WIEGO Women in Informal Employment Globalizing and Organizing

WIEGO Statistical Brief Nº 13 May 2014

Measurement of Employment-Related Income: Concepts, Data Sources and a Test of Methods

Farhad Mehran[1]

Introduction

Most people work to earn an income. Income from employment makes up a large proportion of household income and provides the basic resources of workers to maintain the welfare of themselves and members of their family. Data on income from employment serve a variety of economic and social purposes. In general the data provide information for analyzing the income-generating capacity of different economic activities and for analyzing the economic wellbeing of persons on the basis of the employment opportunities available to them. Particular uses of the data include:

- design, implementation and assessment of employment promotion policies, which aim at creating and developing employment that provides adequate income
- analysis of the informal sector for employment and income generation, and measurement of related concepts such as underemployment, inadequate employment situations, low pay and the working poor
- understanding the changes in employment patterns and remuneration practices that have taken place in different countries
- assessment of the impact of specific economic and social policies such as assistance to agricultural workers and of access of particular workers such as women and rural-urban migrants to the labour market
- appraisal of the consumption capacity of workers and their level of employment-related welfare
- formulation of fiscal policies on adjustment of income taxes and social security contributions and the redistribution of income and social security benefits
- input to other bodies of statistics, in particular to the compilation of labour and national accounts

The international standards on the measurement of employment-related income recognize different sources of data, including labour force surveys (LFS) and household income and expenditure surveys (HIES). Other sources of data include establishment surveys, administrative records (such as income tax and social security records), informal sector surveys, agricultural surveys, surveys of small economic units and population censuses.

CONTENTS

[1] Farhad Mehran is the former Director of the Bureau of Statistics, International Labour Office, Geneva. He is currently an independent consultant on labour statistics.

The ILO-WIEGO project[2] aims at reviewing the various possibilities of measuring income from employment with existing sources and, in particular, assessing the feasibility and accuracy of combining the employment data of labour force surveys and the income data of household income and expenditure surveys, taking advantage of each source while minimizing their disadvantages. Such linkage of existing sources would provide cost-efficient means of joint measurement of employment and income where each component is derived from the most appropriate source. One application of the results would be the calculation of indicators such as the working poor and low earnings not only for formal employees but also for informal employees and own-account workers.

Within the framework of the project, data from the quarterly labour force survey of Iran (2011) and the annual household income and expenditure survey of Iran (2011) have been analyzed and linked using alternative statistical matching techniques. The purpose of the present final report is to put together the various elements in the form of preliminary guidelines on the various possibilities in measuring income from employment using either labour force surveys alone or in combination with household income and expenditure surveys.

The report is organized as follows. In the next section, the international concepts and definitions of employment-related income are briefly described and the basic data sources reviewed. Proposals are then made for measuring income from employment in labour force surveys either in the core survey or as a special added module. The measurement of income from employment in household income and expenditure surveys is considered next. This paves the way to discuss the methods for combining the two sources for the joint measurement of employment and income from employment. The main conclusions are reported in the final section.

Income from Employment[3]

Income from employment consists of the payments, in cash, in kind or in services, which are received by individuals, for themselves or in respect of their family members, as a result of their current or former involvement in paid or self-employment jobs. Income from employment excludes income derived from other sources such as property, social assistance, transfers, etc., not related to employment.

Because of the differences in the nature of income generation in self-employment and paid employment jobs, the definition of "employment-related income" distinguishes between paid employment and self-employment. In the case of paid employment, the concept is defined in terms of its components, namely, remuneration in cash and in kind, profit-related pay and current receipts of employment-related social benefits.

Income Related to Paid Employment

Income related to paid employment includes direct wages and salaries in cash for time worked and work done, remuneration for time not worked, cash bonuses and gratuities, and remuneration in kind and services, profit-related pay and employment-related social security benefits.

It excludes annuities, remittances, gifts, etc. as well as family allowances[4] and other social security benefits or assistance and indemnities or allowances in cash and in kind paid by the employer purely to cover the employees' cost of work-related expenditure,[5] and employers' contributions to social security funds, insurance or other institutional units responsible for social insurance schemes.

Income Related to Self-Employment

In the case of self-employment, the concept is defined as the difference between gross value of output and operating expenses. Income related to self-employment is the profit or share of profit generated by the self-employment activity. It can be calculated as the difference between the value of gross output of the activity and the operating expenses. Income from self-employment includes remuneration received by owner-managers of corporations and quasi-corporations, where relevant. It also includes employment-related social security benefits received by self-employed persons.

Both paid and self-employment related income is measured in surveys in terms of gross values. Net income related to paid employment may be obtained by deducting employees' direct taxes, union dues and other obligations. Similarly, net income related to self-employment is obtained by deducting from gross income related to self-employment, personal direct taxes and other employment-related obligations.

---

[2] The project has been funded by the research network Women in Informal Employment: Globalizing and Organizing (WIEGO). It was conducted in collaboration with Nomaan Majid, senior economist, and Pedro Oluwaseum, research assistant, Economic and Labour Market Analysis Department, International Labour Office (ILO).

[3] The term "income from employment" is used here in the same sense as the more exact term "employment-related income" adopted by the International Conference of Labour Statisticians, *Resolution concerning the measurement of employment-related income*. 1998. Geneva: 16th ICLS.

[4] For example, food stamps, government or community housing, free health-care assistance, etc., when paid by social security schemes or the State without regard to the employment status (e.g. under universal schemes with or without means tests).

[5] For example, tools, equipment, clothing or footwear used exclusively or mainly at work, special housing and meals necessitated by exceptional working conditions, reimbursement of business travel and accommodation expenses, medical examinations or health checks required because of the nature of the work, etc. However, when indemnities take the form of cash payments over and above the reimbursement of expenses incurred by employees, such payments should be considered as income related to paid employment.

There are other particular concepts of income from employment, each serving different purposes, e.g., wages. "Wages" is a narrower concept than employment-related income and is part of income related to paid employment. As shown in the following diagram, wages itself can be viewed in different ways: (a) wages as price of labour, called "wage rate"; (b) wages as income to the worker, called "earnings"; and (c) wages as cost to the employer, called "labour cost".

Wage Rate

Wage rate is the rate of pay per period of time or per unit of production for an employee on a given job. It includes basic wages, cost-of-living allowances and other guaranteed and regularly paid allowances. It excludes over-time payments, bonuses and gratuities, family allowances, and other social security payments by employers, payments in kind, supplementary to normal wage rates (12th ICLS 1973).[6]

Earnings

Earnings is the remuneration in cash or in kind paid to employees, as a rule at regular intervals, for time worked or work done together with remuneration for time not worked such as annual vacation and other paid leave or holidays. It includes direct wages and salaries, remuneration for time not worked, bonuses and gratuities, and payments in kind. It excludes employers' contributions to social security and pension schemes, severance and termination pay (12th ICLS 1973).[6]

Labour Cost

Labour cost is the cost incurred by the employer in the employment of labour. It includes earnings, employers' social security expenditure, cost of vocational training, cost of welfare services, taxes and fees regarded as labour cost, and other expenditures such as transport, clothing and recruitment (11th ICLS 1966).[7]

The relationship between the different concepts of wages with income from paid employment is schematically presented in the following diagram.[8] The diagram also shows the relationship with the concept of compensation of employees used in the system of national accounts (SNA).

Figure 1: Relationship Between the Different Concepts

<table>
<tr><td colspan="5">Income from paid employment</td></tr>
<tr><td rowspan="7">Benefits received by employees from social security schemes</td><td colspan="4">Labour costs</td></tr>
<tr><td colspan="3">Compensations of employees</td><td rowspan="2">Cost of vocational training</td></tr>
<tr><td colspan="2">Earnings</td><td>Employers' social contributions paid directly to employees</td></tr>
<tr><td>Wage rates</td><td rowspan="2">Direct wages and salaries corresponding to periods outside normal hours: "premium" pay for: – overtime, night work, shift work, holiday work, incentive pay</td><td rowspan="2">Employers' contributions to social security and pension schemes</td><td rowspan="2">Other labour cost</td></tr>
<tr><td>Direct wages and salaries corresponding to normal hours of work
– for time worked, for time not worked, e.g. vacations, sickness, study, ...</td></tr>
<tr><td>Guaranteed and regular allowances:
– cost of living, housing and rent allowances, ...</td><td>Regular bonuses and gratuities</td><td rowspan="2">Irregular bonuses and gratuities</td><td rowspan="2">Taxes regarded as labour cost</td></tr>
<tr><td>Wages and salaries in kind</td><td>Family allowances paid directly by the empolyer</td></tr>
</table>

[6] International Conference of Labour Statisticians. 1973. *Resolution concerning an integrated system of wage statistics*. Geneva: 12th ICLS.
[7] International Conference of Labour Statisticians. 1966. *Resolution concerning statistics of labour cost*, Geneva: 12th ICLS.
[8] Reproduced from slide prepared by Adriana-Mata Greenwood, ILO Department of Statistics, Geneva.

