor at American University (winters@american.edu).

[2] The disaggregated components of the principal categories are discussed in the Annex IV, which provides detailed information on each individual survey's income aggregate

[3] Source: ILO, Resolution I "Resolution concerning household income and expenditure statistics"
Available from: http://www.ilo.org/public/english/bureau/stat/download/res/hiestat.pdf

Box 1: ILO Definition of Income

The ILO's *Resolution concerning household income and expenditure statistics* defines income as follows: "Household income consists of all receipts whether monetary or in kind (goods and services) that are received by the household or by individual members of the household at annual or more frequent intervals, but excludes windfall gains and other such irregular and typically onetime receipts. Household income receipts are available for current consumption and do not reduce the net worth of the household through a reduction of its cash, the disposal of its other financial or non-financial assets or an increase in its liabilities." (ILO, 2003).

For the purposes of constructing an income aggregate, we therefore consider as income receipts those that (i) recur regularly; (ii) contribute to *current* economic well-being; and (iii) do not arise from a reduction in net worth. These three criteria are embodied in each of the components of income; as such, irregular payments such as lottery earnings or inheritances; investments and savings and the value of durables are not included in our definition and measure of income.

In order to create income aggregates that are comparable across countries and over time, we apply the following criteria in the estimation of our income measures:

→ All aggregates are estimated at the *household level*.
→ All income and expenditures are *annualized.*
→ All income components are *net of costs*.
→ *Purchases and sales of durable goods, investments* and *windfall gains are excluded* from household income and expenditure calculations since these are not transactions undertaken regularly by households and can result in the significant over- or under-stating of permanent income.
→ All aggregates are reported in *local currency units*.
→ For each survey, only the *rural sample* is used.

III. Components of Aggregate Income

Although the construction of the income aggregates takes into consideration all sources of income reported by the household in the survey, some aggregation of the different sources is necessary and more functional for our purposes. At the least disaggregated level, we can define two categories of income: wage and non-wage. Wage income includes all activities undertaken by persons in which the income received is in the form of a wage paid out by an employer; in other words, wage income includes earnings from dependent activities. Non-wage income is a broader category referring to non-dependent income and includes household crop and livestock production, self-employment earnings, transfer income and other non-labor income sources.

We further disaggregate our income measures into the following six principal categories: wages (also separated into agricultural and nonagricultural wages), self-employment, crop production, livestock production, transfers, and other income. The breakdown proposed follows standard practices and best reflects the specific objective of the RIGA study. In the remainder of this section, we describe in detail each of these categories. The full classification is summarized in

Table 1. Any further country-specific disaggregation or classification is fully described in Annex IV.

Wage Income

Wage income consists of all income received in the form of employee compensation either in cash or in kind. Since it is common for household members to simultaneously hold more than one job or change jobs throughout the survey reference period, all income from primary, secondary and any additional jobs held in a 12-month period is considered to account for individuals' pluriactivities.

Wage employment income is further disaggregated by industry. The classification is based on the United Nations' International Standards Industrial Classification of All Economic Activities (ISIC). As the classification of industries changes over time, the most appropriate revision of the ISIC classification standards is chosen based on the year of the survey.[4] As presented in Table 2, industries are grouped into ten principal categories: (1) Agriculture, Forestry and Fishing; (2) Mining; (3) Manufacturing; (4) Utilities; (5) Construction; (6) Commerce; (7) Transportation, Communications and Storage; (8) Finance and Real Estate; (9) Services; and (10) Miscellaneous. The wage component is further disaggregated into skilled and unskilled; some observations are forcibly classified as unknown when not enough information is provided by the individual in a specific survey. The distinction among these three sub-categories is based on the ILO's International Standard Classification of Occupations (ISCO-88)[5] and, sometimes also on country-specific documentation.

Furthermore, the estimation of total wage employment income is divided between agricultural and nonagricultural wage income (as determined by the ISIC industry categories). The total wage income aggregate is not created and not included in the final dataset: given the objective of the RIGA study, the two sub-components of wage income will always be analyzed separately. Obviously, if needed for other types of analyses, constructing total wage income from the database is trivial.

