



STEP Skills Measurement Surveys

Innovative Tools for Assessing Skills

July 9, 2014

Gaëlle Pierre, Maria Laura Sanchez Puerta, Alexandria Valerio, Tania Rajadel
(World Bank)¹

With contributions from:

Household Survey

Angela Duckworth (University of Pennsylvania), Nancy Guerra (University of California), Michael Handel (Northeastern University), Sergio Urzua (University of Maryland), Irwin Kirsch, Claudia Tamassia, Mary Lou Lennon, Ann Kennedy, Eugenio Gonzalez, and Kentaro Yamamoto (ETS), Valerie Evans (World Bank), Sebastian Monroy Taborda (World Bank), Thomas Sohnsen (World Bank)

Employer Survey

David Margolis (Paris School of Economics), Emanuella di Gropello (World Bank), John Earle (George Mason University)

¹ The World Bank: Human Development Network. Maria Laura Sanchez Puerta (msanchezpuerta@worldbank.org) and Alexandria Valerio (avalerio@worldbank.org). Gaëlle Pierre and Tania Rajadel were consultants at the World Bank at the time of writing.

Abstract

The Skills Towards Employability and Productivity (STEP) program was designed to better understand the interplay between skills on the one hand and employability and productivity on the other. The STEP program developed survey instruments tailored to collect data on skills in low- and middle-income country contexts. The present note is a reference document for readers seeking background information on the STEP surveys and for users of the data, which is publicly available through the World Bank's Microdata Catalog. The note describes the design of the survey instruments and the constructs measured as well as the technical standards and implementation protocols adopted to ensure data quality and comparability across countries. It also provides guidance to users for the construction of aggregated skills indicators and for the use of the reading literacy assessment data.

JEL Classification: E24, I35, J21, J24, J23

Keywords: skills; personality traits; socio-emotional skills; literacy; education; labor markets

This methodology note was prepared by a team comprising G  lle Pierre, Maria Laura Sanchez Puerta, Alexandria Valerio, and Tania Rajadel. The team appreciates the leadership and technical support provided by Arup Banerji, Anush Bezhanyan, Elizabeth King, Harry Patrinos, and David Robalino. Ariel Fiszbein (former Human Development Chief Economist) provided invaluable and sustained technical support since the early stages of the STEP Skills Measurement program.

Comments, which helped shape and improve this note, were received from participants during the World Bank’s Human Development Network Learning Week on February 4-8, 2013, and during the Poverty and Inequality Analysis Course on March 6-7, 2013. Technical inputs were provided by Omar Arias, Reena Badiani-Magnusson, Juan Baron, Christian Bodewig, Aline Coudouel, Shuang Chen, Rodica Cnoblach, Wendy Cunningham, Tazeen Fasih, Marta Favara, Deon Filmer, Robin Horn, Yuki Ikeda, Kevin MacDonald, David Margolis, Natalia Millan, Sebastian Monroy Taborda, David Newhouse, Ana Maria Oviedo, Jorge L. Rodriguez Mesa, Jan Rutkowski, Federica Saliola, Yevgenivya Savchenko, and Jee-Peng Tan.

World Bank regional teams were led by Meskerem Mulatu (Armenia, Azerbaijan and Georgia), Ana Maria Oviedo (Bolivia), Pablo Acosta (Bolivia and Colombia), Peter Darvas (Ghana), Helen Craig (Kenya), Ximena del Carpio (Lao PDR), Johannes Koettl and Indhira Santos (Macedonia and Ukraine), Halil Dundar (Sri Lanka), Christian Bodewig (Vietnam) and Xiaoyan Liang (Yunnan Province, China).

The STEP Skills Measurement program has benefited from the Technical Cooperation established with Irwin Kirsch, Claudia Tamassia, Ann Kennedy, Mary Lou Lennon, Kentaro Yamamoto and Eugenio Gonzalez (Education Testing Services) and Andreas Schleicher, William Thorn, Marta Encinas - Martin, Mark Keese and Glenda Quintini (OECD). The program received technical support from Angela Duckworth (University of Pennsylvania), Nancy Guerra (University of Delaware), Michael Handel (Northeastern University), Kees Hammink (Consultant), and Marguerite Clarke (World Bank). Financial support to implement the STEP Skills Measurement surveys was received from the Bank Netherlands Partnership Program (BNPP), the Multi Donor Trust Fund for Labor Markets, Job Creation and economic Growth (MDTF) and the Russia Education Aid for Development (READ) Trust Fund.