Wage rate is the narrowest concept, often used for collective agreements, arbitral awards or other wage-fixing decisions, which generally specify minimum rates for particular occupations or groups of workers. Labour cost is a broader concept used in establishment surveys to measure the average cost of labour per hour worked (hourly labour cost) or the average cost of labour per unit of output (unit labour cost).

In the LFS, income from employment is generally measured in reference to the concept of earnings as it most closely represents income from employment as perceived by the worker. Also because it excludes irregular bonuses and gratuities, it represents the value of income on which the worker can make decisions for the welfare of the household. In principle, of course, when the objective is to analyze the individual's employment-related wellbeing, the additional employment-related benefits provided by social security or compulsory insurance schemes or by the State should also be included. When the objective is to measure the income-generating capacity of a job, all the components of income related to paid employment provided by the employer should be included.

In the next section, the term earnings is sometimes used to refer to income from employment measurable in labour force surveys, covering both income related to paid employment and income related to self-employment.

Measurement in Labour Force Survey

Because a LFS is designed to measure current employment and unemployment, it provides a suitable vehicle to also measure income from employment. Various possibilities may be considered: measuring income from employment as an integral part of the core labour force survey or measuring income from employment as a separate exercise linked to the labour force survey. Other possibilities may also be envisaged.

Core Labour Force Survey

Measuring income from employment as part of the core LFS has both advantages and drawbacks. One advantage is the possibility of linking the income data to the individual's labour force status, as well as to other characteristics such as volume of employment in terms of hours of work and duration of employment, and type of occupation, skill level, etc. Another advantage is cost-effectiveness: the data collection on income and the subsequent data processing will share the infrastructure of the core labour force survey, so no new basic structures need to be developed except those for collecting and processing the extra elements on income.

There are, however, certain drawbacks. The survey design for collecting data on income should not infringe on the requirements of the core labour force survey, which should take precedence on any other requirements. This means that the scope of persons for which data on income on employment is to be collected and the method of data collection in terms of number and type of questions and choice of respondents should be limited so as to minimize any adverse effect on the results of the core labour force survey.

Mindful of these issues, national LFSs when also measuring income from employment are generally limited to the measurement of weekly or monthly income from employment of paid employees in their main job (option 1). Depending on the availability or non-availability of alternatives, this limited option may be expanded to cover also certain or all self-employed persons (option 2), or to subsidiary as well as main jobs (option 3), and in rare cases to income receiving from past employment as well as from current employment (option 4).

The measurement objective under all four options is the collection of data on weekly or monthly income from employment rather than income from employment an annual or twelve-month basis. The restriction to week or month is to conform to the reference period used for measuring employment in the labour force framework (reference week) and that of measuring hours usually worked, which is generally defined as the average hours actually worked per week over a four-week period (effectively a month).

Examples of question sequences under each of the four options are given in the working paper *Resolution concerning the measurement of employment-related income* cited earlier. The working paper also discusses a range of measurement issues, including in-kind benefits, occupation expenditure of employees, production for own-use, mixed income from self-employment, methods to reduce recall errors, the effect of proxy response against self-response, the use of income intervals and other innovative ways to accessing income data in surveys. An example of questionnaire design and construction of derived variables is reproduced in Annex A of the present report.

Special Added Module

The international standards stipulate that the income related to paid and self-employment should be measured over a long reference period such as a full-year, in order to take into account seasonal variations of activities, the fluctuations in work intensities of individuals, and the combination of multiple activities and periods of activity and inactivity of the population. Such a measurement is particularly appropriate if the objective of the survey is the analysis of the economic wellbeing of the population given the employment opportunities available to them.

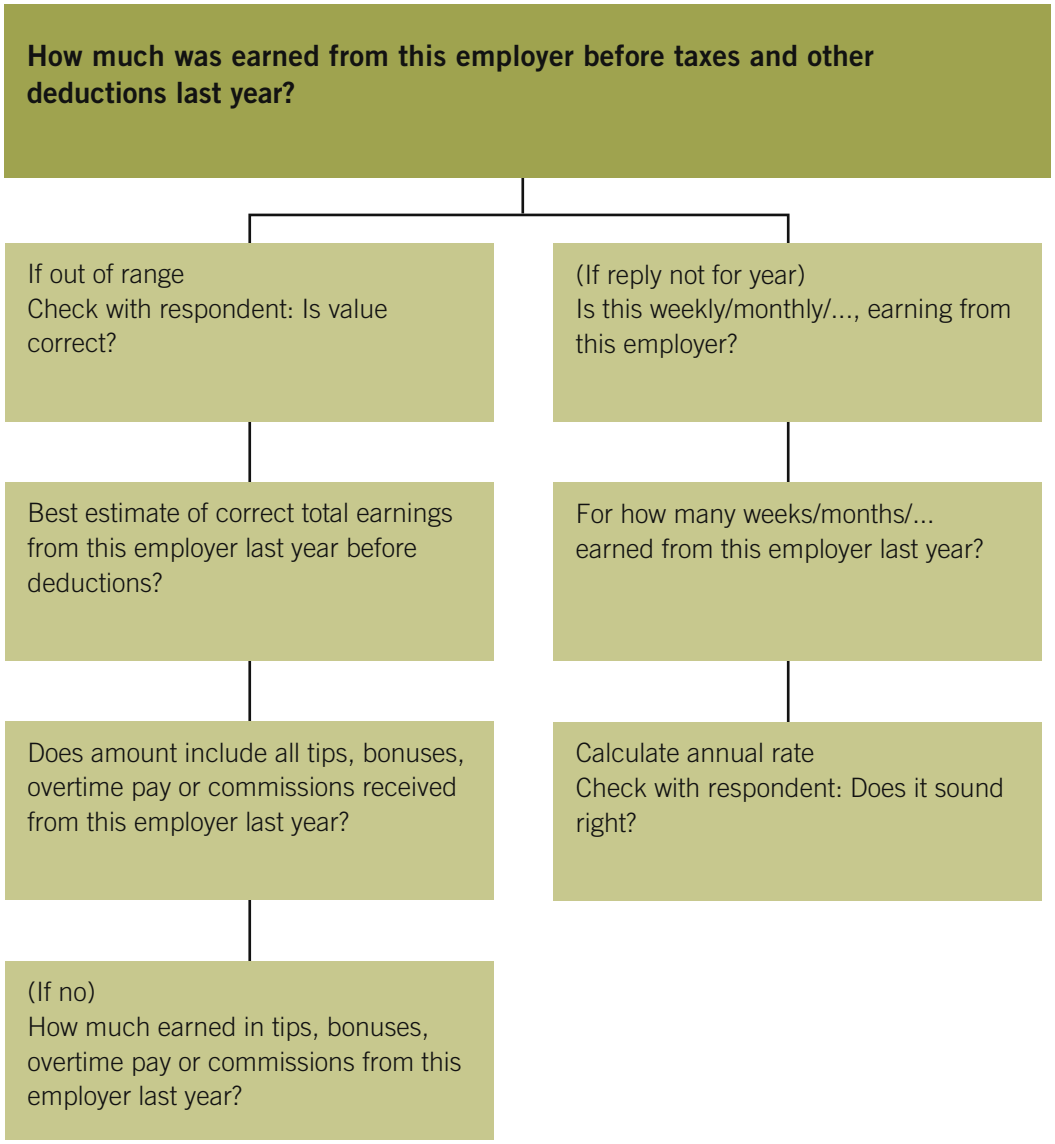
Core LFSs with a short reference period such as a calendar week or the last seven days do not lend themselves to measurement of income from employment over the year. A specialized survey module linked to the core labour force survey may be a more appropriate arrangement for such full measurement of income from employment. The specialized module will provide information on income and work intensity over the year and stable demographic and social characteristics of the individual. Two approaches are considered below.