Self-employment

The self-employment category includes the income earned from all non-farm household enterprises. Self employment income includes all cash and in kind earnings and non-durable, recurrent expenditures for all non-farm businesses operated by any member of the household over a 12-month period. All expenditures for equipment and machinery purchases and other such investment expenses are not included in the aggregate. It must be noted that the purchase of raw materials, although clearly not an investment, are often procured in bulk; consequently, in some instances, they had to be treated differently from other expenses. Annex III elaborates on the issues encountered with this expenditure category and the solutions applied in some countries.

Income from household enterprises can also be decomposed into industries according to the ISIC nomenclature. It should be noted that the first industry category for self-employment represents agricultural processing activities, which should be distinguished from agricultural production activities. Further, when the information is available, the total income is weighted by the share

[4] The classification system can be found at http://unstats.un.org/unsd/cr/family1.asp.

[5] These classifications can be found at http://www.ilo.org/public/english/bureau/stat/isco/isco88/major.htm.

of the enterprise owned by the household since non-farm enterprises may be owned by more than one household.

Crop Production

The estimation of crop income accounts for the sale of crops, crop by-products, sharecropping earnings, the consumption of homegrown crops, net of all expenditures related to these activities, such as seed and fertilizer purchases and the hire of farm labor.

For the valuation of own crop consumption, two different approaches are followed depending on the availability and quality of the data in each country. In the first approach, own crop consumption is calculated based on the quantities consumed of own- produced crops as reported in the agricultural module of the household questionnaire. In cases where the quantities of own consumption are not specifically asked in the questionnaire, this magnitude is estimated for each crop as a residual by subtracting the total amount sold, bartered, lost or used as an input (such as for seed or fodder) from the total amount harvested. The second approach relies on the food expenditure section of the questionnaire to estimate the quantities of food consumed from own production.

In both approaches, to estimate the value of the reported quantities of own-consumption, unit values are imputed using the following procedure:[6] median unit values are estimated from crop sales for every crop-unit combination at the cluster and higher geographic (e.g. district, region, etc) levels. If a unit value can not be adequately estimated for a specific crop using the above method at any of the different levels of geographic clustering, the price questionnaire is used to estimate median unit prices at the cluster or higher geographic levels, which are then applied to quantities consumed by the household. Only in those few occurrences in which no imputation is possible using any of the above procedures, median unit values are estimated using the self-reported values of consumption by households in the cluster or at higher geographic levels, if necessary. The two approaches generally yielded qualitatively similar results. The second approach was followed for most countries, among other things, to improve the comparability with consumption-base welfare measures; in a few cases, however, quantities from the production side were used as they were deemed more accurate.[7]

Livestock

The livestock income category includes income from the sale and barter of livestock, livestock by-products (i.e. milk, eggs, honey etc.), net of expenses related to livestock production and livestock purchases, plus the value of household consumption of own livestock and livestock by-products. The values of own consumption are estimated based on the food expenditure section (or Food Booklet/Diary) of the questionnaire. In cases where this information is not available in that module, the consumption amount is obtained from the agricultural module.[8] Under both

[6] Deaton, A. and Zaidi, S., *Guidelines for Constructing Consumption Aggregates for Welfare Analysis,* LSMS Working Paper No. 135, The World Bank, Washington, DC, 2002

[7] The first approach was preferred for the following surveys: Bangladesh 2000; Indonesia 1992 and 2000; Vietnam 1992-93.

[8] Livestock own consumption reported in the agricultural module would also be used to maintain consistency with the approach taken for crop income: in cases where crop own consumption is taken from the agricultural module (see Footnote 7 for the list of surveys), the same approach is followed for livestock own consumption.

circumstances, the approach for valuing own consumption is the same as the one followed for the valuation of crop own-consumption (see above).

Transfers

This category refers to both private and public transfers received by the household, both in cash or in-kind[9]. Private transfers primarily refer to incoming remittances, but they can also include benefits from private organizations and/or associations as well as forms of gifts and contributions not associated with the performance of a job or the provision of a service. Public transfers are divided into state-funded pensions and social benefits, which include welfare support, maternity benefits, and educational transfers. Pensions and social benefits reported in this section do not include benefits received from employers, as those are included under the wage employment component.