The team acknowledges the support from Melvina Clarke, Elise Egoume-Bossogo, Lorelei Lacdao and Marie Madeleine Ndaw. The written pieces contained within this study were edited by Marc DeFrancis (DeFrancis Writing & Editing).

Table of Contents

I. BACKGROUND	7
II. HOUSEHOLD SURVEY	11
1. Overview	11
2. Description of the Questions Module-by-Module	14
<i>Household Level Information</i>	14
<i>Individual Respondent Information</i>	16
III. EMPLOYER SURVEY	48
1. Overview	48
2. Description of the Questions Module by Module	50
IV. STANDARDIZED IMPLEMENTATION	56
1. Implementation of the STEP Household Survey	56
<i>Set-up</i>	57
<i>Fieldwork Preparation</i>	57
<i>Data Collection</i>	59
<i>Data Processing</i>	59
<i>Sample Size and Response Rates</i>	61
2. Implementation of the STEP Employer Survey	62
<i>Set-up</i>	63
<i>Fieldwork Preparation</i>	64
<i>Data Collection</i>	66
<i>Data Processing</i>	66
<i>Sample Size and Response Rates</i>	67
V. USING THE STEP DATA IN ANALYSES	68
1. Skills Aggregation Methodology	68
<i>Aggregation Principles</i>	68
<i>Identifying Relevant Sub-domains of Skills</i>	69
<i>Translating Complex Scoring Scales into Interpretable Objects</i>	72
<i>Aggregating Self-reported Cognitive and Job-relevant Skills</i>	74
<i>Aggregating Behavioral and Personality Trait Measures</i>	77
2. Direct Reading Literacy Assessment Data	79
<i>Reading Components Results: Accuracy and Rate</i>	79
<i>Core Assessment</i>	80
<i>Exercise Booklets: the Literacy Proficiency Scale & Proficiency Levels</i>	80
3. Matching Skills from the STEP Household and Employer Surveys	84

VI. RESOURCES AND WAY FORWARD.....87

1. Resources	87
<i>Country Reports</i>	87
<i>Materials</i>	88
2. Way Forward.....	88

VII. REFERENCES89

VIII. APPENDIX92

Appendix 1. STEP Reading Literacy Assessment Sample Items	92
Appendix 2. STEP Stata Module	95

LIST OF BOXES

Box 1. Survey Instruments Development Timeline	8
Box 2. Construction of the Asset-based Wealth Index.....	15

LIST OF TABLES

Table 1. STEP Household Survey: Topical Content, by Module.....	12
Table 2. STEP Employer Survey: Topical Content, by Module.....	51
Table 3. Cognitive, Socio-emotional, and Job-relevant Skills	54
Table 4. STEP Household Surveys: Sample Sizes and Response Rates	62
Table 5. STEP Employer Survey Sampling Options	65
Table 6. STEP Employer Survey: Sample Sizes and Response Rates for Selected Surveys.....	67
Table 7. How to Generate the Sub-domain “Extraversion”	68
Table 8. Self-reported Cognitive Skills.....	75
Table 9. Selected Job-relevant Skills.....	75
Table 10. Behavioral and Personality Trait Measures	78
Table 11. Reading Proficiency Levels.....	83
Table 12. Skill Domains and Matching Questions in the STEP Household and Employer Surveys	86

LIST OF FIGURES

Figure 1. STEP Household Survey: Structure	11
Figure 2. Random Selection of the Individual Respondent	16
Figure 3. Workflow for the STEP Reading Literacy Assessment	39
Figure 4. Word Meaning: Sample Item	40
Figure 5. Sentence Processing: Sample Items	41
Figure 6. Reading Comprehension: Sample Passage	42
Figure 7. Exercise Booklet Design for STEP Literacy Items	43
Figure 8. Reading Literacy Performance / Scale from 1 - 5	45
Figure 9. STEP Employer Survey Structure	49
Figure 10. STEP Household Survey: Implementation Stages.....	56
Figure 11. STEP Employer Survey: Implementation Stages.....	63
Figure 12. Early Childhood Education and Reading Literacy in Urban Lao, Sri Lanka, and Vietnam....	71
Figure 13. Reading Skills Use and Educational Attainment in Urban Ghana.....	74

I. BACKGROUND

The Skills Towards Employability and Productivity program (henceforth referred to as STEP) provides a set of core surveys and implementation materials to build comparable country databases on skills that can be used for country-level policy analysis. This methodology note describes the survey instrument design, the constructs that are measured, and the technical standards and implementation protocols that have been designed and applied to ensure comparability of data. The note provides useful background to readers who may want to implement such surveys in their own countries, but it is especially targeted to users of the datasets that have been collected with these surveys. In particular, it explains the skills concepts that are measured in the surveys and provides guidance for the construction of aggregated skills indicators.