Annual Income from Employment

One approach is to let the respondent provide a spontaneous response on his or her income from employment and then, if necessary, calculate the corresponding yearly amount and get confirmation from the respondent. Once the basic value is thus determined, the next step is to ensure that it covers the relevant components that may have been omitted from the spontaneous response.

The following question sequence concerns income from paid employment in the main job during the reference year. Similar questions are designed for measuring income earned after expenses from own business or farm and any secondary work performed during the year.

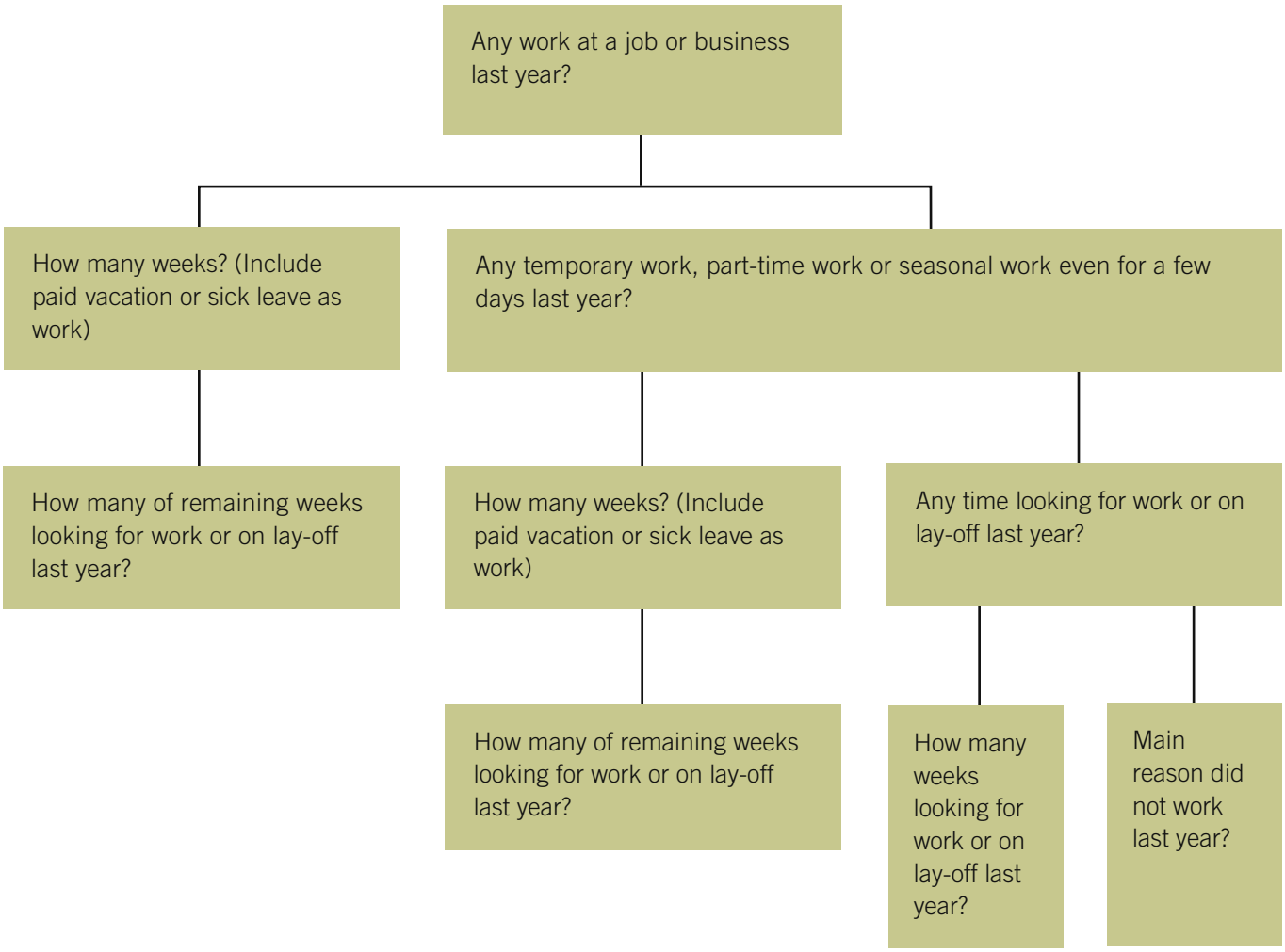
Figure 2: Income from Paid Employment in Main Job During Reference Year



Month-to-Month Work Experience

Another approach is to measure both employment and income from employment using a common long reference period: the reference year. Two variants to this approach may be considered, both based on retrospective questions. One variant follows the question sequence suggested for the core labour force survey, except that a year rather than a week is used as the reference period. The essential elements of the question sequence are reproduced in the flowchart below. Follow-up questions ask about the number of hours usually worked per week in the week or weeks the person worked during the reference year.

Figure 3: Measuring Employment and Income from Employment Using Reference Year



The second variant is based on a month-to-month recall of economic activity of the individual. The following diagram illustrates its application in a survey conducted, for example, in April year t with retrospective questions on the past twelve months.

Figure 4: Month-to-Month Recall of Economic Activity

Q. For each month in the last twelve months, did you have a job and report the corresponding income from employment? Include any little occasional jobs, too, even if they were only for 1 hour, as well as unpaid work in family business. Also, if job kept during an absence (holidays, illness, maternity leave, military service, etc.) enter code 1 "Yes", otherwise mark 2 "No".

Month	Year	1. "Yes"	2. "No"	If "Yes" report corresponding income from employment
March	t			
February	t			
January	t			
December	t-1			
November	t-1			
October	t-1			
September	t-1			
August	t-1			
July	t-1			
June	t-1			
May	t-1			
April	t-1			

The first variant has the advantage of being consistent with the labour force framework in measuring both work and job-search in terms of number of weeks as unit of measurement. The second variant, while consistent with the monthly concept of earnings, has the drawback of ambiguity with respect to the measurement of employment. It is not clear what a month of employment means. Should one hour of work during the month be considered as employment over the month? The month-to-month questioning has, however, the advantage of helping recall and limiting memory errors. This also permits the use of a reference year defined as the last 12 months rather than the last calendar year.

Measurement in Household Income and Expenditure Survey

The principle objective of HIESs is to assess the level, structure and trends of the economic wellbeing of households and individuals in terms of the distribution of income and consumption expenditure pattern for various population sub-groups of interest.[9]

Household income covers income from paid employment, income from self-employment, and income from other sources of the household, including property income, pension and social security benefits, remittances from abroad and other income such as royalties, alimony, etc. Some surveys record the income information for individual members of the household separately, and some record the information for the household as a whole.

Questionnaires for HIESs are generally extensive. They include a cover page (for identifying the sample household and reporting on the outcome of the survey interview), a household roster (for listing the household members and recording their socio-demographic characteristics), and a battery of questions on some twenty or so sources of household income and more than several hundred items of food and non-food expenditure.

In many HIESs, income from paid employment is measured for each currently held job by first identifying the occupation, industry, status in employment (public, private and cooperative sector), hours of work per day and days of work per week, and then by asking the monthly and annual gross earnings from that job before calculating net monthly and annual earnings by deducting monthly and annual taxes and contributions to pension funds, and any regular or irregular benefits linked to the job, such as housing allowance, family or child support, overtime, bonuses, end-of-year gratuities, etc.

Similarly, income from self-employment is measured for each currently engaged activity distinguishing between agriculture and non-agriculture activities. First, data are obtained on annual gross receipts of the business and on the different items of business expenditures: cash and in-kind salaries paid for labour engaged; purchases of products for sale or intermediary products such as fertilizers, pesticides and fuel; expenditures on equipments, maintenance, depreciation, etc.; other business expenses such as taxes, rent, transport and interest on loans and commissions. The net receipt is then calculated by deducting the business expenditure from total gross receipt. For agriculture activities, the questionnaire includes a special worksheet for collecting data on the nature and characteristics of the holding and calculating the net receipt for each type of agricultural products of the farm.