Other Sources

All other non-labor income components that do not fall into the previous five categories are considered in this last grouping. Other income consists of gross non-labor income from farm land rental, non-farm real estate rental, rental of owned assets, and other miscellaneous sources not specified in the questionnaire. Some caution should be exercised when comparing this income category across countries because some surveys do comprise different sources. However, in the vast majority of cases, it only represents an insignificant portion of total income.

Table 1

Principal Income Categories	Disaggregated Components
Wage Employment- Agricultural	Agriculture & Fishing, Unskilled Agriculture & Fishing, Skilled Agriculture & Fishing, Unknown
Wage Employment- Nonagricultural	Mining- Unskilled Mining- Skilled Mining- Unknown Manufacturing- Unskilled Manufacturing- Skilled Manufacturing- Unknown Electricity & Utilities, Unskilled Electricity & Utilities, Skilled Electricity & Utilities, Unknown Construction, Unskilled Construction, Skilled Construction, Unknown Commerce, Unskilled Commerce, Skilled Commerce, Unknown Transport, Storage and Communication,

[9] Outgoing transfers, which are usually private, are not considered since it is not possible to determine if (1) the money is to be returned, in which case it should be classified as a loan, which is not accounted for in an income aggregate; or (2) the transfer is to be used for consumption or investment, the latter of which is also not considered in the income aggregate since it would not be a regular expenditure on the part of the recipient household.

	Unskilled Transport, Storage and Communication, Skilled Transport, Storage and Communication, Unknown Finance, Insurance and Real Estate, Unskilled Finance, Insurance and Real Estate, Skilled Finance, Insurance and Real Estate, Unknown Services, Unskilled Services, Skilled Services, Unknown Other, Unskilled Other, Skilled Other, Unknown
Self Employment	Agricultural Processing Mining Manufacturing Electricity & Utilities Construction Commerce Transport, Storage and Communication Finance, Insurance and Real Estate Services Other
Crop Production	*Total net income.*
Livestock Production	*Total net income.*
Total Transfers	Private Transfers Public Transfers Pensions Social Transfers
Other Income Sources	Non-farm Rental Income Farm Rental Income Other, Miscellaneous Income

IV. Data Transformation and Calculation Procedures

Survey data vary considerably across countries in terms of quality. Consequently, to enhance comparability of the different income components, we implemented comprehensive data cleaning and imputation procedures at different levels of aggregation.

Outlier Checks

Raw data are assumed to have been already checked for consistency and outliers; therefore, in creating the income aggregate components, checks for outliers are performed only on constructed variables. Such transformations include, but are not limited to: annualizing the original variable; collapsing the data from person to household level; and multiplying the price received by the quantity sold.

The outlier check is performed by dividing the variable by one to two relevant subgroups, for example, plot size and crop code for checking the values of crop sales, or industry sector for wage employment income. For variables without a logical sorting variable and for variables in which the logical sorting variable does not include a sufficient number of observations in the sorting category,[10] an administrative unit variable, such as district, is used as the sorting variable. Outliers are identified as values greater or less than three standard deviations from the median value of the variable for that specific group. The specific syntax used for flagging and imputing outliers is described in Annex II.

After completing the first check for outliers, the data is summed at the household level (if not already done so) and final aggregate variables are computed. The outlier check is repeated on these variables using an administrative or geographic unit as the categorical sorting variable. For all outlier checks, zeroes and missing values are excluded from the computation of the median, standard deviation and identification of outliers in order to achieve accurate imputations. This ensures the medians and standard deviations are not skewed by zeros and that households with missing values are not erroneously assigned values.

One final outlier check was imposed after the construction of the income aggregate to deal with extreme income *shares* that arose following the aggregation of all income components. After estimating the shares of the seven principal income categories (agricultural wages, non-agricultural wages, crop, livestock, self employment, transfer and other) observations whose share was greater than or less than 3 (indicating a percentage share of +/- 300%) were dropped from the aggregate.