STEP consists of two survey instruments that collect information on the supply and demand for skills. Both surveys drew on similar surveys fielded in Peru, Lebanon, the United States, and other OECD countries and on extensive consultations with a panel of experts.² They were developed, piloted and fine-tuned over a period of one year before being implemented in a first wave of seven countries in 2012 and a second wave of six countries in 2013 (see Box 1).

An important aspect of the STEP surveys is the use of a multi-dimensional concept of skills that goes beyond educational attainment to capture human capital more comprehensively. Three broad types of skills are measured. **Cognitive skills** are defined as the “ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking

² For the household survey, the panels of experts included Marguerite Clarke (World Bank), Angela Duckworth (University of Pennsylvania), Peter Elias (University of Warwick), Nancy Guerra (University of California, Los Angeles), and Michael Handel (Northeastern University). For the employer survey, the panel included David Margolis (Paris School of Economics), John Earle (George Mason University), Francis Green (Institute of Education), Nathalie Greenan (Centre d’études sur l’emploi, France), Hartmut Lehmann (University of Bologna), and David McKenzie (DEC, World Bank).

thought.”³ Literacy, numeracy, and the ability to solve abstract problems are all cognitive skills. **Socio-emotional skills**, sometimes referred to in the literature as non-cognitive skills or soft skills, relate to traits covering multiple domains (such as social, emotional, personality, behavioral, and attitudinal). **Job-relevant skills** are task-related (such as computer use) and build on a combination of cognitive and socio-emotional skills.

Box 1. Survey Instruments Development Timeline

Development. The development of the surveys started in September 2010, when a group of experts were solicited to provide drafts for each skills module of the household survey and of the employer survey to measure skills supply and demand in developing countries. The initial drafts of each survey were extensively reviewed and revised by a wider group of specialists within and outside the World Bank before being tested in the summer of 2011.

First Test of Qualitative Surveys. Qualitative tests of the STEP household survey were undertaken in August 2011 in Bolivia, Sri Lanka, and the Yunnan Province of China. The aim of these tests was to try out a subset of innovative and critical questions deemed possibly hard to understand and therefore at risk of eliciting inaccurate or inconsistent responses. These surveys were carried out among respondents who would qualify for participation in the full survey; they consisted of about 25 interviews in each country. The analysis of their results led to some revisions.

Piloting of Full Surveys. The revised full surveys were piloted in Bolivia and Sri Lanka in October and November 2011. Particular attention was paid to time, flow, and clarity of the questions while administering the surveys to a diverse group of respondents. Suggestions were made to improve the survey instruments, in particular in order to shorten administration time. The employer survey was pre-tested in Sri Lanka and Vietnam in September 2011. The household and employer survey instruments, as well as related implementation materials, were finalized in December 2011.

Implementation. Data collection started in March 2012 for the first wave of countries. In this wave, both the household and employer surveys were implemented in Lao PDR, Sri Lanka, Ukraine, Vietnam, and Yunnan (province of China). In Bolivia and Colombia only the household survey was implemented. The second wave of implementation started in 2013. Both the household and employer surveys were implemented in Armenia, Georgia, and Macedonia. The household survey alone was administered in Ghana and Kenya, and the employer survey alone was implemented in Azerbaijan.

³ Neisser et al., 1996, p. 77

STEP's goal is also to measure human capital stocks, that is, skill supply. All adults, whether they work or not, are therefore asked a similar set of questions to measure labor force potential as well as skills used. The STEP household survey therefore collects background information on a participating household as well as detailed information on a randomly selected individual within the household (ages 15 to 64) regarding his or her skills acquisition history, educational attainment, work status and history, family background, and health. (Under skills acquisition history, the survey gathers detailed information on the individual's field of study of all reported degrees and certificates and any participation in apprenticeships, continuing education, or training). The household survey includes three unique modules to measure different types of skills: (i) an assessment of reading literacy designed to identify levels of competence at accessing, identifying, integrating, interpreting, and evaluating information; (ii) a battery of self-reported information on personality traits and behavior (conscientiousness, extraversion, self-control, decision making, and aggressive behavior) as well as risk and time preferences; and (iii) a series of questions on task-specific skills that the respondent possesses or uses in his or her job.