Total household income is then derived by aggregating the income from employment of all household members to which is added the income of household members received from other sources. Household income from other sources includes: past employment-related income, such as pension, severance and early retirement payments; property income, such as rental income received, royalties, interest and dividend receipts; transfer income received from other households and from non-profit institutions in the form regular gifts or financial support, such as scholarships, union strike pay, union's sickness benefits, relief payments; and finally income generated from household production of services for own consumption, calculated on the basis of a special worksheet incorporated as a separate part of the questionnaire.

The relative strength of HIESs in measuring income from employment is somewhat offset by the relative weakness in the measurement of employment characteristics. Because the principal objective of income and expenditure surveys are not the measurement of employment, questions not sufficiently probing are generally incorporated in the questionnaire to detect all types of employment activities, especially those that are unpaid, irregular and casual. For the same reason, there is often a relatively high extent of missing values and coding errors on employment characteristics, such as status in employment, branch of economic activity, occupation and hours of work. Also, HIES questionnaires generally do not include questions to identify the informal sector and informal employment.

The World Bank Living Standards Measurement Study (LSMS) also collects data on income and employment within a broad framework that covers many other dimensions of household well-being, including consumption, savings, health, education, fertility, nutrition, housing and migration.[10] The survey includes extensive quality control features and the questionnaire attempts to measure employment and income from employment in a consistent way, covering an entire year. However, the sample sizes of LSMS surveys are generally small and the resulting labour force indicators often differ from the corresponding figures derived from national labour force surveys.

---

[9] International Conference of Labour Statisticians. 2003. *Resolution concerning statistics of household income and expenditure*. Geneva: 17th ICLS.

[10] Grosh, Margaret E. and Paul Glewwe. 1995. *A guide to living standards measurement study surveys and their data sets*. Living standards measurement study (LSMS) working paper, No. LSM 120. Washington, D.C.: The World Bank. Available at http://documents.worldbank.org/curated en/1995/09/697050/guide-living-standards-measurement-study-surveys-data-sets (accessed 10 December 2014).

Measurement by Linking Labour Force Surveys and Household Income and Expenditure Surveys

LFSs in many countries do not collect data on income from employment and household income. Where they do, as mentioned earlier, the income data are generally restricted to earnings of employees, leaving out income from self-employment. By contrast, HIESs generally collect detailed data on household income as well as income from paid and self-employment of individual household members, but they are relatively less precise in the measurement of employment and its characteristics.

Can the two surveys be linked to take advantage of each source while minimizing their disadvantages? Three situations should be considered: (1) the two surveys are conducted on the same sample of households or the sample of one is a sub-sample of the other; (2) the two surveys are conducted on different households but in the same primary sampling units; and (3) the sample households and the primary sampling units of the two surveys are all different.

Full or Partial Sample Overlap, Common Sample Households

Where the income data are measured for each individual and the HIES is conducted on the same sample of households or a sub-sample of the households used for the labour force survey, the employment data of the labour force survey and the income from employment data of the household income and expenditure survey can be linked at the individual record level. In such cases and on the assumption that the questionnaire design provides for reliable household and individual identification, the main issue will be the calculation of appropriate sampling weights so that consistent aggregate estimates for the employed population can be made.

Let w be the extrapolation weight calculated for individual i in the LFS sample, and w the extrapolation weight for household h in the household income and expenditure survey sample. Explain the possible inconsistency that may arise between the aggregate number of persons derived based on the labour force survey weights and the aggregate number derived from the HIES weights.

Then, describe how consistency can be achieved by adjusting the weights of the LFS based on the values of the HIES weights. Let wi represent the weight of person i in the sample labour force survey (s_{LFS}) and wk the weight of household k in the sample of the household income and expenditure survey (s_{HIES}). The estimate of the total population derived from the labour force survey may be expressed by

$$P_{LFS} = \Sigma_{i \in sLFS}\ w_i$$

The corresponding estimate from the household income and expenditure survey is given by

$$P_{HIES} = \Sigma_{k \in sHIES}\ w_k n_k$$

where $\mathbf{n}_k$ is the number of household members in the sample household k.

The two population estimates, P_{LFS} and P_{HIES}, are generally different, and so are the two corresponding estimates of the total number of households. A method to ensure consistency between the two sets of population and household estimates is to adjust the labour force survey weights using the following proportional adjustment procedure.[11] Consider the set of all heads of households in the LFS sample,

$$S_{HLFS} = \{\ j\text{: if j is head of household in sample LFS}\}$$

and

$$S_{NLFS} = \{\ j\text{: if j is not head of household in sample LFS}\}$$

Then define the adjusted weight of individual i in the labour force survey sample by

$$w_i^* = \begin{cases} w_i\,(H_{HIES}/H_{LFS}) & \text{if } i \in S_{HLFS} \\ w_i\,(P_{HIES}-H_{LFS})/P_{NLFS} & \text{if } i \in S_{NLFS} \end{cases}$$

[11] Verma, Vijay. *Sampling Methods, Manual for Statistical Trainers Number 2*, Statistical Institute for Asia and the Pacific (SIAP), Tokyo, Revised 2002 (Section 6.9 Consistency between the estimated numbers of households and persons).

where HHIES is the estimated total number of households derived from the household income and expenditure survey, $H_{HIES} = \Sigma_{k.sHIES} w_{k;} H_{LFS}$ the corresponding estimate from the labour force survey, $H_{LFS} = \Sigma_{j.sHLFS} w_{j;} P_{HIES}$ the estimate of total population from the household income and expenditure survey; and P_{NLFS} the estimate of the non-head of household population from the labour force survey.

Common Primary Sampling Units, Different Sample Households

Where there is no overlap between the samples of the two surveys, linkage between the employment data from the labour force survey and income data from the household income and expenditure survey may not be achieved at the individual record level. Depending on the situation, linkage may nevertheless be achieved at a more aggregate level, for example, for the same primary sampling units or for higher levels of geographical divisions depending on the sample designs of the two surveys.

In such cases, the joint analysis of employment and income from employment is performed on the common sample areas. Suppose there are K common sample areas denoted by $a_1, a_2, \ldots, a_k, \ldots, a_K$. The employment information, $E_1, E_2, \ldots, E_k, \ldots, E_K$, will be coming from the labour force survey and the income information, $I_1, I_2, \ldots, I_k, \ldots, I_K$, from the household income and expenditure survey.

The use of the joint information to make inference about the country as a whole requires careful considerations. Like in the case of common sample households, the sampling weights of the common sample areas should be adjusted as the probability of selection of the areas may have been different under the sample designs of the two surveys.

Also, the joint analysis of employment and income from employment will be limited to cross-tabulations by geographic area as no other information is available at the level of the sample areas. For analysis by other characteristics such as sex, age group, and educational attainment, the area information should be initially produced in terms of those variables before cross-tabulation. Referring to the earlier notation, this means that the employment data Ek now refers, for example, to the number of male employed persons in sample area k and Ik the corresponding average income from employment of male employed persons in sample area k.