Imputations

One major problem in working with survey data is the presence of missing and misreported values. Missing values can be identified in three categories, following the NAS[11]: under-coverage, unit non-response and item non-response. Misreported/miscoded values arise either when the answer provided by the respondent does not logically correspond to the question, or in the data entry stage of the survey. In these instances, recoding and imputation procedures may be necessary.

Recoding. As it is always the case, some respondents choose not to answer or are unable to answer some questions in the questionnaire and as a result the answer is coded as missing. In other cases, the missing value is a natural result of the skip pattern in the questionnaire: for example, if a respondent answers “zero” when asked for the quantity of milk sold, a subsequent question regarding the price at which the milk was sold would be skipped and coded as missing. Missing values of this sort are preserved and not recoded. However, if the amount of milk sold is reported but the price is reported as missing, the price must be recoded to a logical, imputed

[10] For this study, outlier checks must be performed by groups of 50 observations minimum in order to ensure that outliers are accurately identified. This criteria was adhered to in most cases; however, in some instances in which too few households participate in a given activity, the groups by which outliers are checked may be somewhat smaller.

[11] National Academy of Sciences (1983) in: Madow, Olkin and Rubin, eds., *Incomplete Data in Sample Surveys*. New York: Academic Press, Vol. 1-3.

value, a procedure described below. For the purpose of recoding is thus important to first distinguish between missing values resulting from intentional or unintentional omission and missing values caused by the skip pattern of the questionnaire

Imputation Procedures. Following the recoding procedure, we use imputations to (1) correct values that are erroneously coded or reported and (2) appropriately assign values where they have been intentionally or unintentionally omitted. For the first case, outlier checks, as described above, are one form of imputations undertaken to deal with extreme values. Also corresponding to the first objective is the reassigning of miscoded values. For example, if the answer to a filter question (e.g. "Did you sell...?") is coded "No" yet all subsequent questions (which would correspond to quantity sold, price received, etc.) were answered, indicating the answer to the filter question should, in fact, be "Yes", then the answer to the filter question would be modified. This kind of imputation requires a meticulous analysis of each questionnaire module and the answers provided.

The second case for which imputations are used to replace missing values corresponds to estimating prices for variables that represent income or expenditures but are not given a monetary value or do not contain any information on price per unit. To value these income sources, we apply the price imputation approach proposed in Deaton and Zaidi (2002).[12] An example of a variable that would require a price imputation is bartered crops. In many countries, the questionnaire may ask for the quantity of crops bartered, but not their monetary value, which is necessary information if an income value is to be estimated from this transaction. In order to do so, a unit value can usually be obtained from information on the quantity of crop sold and amount received from similar types of sales. A median unit value can than be imputed to value the bartered crops. As shown in the Stata codes found in Annex II, for each crop and unit, a median value is calculated at different levels of the sample: cluster, district, region, urban/rural and across the whole sample at the crop-unit level. In the second step, the amount of bartered crops is multiplied by the imputed price to assign bartered crops monetary value.

Annualization.

The survey questionnaires collect data using different reference periods, i.e. day, week, month or year. In order to aggregate these variables, we first transform into annual terms those values reported for other reference periods. Annualization is then simply a matter of multiplying the amount received by the corresponding number of times it was received (or if dealing with expenditures, the amount spent by the number of times the expense was made). This means that "annualized" figures represent the total revenues received and expenses made by the household over the course of one year (generally, the year preceding the survey) and account for the frequency at which income is received and expenditures are made. When such specific information is not available to annualize data, we assume 313 working days (6 working days per week), 52 weeks and 12 months per year.[13]

[12] See footnote 6 for the citation.

[13] This assumption is imposed so long as it is relevant to the income source. If a household reports an income that should not be annualized in this way (e.g. a transportation subsidy given to students in the household would have to consider that during several months of the year, students do not attend school), then the annualization takes into consideration the context of the income received.