On the employer's side, STEP measures both work requirements and reported skill difficulties as indicators of the demand for skills, potential skill shortages, and work performance for sampled sectors of activity. National economic well-being is the outcome of the relative quality of the levels and match between the population and employment opportunities. The employer survey gathers information from a random sample of employers on hiring, compensation, and termination and training practices, as well as enterprise productivity. The survey includes questions to identify (i) employers' skill needs and utilization; (ii) the types of skills employers consider most valuable and the hiring mechanisms; and (iii) the tools used to screen prospective job applicants. The survey uses the same skills concepts and definition as those used in the household survey, a feature intentionally designed to facilitate analysis of skills gaps and mismatches.

The simultaneous measurement of skill stocks and job demands on both household and employer surveys is designed to give some indication of the levels of skill utilization and

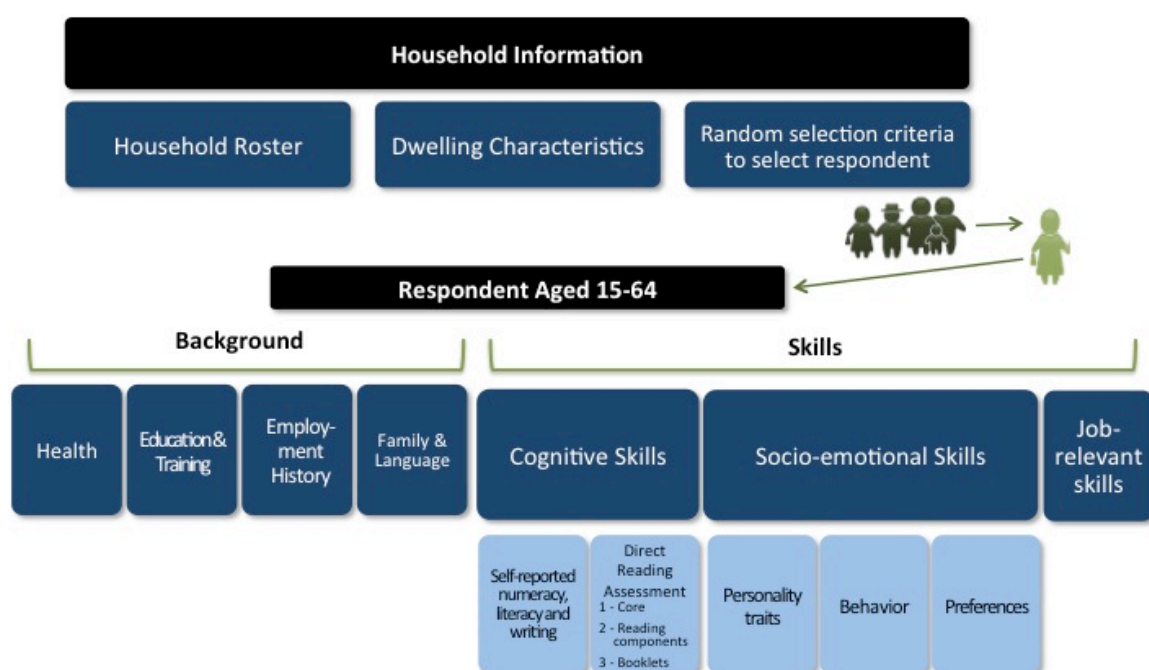
mismatch using comparisons of parallel measures relating to persons and jobs. Thus, both the household and employer surveys contain detailed measures of required education and experience and of the required skills in reading, writing, math, problem solving, interpersonal/socio-emotional traits, technology use, and manual work required by jobs. Comparing the worker- and job-side results will give some indication of the extent of any mismatch between the skills workers possess and those demanded by employers.

II. HOUSEHOLD SURVEY

1. Overview

The household survey seeks to obtain a wide range of information on personal background, education, employment and compensation, household wealth, household size and composition, personality, and personal health. The topics are summarized in Figure 1, and a detailed description of the information obtained is provided in Table 1.