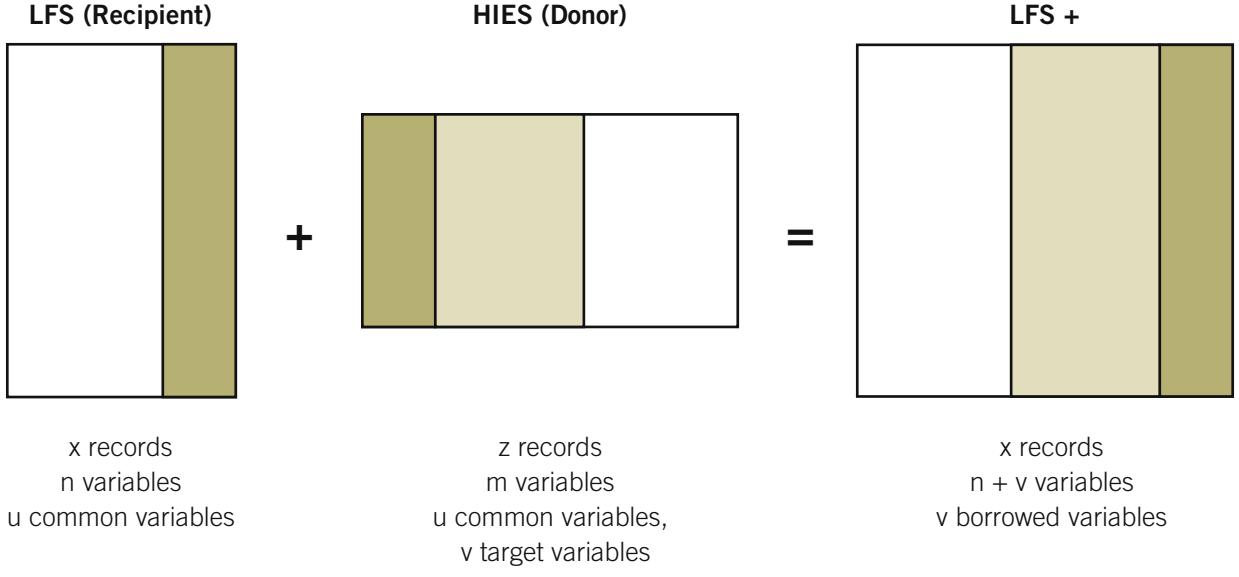
Completely Independent Samples

When the labour force survey (LFS) and income and expenditure survey (HIES) samples are completely independent, there are essentially no common sample units in the two samples. It may nevertheless be possible to link the two sources using statistical matching methods.

Statistical Matching

Statistical matching may be formulated in terms of data donor-recipient relationship as shown in the following diagram. In the present context, the donor is the household income and expenditure survey and the recipient is the labour force survey.

Figure 5: Schematic Representation of Statistical Matching



Suppose the LFS survey contains x records and n variables of which u variables are common with HIES. The u common variables are shown in dark grey in the diagram. The other LFS-specific variables are shown in white. Similarly, suppose the HIES contains z records and m variables. There are u common variables with LFS, shown in dark grey, and m-u HIES-specific variables. Among them v variables are targeted for merging with the LFS. The v target variables are shown in light grey and the other HIES-specific variables in white.

Statistical matching of LFS and HIES produces a new enhanced dataset LFS+ of synthetic values with the same number of records as the original LFS dataset (x records) but with more variables (n + v variables), the original n variables plus the v target variables borrowed from the HIES. The success of the statistical matching operations depend on several factors, including the quality of the original data, both the recipient and donor data sources, the properties of the matching algorithm and the procedures used to analyze the synthetic LFS+ data resulting from the statistical matching operations.

Statistical matching is a form of record linkage. It is increasingly used, for example, to impute missing or outlier values in surveys[12] or to link data from different sources.[13, 14]

A large part of the work of the project consisted of experimenting with the choice of common variables for linkage and different matching methods to determine if a match exists or not. The experimentation was made with data from the quarterly labour force survey of Iran (2011) and the annual household income and expenditure survey of Iran (2011).[15] The basic elements of the two surveys are described in Annex B of the present report.

In the LFS data file, a record corresponds to a *person*. Thus each line of the rectangular-shape shown in the left panel of the diagram above is a record corresponding to a member of the sample household covered by the LFS. The record contains the information obtained on the person in the course of the survey, including any derived variables and sampling weight associated to that person. Each employed person may have more than one job during the reference. For example, the test data on Iran for the province of Isfahan shows that 90 per cent of the employed persons reported to have been engaged in only one job during the reference week in spring 2011 (1,390 Q1), eight per cent reported two jobs and two per cent reported three or more jobs.

For each job reported by the employed person in the LFS, a corresponding income from employment is to be found in the HIES. For this purpose, the appropriate definition of a record in the HIES file is a "job", or more precisely, a "person-job". Accordingly, each line in the middle HIES rectangle corresponds to a person-job. Persons not employed or below 10 years of age do not have a job and the target variables on income from employment for these persons are empty. The target variables on income from employment of persons employed as contributing family workers are also empty.

Common Variables

"Common variables" are variables used for statistical matching that appear both in the LFSs and HIESs. In fact, the set of variables that appear in both LFS and HIES may be greater than the set of common variables in the sense used here. Certain variables such as previous residence or current school attendance may appear in both surveys but may be irrelevant for statistical matching problem at hand. They are therefore not considered as common variables in our context.

Among the common variables, there is a geographic variable that defines the address of the household in terms of urban-rural area, province and finer geographical location up to urban block or rural village. For the Iran data, it is defined in terms of a 10-digit identifier, where the first digit specifies the urban-rural location, the next two the province, the next four the serial number of the primary sampling unit in the master frame, and the final three the serial number of the household in the primary sampling unit.

The remaining common variables are selected in relation to income from employment. The choice is based on the Mincer's earnings equation.[16] In broad terms, the level of earnings adjusted for hours of work is a function of the occupation performed and the skill level of the incumbent. Skill level of the incumbent is itself a function of formal training received (measured here by educational attainment) and on-the-job training (measured, for example, by job tenure or simply in terms of age). The variable sex is generally also included to allow for earnings differential between men and women. Analysis of the Iran HIES data on income from employment has shown that the Mincer equation closely fits the data, especially for paid employment jobs.[17]

---

[12] Roderick, J.A. Little and Michael E. Samuhel. 1983. "Alternative models for CPS income imputation." US Bureau of the Census.
[13] Rubin, D.B. 1986. "Statistical matching using file concatenation with adjusted weights and multiple implementation." *Journal of Business and Economic Statistics*, 4, 87-94.
[14] Gupta, A., J.N. Kok and P. Van Der Puttan. 2002. *Data Fusion through Statistical Matching*, Paper 185. Boston: Center for eBusiness @ MIT.
[15] The data have been kindly provided by Statistical Centre of Iran, Ms. Asieh Abasi, Specialist, Economic Statistics, Bureau of Population and Labour Force Statistics (Household Income and Expenditures Group).
[16] Heckman, James J., Lance J. Lochner, and Petra E. Todd. 2003 *Fifty years of Mincer earnings regressions*, No. w9732. National Bureau of Economic Research.
[17] Salehi-Isfahani, Djavad. 2009. Education and earnings in the Middle East: A comparative study in Egypt, Iran, and Turkey. Working Paper No. 504, Economic Research Forum, September 2009.

Target or Borrowed Variables

In this project, only one target variable was considered, namely, income from employment. The target variable was borrowed from HIES to enhance the LFS. The use of different terms – target variable when it has to do with HIES and borrowed variable in the case of LFS – is deliberate. The value of the target variable in HIES – income from employment – may change in the process of borrowing when implanted in LFS. This is because hours of work may differ in the two data sources.

Let y_{HIES} denote the annual income from employment of a sample person in HIES at a job with usual hours of work per week, h_{HIES}. Assuming the reported weekly hours of work applies for all weeks of the month and all months of the year, the average hourly income from employment may be estimated as

$$\frac{y_{HIES}}{52 \times h_{HIES}}$$

The numerical factor in the denominator should be replaced by (52/12) if y_{HIES} is monthly income from employment.

If this target variable in HIES is borrowed for a matching employed person in LFS with usual hours of work per week, $h_{LFS,}$ it should be adjusted to become

$$\hat{y}_{LFS} = 52 \times h_{LFS} \times \frac{y_{HIES}}{52 \times h_{HIES}} = \frac{h_{LFS}}{h_{HIES}} \times y_{HIES}$$

Unless $h_{LFS} = h_{HIES}$ the values of the target variable y_{HIES} and the borrowed variable y_{LFS} differ. In the case of a self-employed person working with contributing family workers, income from employment must take into account the hours of work of the contributing family workers.[18] The relationship between the borrowed and target variables may thus be expressed as

$$\hat{y}_{LFS} = \frac{\sum h_{LFS}}{\sum h_{HIES}} \times y_{HIES}$$

where the summations are over the hours of work of the self-employed person and the contributing family workers in the two datasets, respectively.