ANNEX I: List of surveys in RIGA database

Table 1

Country	Name of Survey	Year of Survey
Africa		
Ghana	Ghana Living Standards Survey- Round Two	1992
Ghana	Ghana Living Standards Survey- Round Three	1998
Madagascar	Enquête Permanente Auprès des Ménages	1993-1994
Malawi	Integrated Household Survey-2	2004-2005
Nigeria	Living Standards Survey	2004
Asia		
Bangladesh	Household Income-Expenditure Survey	2000
Indonesia	Family Life Survey- Wave 1	1992
Indonesia	Family Life Survey- Wave 3	2000
Nepal	Living Standards Survey I	1995-1996
Pakistan	Integrated Household Survey	1991
Pakistan	Integrated Household Survey	2001
Vietnam	Living Standards Survey	1992-1993
Vietnam	Living Standards Survey	1997-1998
Eastern Europe		
Albania	Living Standards Measurement Survey	2002
Albania	Living Standards Measurement Survey	2005
Bulgaria	Integrated Household Survey	1995
Bulgaria	Integrated Household Survey	2001
Latin America		
Ecuador	Estudio de Condiciones de Vida	1995
Ecuador	Estudio de Condiciones de Vida	1998
Guatemala	Encuesta de Condiciones de Vida	2000
Nicaragua	Encuesta de Medición de Niveles de Vida	2001
Panama	Encuesta de Condiciones de Vida	1997
Panama	Encuesta de Condiciones de Vida	2003

ANNEX II: Stata Syntax

Outlier Checks.
When checking for outliers, each variable that is identified as an outlier is flagged. The flags are named like the variable being checked for outliers, plus an “m”(e.g. the flag for “exp” is “expm”). In our sample Stata syntax (below), **expm** is either a missing value or, if the value in **exp** was an outlier, the value to be used as the imputation. Besides each variable and its flag, the dataset includes the imputed variable, **expimp** in this example. The variable **expimp** contains the original values when no imputations are necessary and the imputed values (obtained from **expm**) when an outlier is identified.

In order to flag and impute outliers in the variable **exp,** the variables **expm** and the variable **expimp** are created in lines (2) and (3). At this stage, the variable **expm** consist only of missing values and the variable **expimp** is equal to the original variable, **exp**. Lines (4) and (5) calculate the median and standard deviation of crop expenditures by the relevant category of expenditures (itemcd). Line (7) flags any outliers in **exp** by replacing in **expm** the missing value with the median of **exp**. In the last line of the program (8), the original value (in **expimp**) is replaced by the median value (in **expm**) if an outlier is identified.

```
(1) sort region
(2) gen expm = .
(3) expimp = exp
(4) by itemcd: egen expme = median(exp) if expm != . & expm != 0
(5) by itemcd: egen expsd = sd(exp) if expm != . & expm != 0
(6) replace expsd = 0 if expsd == .
(7) replace  expm = expme if  !(exp >= (expme - 3*expsd) & exp <= (expme + 3*expsd)) &  expm == . & expm != . & expm != 0
(8) replace expimp = expm if expm != .
```

Price Imputations. The following lines of Stata codes demonstrate the approach used to estimate a standard set of prices by which to value income and expenditures that were reported in quantities rather than monetary units. Lines 1 through 6 demonstrate how crop-unit prices are obtained for each cluster, district, region, and urban/rural area. Lines 7 through 11 indicate how the prices are then multiplied by the quantity variable in order to obtain a monetary equivalent of the quantity bartered, as presented in this example.

```
(1) gen price = salesvalue / soldquantity
(2) bysort crop unit urban region distr clust: egen pricedata1 = median(price)
(3) bysort crop unit urban region distr: egen pricedata2 = median(price)
(4) bysort crop unit urban region: egen pricedata3 = median(price)
(5) bysort crop unit urban: egen pricedata4 = median(price)
(6) bysort crop unit: egen pricedata5 = median(price)

(7) bysort crop unit : gen barter = (bartquantity * pricedata1)
(8) bysort crop unit : replace barter = (bartquantity * pricedata2) if barter == .
(9) bysort crop unit : replace barter = (bartquantity * pricedata3) if barter == .
(10) bysort crop unit : replace barter = (bartquantity * pricedata4) if barter == .
(11) bysort crop unit : replace barter = (bartquantity * pricedata5) if barter == .
```