Figure 1. STEP Household Survey: Structure



Note: See Figure 2 for a description of the random selection process

After implementing a relatively short household questionnaire, interviewers randomly select an individual within the household to answer the individual questionnaire. The respondent is then asked to take a literacy assessment at the end of the interview. Depending on whether they pass a basic assessment or not (and on whether the country is implementing a full literacy assessment or not), the respondent continues by taking an extended literacy assessment.

The individual questionnaire contains seven different modules, including three modules focused on skills, and is followed by a reading literacy assessment. The individual questionnaire modules cover a wide variety of personal characteristics and detailed question on skills (Table 1). Although many questions come directly from surveys that have been implemented in the past, a large number of questions have never been implemented in developing countries so far, and the more “traditional” modules are designed to obtain finer details on education pathways and employment trajectories. The rest of this section reviews the main innovative aspects of each module.

Table 1. STEP Household Survey: Topical Content, by Module

<u>Household Level Information</u>	
(a) Household Roster (Module 1a)	Names, age, gender, relationship to head for all household members Education status and self-reported literacy of all members aged 6 and over Marital and labor force status of all members aged 15 and over
(b) Dwelling Characteristic (Module 1b)	Dwelling construction materials, number of rooms, source of water and energy, toilets Tenure status Inventory of household consumer goods, appliances, and vehicles, number of books Ownership of bank accounts, receipt of social benefits
<u>Individual Respondent Information</u>	
(c) Education and Training (Module 2)	Participation in early childhood education Level of formal education and whether academic or vocational Field of study for highest qualification (13-15 categories) Reasons for dropping out (if applicable), Reason for interrupting schooling (if applicable) Apprenticeship (y/n) and trade Number of training courses, participation in literacy courses School class rank, parental encouragement
(d) Health (Module 3)	Overall life satisfaction Height, weight, health in last 4 weeks, chronic health problems and severity Health insurance coverage

(Table 1. cont'd.)

(e) Employment (Module 4)

Employment status, whether work on own account and casual work
Reason not working, job search methods, reason not looking for work (if not working)
Reservation wage, occupations for which qualified (if not working)
Occupation, tenure, industry, hours worked, other occupations for which qualified
Class of worker (wage/salary, daily or piecework, self-employed with/out employees)
Wage, salary, or profits per time period, in-kind payments
Employer (government, individual, domestic or foreign firm, NGO)
Establishment size, social benefits coverage

(f) Self-reported Cognitive Skills and Job-relevant Skills (Module 5)

Inventory of reading tasks performed on job (or in general), length of longest document read
Inventory of writing tasks performed on job (or in general), length of longest written document
Inventory of math tasks performed on job (or in general)
Whether lack of reading and writing skills hindered employment, promotion, or pay raise
Frequency of difficult problem solving on job
Level of involvement with customers, clients, students, or public on job
Make formal presentations as part of job
Supervisory responsibilities, job autonomy, repetitiveness, continuous learning
Level of physical job demands
Inventory of technology use on job (including computer use and inventory of software use)
Computer use outside work and inventory of software use
Whether lack of computer skills has hindered employment, promotion, or pay raise
Usefulness of own studies at school for current job
Level of education and related job experience required for current job, job learning time
Job search skills, whether employer required formal credentials or other proof of skills

(g) Personality, Behavior and Preferences (Module 6)

Thirty-one personality items on the frequency of diagnostic behaviors (e.g., extraversion)
Seven-item risk preference scale

(h) Language and Family Background (Module 7)

Native language, other specific language proficiency
Mother's and father's educational attainment
Family size, composition, and socio-economic status when 12 years old, adverse family events
Experience as child laborer, occupation

(i) Reading Literacy Assessment (Assessment booklets)

Core
Reading Components
Exercise booklets

(j) Interviewer Impressions (Modules 8-11)

Comprehension of questions, reliability and candor, distractions

2. Description of the Questions Module-by-Module

Household Level Information

Household Member Characteristics (Module 1a)

A simple household roster is complemented with age-appropriate questions on the educational attainment and labor market status of all household members. The idea is to get a full picture of the characteristics of household members that could influence the outcome of interest (such as obtaining a job) for the individual who will later respond to the full questionnaire. For example, it is well known that parents' educational attainment is correlated with children's attainment and labor force participation; and that household size is correlated with labor force participation.