Matching Method and Test Results

Many methods have been developed for statistical matching of two or more datasets. One series of methods views statistical matching as a missing value problem. The values of income from employment are missing in the LFS file and may be imputed from the HIES file using various imputation methods. The most common and simplest imputation methods are linear regression and hot deck.[19] More complex methods are also available, such as non-parametric regression, log-linear models[20] and multiple imputations.[21] Another series of methods regard statistical matching as a data fusion problem. The simplest algorithms are k-nearest neighbor prediction[22] and classification and regression trees.[23] More complex algorithms are based on neural networks and other data mining methods.

The procedure in all methods essentially consists of three steps. First, given some elements of the recipient file, the set of best matching donor elements is selected. Then, a matching distance is calculated over some subset of the common variables. And, finally, the predictions of the target variables are derived using the donor elements within the shortest distance. Often some constraints are imposed in the procedure. For instance, it may be desirable that men are never matched with women for variables known to be gender specific.

After experimenting with alternative matching techniques, hot deck imputation has been chosen as the most reliable. The basic methodology is described in Annex C. Its implementation has been tested with the Iran data for imputation of paid employment income and self-employment income separately.

---

[18] Chiswick, Carmel U. 1983. "Analysis of earnings from household enterprises: Methodology and application to Thailand." *The Review of Economics and Statistics*, Vol. 65, 658-62.
[19] Singh, A.C., H. Mantel, M. Kinack, and G. Rowe. 1993. "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption." *Survey Methodology*, 10, 59-79.
[20] Singh, A.C. 1989. "Log-linear imputation." *Proceedings of the 5th Annual Research Conference*. Washington D.C.: U.S. Bureau of the Census.
[21] Rubin, D.B. 1986. "Statistical matching using file concatenation with adjusted weights and multiple implementation." *Journal of Business and Economic Statistics*, 4, 87-94.
[22] Gupta, A., J.N. Kok, and P. Van Der Puttan. 2002. *Data Fusion through Statistical Matching*, Paper 185. Boston: Center for eBusiness @ MIT.
[23] Sariyar, M. and A. Borg. "The Record Linkage Package: Detecting Errors in Data." *Contributed Research Articles*. Available at http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf (accessed 10 December 2014).

In testing, the HIES file of paid employment records was randomly divided into two parts, one with 15,225 records regarded as the donor file and the other with 7,584 records as the recipient file. The following common variables were retained:

- Location with 2 categories (Rural and Urban)
- Sex with 2 categories (Male and Female)
- Age with 6 categories (15-19 yrs, 20-29 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, and 60+ yrs)
- Educational attainment with 3 categories (Primary, Secondary, and Tertiary)
- Sector of employment with 2 categories (Public and Private)
- Occupation with 9 categories (ISCO major groups)

The different categories of the common variables generated a total of 1,296 imputation cells (1,296=2x2x6x3x2x9). Some 745 cells were empty, i.e., there were neither recipients nor donors for those particular combination categories.

For each record in the recipient file, the hot deck imputation procedure first searches in the donor file records that match the same combination of categories of the common variables. Among the records found, it then picks a particular record at random (or by nearest-neighbor method) if there is more than one and assigns the corresponding income from paid employment to the recipient record. The final imputed income from paid employment of the recipient is then obtained by multiplying the donor's half-daily income from paid employment with the number of half-days worked by the recipient.

Comparing the imputed value with the actual value for all the 7,584 records of the recipient file, the results show that the average and median error rates of imputing income from paid employment by the hot deck method were, respectively:

Average error rate = 5.1 per cent
Median error rat = 3.5 per cent

The application of the same procedure for self-employment jobs was, however, disappointing. Because of the wide variation of reported income (in many cases even negative income responses), the error rate of imputing self-employment income was more than three times that of imputing paid employment income. Also, the post-imputation adjustment for hours of work and contributing family workers created complications that could not be resolved within the time frame of the project.

It appears that a more promising approach regarding self-employment jobs would have been to limit the imputation procedure to those self-employment jobs that have paid employment counterparts and to pool the two donor files together for imputation. For example, one may be able to assume that low-income self-employment jobs in non-agriculture occupations that do not involve contributing family workers and corresponding paid employment jobs in the same occupation categories are similar in terms of income from employment. It is envisaged to test this hypothesis in the future outside the framework of this project.

Main Conclusions

The measurement of income from employment along with employment should be an important objective of national statistical programmes. It is necessary in particular for compiling data on low pay workers and the working poor, two decent indicators recommended by the ILO.[24]

Here various methodologies have been described for application under different circumstances depending on the design of existing LFS and HIES when conducted at about the same time. The following cases were considered:

(a) part of core LFS
(b) part of special module attached to LFS
(c) exact match in overlap sample of LFS and HIES
(d) geographic match of common primary sampling units of LFS and HIES
(e) statistical match in independent LFS and HIES

Under (a), the most feasible option is to limit the measurement to weekly or monthly income from employment of paid employees in the main job. But other options are possible – in particular, expanding the measurement to cover also certain or all self-employed persons, or to secondary as well as main jobs, or perhaps to cover income received from past as well as current employment.

Under (b), the most appropriate measurement objective is to collect data on income and work experience over an entire year. Two particular approaches are proposed: the measurement of annual income from employment by spontaneous response of the respondent followed by a sequence of questions for control; and month by month recall of work experience and reporting of the corresponding income from employment.

Under (c), the employment data of the labour force survey are linked to the income from employment data of the household income and expenditure survey at the individual record level by exact matching procedures. Adjusted sampling weights must be calculated as the sample design of the two surveys may differ and the sample overlap may only be partial. A method for deriving appropriate sampling weights in such cases has been described.

Under (d), there is no overlap between the samples of the two surveys, but the sample designs are based on common primary sampling units, generally geographical areas. In such cases, the joint analysis of employment and income from employment is performed not at the individual record level but by matching the common sample areas.

Under (e), the joint measurement of employment and income from employment is done by statistical matching of LFS records as recipients of income data from HIES records as donors. Statistical matching is the most fragile methodology for joint measurement but still viable under certain circumstances. The test made with the Iran data showed that hot deck imputation using sex, age, educational attainment, occupation and urban/rural residence for defining the imputation cells lead to relatively small error rates in the case of paid employment jobs (average error rate 5.1 per cent and median error rate 3.5 per cent).

In the case of self-employment jobs, the results were, however, poor for several reasons – in particular, great variation of reported income from self-employment (in many cases even negative income responses), and the complications of adjusting for hours of work and contributing family workers. The paper proposes to limit the imputation of income from employment for self-employment jobs to self-employment jobs that have paid employment counterparts and to pool the two donor files together for imputation. Such jobs are likely to be informal jobs in non-agriculture occupations that do not involve contributing family workers, such as street vendors, taxi drivers, and, in general, employers and own-account workers in the informal sector.

It should finally be mentioned that two other relevant topics could not be covered in the project, namely, the joint measurement of employment and of total household income and of the use of administrative files as sources of data on income from employment.

Joint data on employment and total household income is needed for measuring the working poor, broadly defined as working persons who are unable to earn enough to maintain the welfare of themselves and their families. In addition to the points made earlier in the context of linking data on employment and income from employment, a number of other issues

[24] ILO. 2008. *Measuring Decent Work*. Discussion Paper for Tripartite Meeting of Experts on the Measurement of Decent Work, Geneva, September 8-10.

should be considered when data on total household income for poverty measurement are also required. In particular, as poverty and geography are highly correlated, it is important that the grid used for linking the LFS and HIES surveys be at the highest level of details as possible.

Administrative files as possible sources of data on income from employment include income tax returns, records of social security organizations, pension funds, health insurance institutions and business registration systems. Income data from administrative sources are increasingly being used in combination with survey data for a variety of purposes: to enrich the survey data with administrative income data not found in the survey; to assess the accuracy of survey data on earnings of employees and to adjust for errors in the income statistics produced from survey data, or to impute missing values or outliers of survey income data, or still to examine the impact of participation in labour market programmes or the validity of simulated tax models.

Annex A

An Example of Questionnaire Design and Calculation of Income from Employment in Core Labour Force Survey

The following example refers to a simplified set of questions proposed for measuring income from paid employment in the main job and secondary jobs in a core labour force survey of an African country.[25] Several particular issues are addressed for the measurement of income from employment as part of the core labour force survey in developing countries. One is the problem of gross income versus net income in the case of self-employment jobs and the importance of subsistence activities among the self-employed workers, particularly in the agriculture sector.

The other is the measurement of income from secondary jobs or activities concerning contributing family workers. These workers are unpaid in their main jobs, but may be receiving income in their other jobs or activities. A similar issue arises in the case of subsistent agriculture workers who may be receiving income from their secondary jobs or activities.

The question sequence distinguishes between paid employment jobs, self-employment non-agriculture jobs and self-employment agriculture jobs.

Figure A1 For Paid Employment Jobs

D13. How much did you earn at your main job last time you were paid, in cash and in kind, before taxes and other deductions?

Enter amount _____
Don't know

D14. What period did it cover?

1. Last month
2. Last week
3. Last day
4. Other (specify)

If "Don't know" in D13:

D17. Would you say the monthly amount was in the range ...?

1. Less than 1,000
2. 1,000 – 10,000
3. 10,000 – 50,000
4. 50,000 – 100,000
5. 100,000 or more

[25] "Malawi Labour Force Survey Programme: A Proposal," Farhad Mehran, Consultant, African Development Bank, 12 August 2011.

Figure A2 For Self-Employment in Non-Agriculture Jobs

D16. How much was your net earnings from your main activity (shop/business) after expenses last month?

Enter amount _____
Don't know

If "Don't know" in D16:

D17. Would you say the monthly amount was in the range ...?

1. Less than 1,000
2. 1,000 – 10,000
3. 10,000 – 50,000
4. 50,000 – 100,000
5. 100,000 or more

Figure A3 For Self-Employment in Agriculture Jobs

D15. In general, are the products obtained from this work for sale/barter or mainly for own family use?

1. Only for sale/barter
2. Mainly for sale/barter but partly for own or family use
3. Mainly for own or family use but partly for sale/barter
4. Only for own or family use

And then D16 and D17 as above

The interval responses (D17) and the qualitative responses on work for sale or barter (D15) require conversion to full income values for further data processing. For the interval response, it is proposed to use the mid-point of the interval (arithmetic average of the interval limits) in the low-income ranges, the geometric average of the interval limits for the middle-income ranges and the harmonic average of the upper income ranges.[26] For the sale/barter question, it is proposed to assign a weight 1 to the first answer category, 2/3 to the second, 1/3 to the third and 0 to the fourth and to analyze the results as indicated in the general methodology of calculation of monthly income from employment at main job and other jobs described in the following table.

[26] Mehran, Farhad. 1975. *Dealing with Grouped Income Distribution Data*. World Employment Programme. Working Paper WP 23-20. Geneva: International Labour Office.

Table A1 Calculation of Monthly Income from Employment at Current Main and Other Jobs

Staus in employment		Response	Monthly income from employment at main job		Monthly income from employment at other current jobs		Total
(1)	(2)	(3)	(4)		(5)		(6)
Employee	D12 = 1	D13= Amount	D13 x 1	if D14=1	1000 x 3/2	if D19=1	(4)+(5)
			D13 x 52/12	if D14=2	(1000 + 10000)/2	if D19=2	(4)+(5)
			D13 x 365/12	if D14=3	√10000 x √50000	if D19=3	(4)+(5)
			D13 x ?	if D14=4	2/(1/50000 + 1/100000)	if D19=4	(4)+(5)
					100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)
		D13 = Don't know	1000 x 3/2	if D17=1	1000 x 3/2	if D19=1	(4)+(5)
			(1000 + 10000)/2	if D17=2	(1000 + 10000)/2	if D19=2	(4)+(5)
			√10000 x √50000	if D17=3	√10000 x √50000	if D19=3	(4)+(5)
			2/(1/50000 + 1/100000)	if D17=4	2/(1/50000 + 1/100000)	if D19=4	(4)+(5)
			100000 x 5/3	if D17=5	100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)
Employer or Own-account worker in non-agriculture activities	D12=2 D12=3	D16 = Amount	D16		1000 x 3/2	if D19=1	(4)+(5)
					(1000 + 10000)/2	if D19=2	(4)+(5)
					√10000 x √50000	if D19=3	(4)+(5)
					2/(1/50000 + 1/100000)	if D19=4	(4)+(5)
					100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)
		D16 = Don't know	1000 x 3/2	if D17=1	1000 x 3/2	if D19=1	(4)+(5)
			(1000 + 10000)/2	if D17=2	(1000 + 10000)/2	if D19=2	(4)+(5)
			√10000 x √50000	if D17=3	√10000 x √50000	if D19=3	(4)+(5)
			2/(1/50000 + 1/100000)	if D17=4	2/(1/50000 + 1/100000)	if D19=4	(4)+(5)
			100000 x 5/3	if D17=5	100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)
Employer or Own-account worker in agriculture activities	D12=4	D16 = Amount	D16 x 1	if D15=1	1000 x 3/2	if D19=1	(4)+(5)
			D16 x 3/2	if D15=2	(1000 + 10000)/2	if D19=2	(4)+(5)
			D16 x 3/1	if D15=3	√10000 x √50000	if D19=3	(4)+(5)
			D16 x ?	if D15=4	2/(1/50000 + 1/100000)	if D19=4	(4)+(5)
					100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)
		D16 = Don't know	1000 x 3/2 x 3/(4-D15)	if D17=1	1000 x 3/2	if D19=1	(4)+(5)
			(1000 + 10000)/2 x 3/(4-D15)	if D17=2	(1000 + 10000)/2	if D19=2	(4)+(5)
			√10000 x √50000 x 3/(4-D15)	if D17=3	√10000 x √50000	if D19=3	(4)+(5)
			2/(1/50000 + 1/100000) x 3/(4-D15)	if D17=4	2/(1/50000 + 1/100000	if D19=4	(4)+(5)
			100000 x 5/3 x 3/(4-D15)	if D17=5	100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)
Contributing family worker	D12=5		0		1000 x 3/2	if D19=1	(4)+(5)
					(1000 + 10000)/2	if D19=2	(4)+(5)
					√10000 x √50000	if D19=3	(4)+(5)
					2/(1/50000 + 1/100000)	if D19=4	(4)+(5)
					100000 x 5/3	if D19=5	(4)+(5)
					0	if D19=6	(4)+(5)

Annex B

Description of National Data Used for Experimenting Different Statistical Matching Techniques

The Statistical Centre of Iran has been conducting regular labour force surveys and household income and expenditure surveys as part of its national statistical programme. After a first attempt in 1373 Shamsi calendar (1994), the labour force survey became annual in 1376 (1997) and quarterly since 1383 (2004). The household income and expenditure survey is annual with quarterly samples and has been in operation for about fifty years, since 1342 (1963) in rural areas and since 1347 (1968) in urban areas.

The labour force survey collects detailed information on the current employment and unemployment characteristics of the population, including secondary jobs. It does not, however, collect data on earnings and household income. The household income and expenditure survey also collects data on the employment characteristics of household members but in less detail than the labour force survey. It does, however, collect detailed information on household income by source of income as well as income from employment of household members for each job separately.

The questionnaire contents and sample designs of each survey are briefly described below.

Quarterly Labour Force Survey of Iran (2011)

The questionnaire of the labour force survey of Iran, 1390 (2011) is composed of three parts: a cover page recording information identifying the sample household and the date and outcome of the survey interview; a household roster listing the household members and recording their principal demographic characteristics; and an individual questionnaire recording the labour force characteristics of each household member 10 years old and over. There are in total 19 questions in the household roster and 50 in the individual questionnaire.

The LFS questionnaire includes six probing questions to measure employment (Q1-Q6) and asks about usual hours of work for the main and secondary jobs (Q16). Occupation and industry in main (Q9-Q10) and secondary jobs (Q28-Q29) are coded at 4-digit level. Status in employment in main and secondary jobs distinguishes employers, own-account workers, contributing family workers, employees and trainees. Informal employment in the main job may be assessed through the question on size of establishment (Q12) and health insurance coverage in the job (Q13).

The LFS is conducted on a quarterly sample of about 45,000 urban and rural households drawn in two stages. In the first stage, a master sample of primary sampling units (PSUs) is constructed from the sampling frame of enumeration areas of the population and housing census of 1385 (2006). A primary sampling unit may be a census enumeration area, or part of it, or a collection of them so that it contains about 200 to 400 target households. The PSUs are stratified by certain geographic and socio-economic characteristics and the sample size is allocated in proportion among the strata before drawing a random sample of PSUs with probability proportional to size where size is measured in terms of number of households according to the 2006 census.

In the second stage of sampling, a sample of target households is drawn from each selected primary sampling unit by systematic random sampling with equal probability from the list of households prepared based on the 1385 (2006) population census. The final sample of households is rotated through time according to the 2-2-2 scheme so that each sample household is in the sample for two consecutive quarters, removed from the sample in the two subsequent quarters and brought back in the sample for the next two quarters before leaving the sample permanently.

Annual Household Income and Expenditure Survey of Iran (2011)

The 2011 HIES questionnaire is extensive as it includes not only a cover page (for identifying the sample household and reporting on the outcome of the survey interview) and a household roster (for listing the household members and recording their socio-demographic characteristics), but also a large number of questions on some twenty or more sources of household income and more than 500 items of food and non-food expenditure.

Income from paid employment is measured for each currently held job by first identifying the occupation, industry, status in employment (public, private and cooperative sector), hours of work per day and days work per week, and then asking the monthly and annual gross earnings from that job before calculating net monthly and annual earnings by deducting monthly and annual taxes and contribution to pension funds and any regular or irregular benefits linked to the job such as housing allowance, family or child support, overtime, bonuses, end-of-year gratuities, etc.

Income from self-employment is measured for each currently engaged activity distinguishing between agriculture and non-agriculture activities. First, data are obtained on annual gross receipts of the business and on the different items of business expenditures: cash and in-kind salaries paid for labour engaged; purchases of products for sale or intermediary products such as fertilizers, pesticides, fuel; expenditures on equipments, maintenance, depreciation, etc.; other business expenses such as taxes, rent, transport, interest on loans and commissions. The net receipt is then calculated by deducting the business expenditure from total gross receipt. For agriculture activities, the questionnaire includes a special worksheet for collecting data on the nature and characteristics of the holding and calculating the net receipt for each type of agricultural products of the farm.

Household income is measured by aggregating the income from employment of all household members to which is added the income of household members received from other sources, including the monthly cash transfer received by the head of household for all members of the household registered under the national programme, *Yaraneh*. Household income from other sources include past employment-related income such as pension, severance and early retirement payments; property income such as rental income received, royalties, interest and dividend receipts; transfer income received from other households and from non-profit institutions in the form of regular gifts or financial support such as scholarships, union strike pay, union's sickness benefits, relief payments; and finally income generated from household production of services for own consumption, calculated on the basis of a special worksheet incorporated as a separate part of the questionnaire.

The HIES is conducted on a sample of approximately 38,500 households, about equally distributed in urban and rural areas. The sample is selected in three stages. In the first stage, samples of census enumeration areas are drawn within urban and rural strata, with probabilities proportional to size measured in terms of the number of households according to the 1385 (2006) population census. In the second stage, sample of clusters (urban blocks and rural villages) are similarly drawn from the list of urban and rural clusters in each sample enumeration area selected in the first stage of sampling. The sample consists of about 3,700 urban blocks and 3,900 rural villages. In the third stage of sampling, a sample of five households is drawn with equal probability by systematic random sampling in each urban and rural cluster from the list of households based on the 1385 (2006) population census.

Annex C

Hot Deck Imputation

Hot deck imputation is generally used for handling missing data in surveys.[27] In the context of the present problem, hot deck imputation means borrowing for each unit in the recipient pools (LFS) the income from employment of a "similar" unit in the donor pools (HIES). Donor pools, also referred to as imputation classes or adjustment cells, are formed based on auxiliary variables that are observed for both donors and recipients.

For the choice of auxiliary variables, two cases are distinguished: paid employment and self-employment. In the case of paid employment, the auxiliary variables are chosen on the basis of theoretical knowledge regarding earning functions (Mincer equations) as specified in the introduction. In the case of self-employment, additional variables are required, taking into account the portion of self-employment income attributable to labour and the existence of contributing family workers who are not directly remunerated paid for their involvement in the production of the household enterprise.

In general, consider q auxiliary variables $X_{1,}$ $X_{2,}$ …, X_q observed for both recipients and donors. Let $xi=(X_{1i,}$ $X_{2i,}$ …, $X_{qi})$ denote the sect of values of the auxiliary variables for subject i in the recipient pool. Similarly, let $xj=(X_{1j,}$ $X_{2j,}$ …, $X_{qj})$ denote the corresponding set of values for a given subject j in the donor pool. Further, let $C(x_i)$ represent the cells in the cross-classification of the q auxiliary variables in which subjects i falls.

Then, matching the recipient i to donors j in the same adjustment cell is the equivalent as using a distance function for matching where the distance function is defined as

$$d(i,j) = \begin{cases} 1 & j \in C(x_i) \\ 0 & j \notin C(x_i) \end{cases}$$

Other measures of distance of potential donors to recipients have been defined in the literature. A particular distance function that handles both categorical and continuous auxiliary variables is the predictive mean defined as

$$d(i,j) = [\hat{Y}(x_i) - \hat{Y}(x_j)]^2$$

Where the first term is the fitted value of the target variable Y (i.e., income from employment) from the regression of Y on the auxiliary variables using only the donors, and the second the predicted value of the missing variable Y for the recipient i from that regression.

In practice, it may be advantageous to impute a function of Y and x instead of Y itself. For example, in the context of imputation of income from employment, the natural variable for imputation is the logarithm of income adjusted for hours of work. Thus the regression and imputation procedures are carried out for the function ln(Y/S) where S is the hours of work performed during the reference period of the measurement of income Y.

There are several methods to define the set of donors j in the donor pool (HIES) that are potential donors for a given unit i in the recipient pool (LFS). One method selects as donor the element j in the donor pool that is nearest to the recipient i in terms of the distance d(i,j). The method is called a deterministic or nearest-neighbor hot deck and may be expressed as select j_o such that

$$d(i,j_o) = \min_j [d(i,j)]$$

Another method defines the donor set as all elements that are within a certain distance from the recipient and then selects a specific donor by a random draw. In other words, the donor set for a given recipient i is defined by

$$C(x_1) = \left\{ j : d(i,j) \leq \Delta \right\}$$

where Δ is a pre-specified threshold defining the maximum distance between the potential donors and the recipient. A specific donor is then selected by a random draw from the set of potential donors $C(x_i)$.

[27] Andridge, Rebecca R., and J. A. Little Roderick, J.A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review*, 78(1), 40-64.

Still another method is to consider all elements in the donor pool (HIES) as potential donors and carry the random draw with unequal probabilities according to which the donors are selected with probability inversely proportional to their distance from the recipient. The probability of selection of a donor j for a given recipient i may thus be expressed as

$$P(j) \cong \frac{1}{d(i,j)}$$

In the case where the distance d(i,j) is zero, the probability of selection is one and the result is identical to that of nearest neighbor hot deck.

In nearest neighbor hot deck and, for that matter, also in random hot deck methods, the same element in the donor pool may be used several times as donor. In order to avoid multiple use of the same donor, the donor file is generally sorted and donors are assigned to recipients in a sequential manner, so that each donor is used at most once in the recipient file.

Another issue is the role of sampling weights. If donors are selected by a deterministic or nearest neighbor hot deck, the sampling weight of the recipient in the recipient pool (LFS) is unaffected by the imputation method. However, if donors are selected by simple random samples from the donor pool, estimators are subject to bias if their sampling weight in the donor pool (HIES) is ignored. One approach that removes the bias in selections with equal probability within adjustment cells is to inflate the donated value by the ratio of the sample weight of the donor to that of the recipient.[28]

[28] Another approach that avoids imputations of implausible values in the case of integer-valued variables is described in Rao, JNK. 1996. "On variance estimation with imputed survey data," *Journal of the American Statistical Association*, Vol. 91, 499-506.

About WIEGO: Women in Informal Employment: Globalizing and Organizing is a global network focused on securing livelihoods for the working poor, especially women, in the informal economy. We believe all workers should have equal economic opportunities and rights. WIEGO creates change by building capacity among informal worker organizations, expanding the knowledge base, and influencing local, national and international policies. Visit www.wiego.org.