Dwelling Characteristics (Module 1b)

The survey collects information on dwelling characteristics and household assets in order to construct an asset index to be used as a proxy for wealth. Since the focus of the survey is to obtain detailed information at the individual level, the household-level information is kept to a minimum. In particular, the survey interview procedure does not allow for sufficient time to collect information on consumption and expenditure. The survey does include questions on the quality of housing (type and ownership of dwelling; material of walls, floor and roof; source of drinking water; type of toilet; energy used for cooking and lighting), a common set of assets, and land and livestock ownership. The module is adaptable to country-specific circumstances; in particular the multiple-choice answers can be changed to reflect local customs and level of development. (See Box 2 for an explanation of the statistical procedure employed to derive asset index values from the reported information).

Box 2. Construction of the Asset-based Wealth Index

The asset index was constructed for urban areas and uses the information collected in Module 1b of the STEP Household Questionnaire.

The asset index was generated using Factor Analysis over a set of indicator variables for the different types of assets and dwelling characteristics. The use of indicator variables is required to maintain equal variances across different types of assets and characteristics. All national-level estimations were weighted using each country's sample weights.

Asset composition varies across countries in order to reflect underlying measures of welfare. However, the selection of the variables was systematically performed based on the following criteria:

- The variables with extremely skewed distributions were dropped. All variables with means across assets and dwelling characteristics below 0.02 and above 0.98 were considered extremely skewed and consequently excluded from the analysis.
- Not all agricultural assets were included. In this scenario, agricultural assets were considered as productive assets and not as an indication of wealth per se. An agricultural index was constructed, whenever possible, but these indexes showed low to poor correlation with an overall asset index.
- The variables with low factor loading (less than 0.1) on the un-rotated first factor of the overall asset index were excluded for the final asset index.

The asset index construction was performed on a country-by-country basis according to the following process:

Step 1. An indicator variable was created for each of the dwelling characteristics and assets available in Module 1b of the STEP Household Questionnaire.

Step 2. The variables that were not in compliance with the first selection criteria were dropped.

Step 3. An overall asset index was generated using factor analysis and it included all the available asset and dwelling related variables. In this stage, the factors with an Eigen value of more than 1 were selected.

Step 4. A varimax rotation was employed using the selected factors from the previous step.

Step 5. A Cronbach's Alpha (or Scale Reliability Coefficient) was estimated for this overall asset index.

Step 6. Indexes for each domain (dwelling characteristics, primary assets, and secondary assets) were constructed by following the same procedure from steps 3 to 5.

Step 7. A pairwise correlation was estimated for each of the domain indexes compared to the overall asset index to determine the level of association.

Step 8. Variables that did not meet the third selection criteria were dropped.

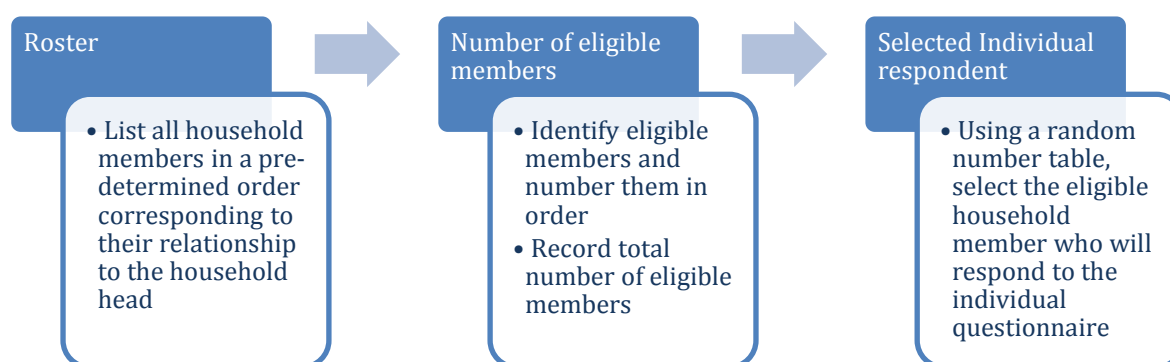
Step 9. A final asset index was constructed based on the factors with an Eigen value of more than 1.

By T. Sohnsen; text adapted by S. Monroy-Taborda.

Individual Respondent Information

The selection of the individual who will go on to answer the full questionnaire is closely monitored in order to avoid biases that can be associated with an inadequate selection of respondents by the interviewers. In particular, there may be a tendency by interviewers to simply continue the interview with any household member available at the time of the visit. The random selection therefore follows a specific protocol that can be easily monitored, both by supervisors in the field and by project managers at the central office location (see Figure 2).

Figure 2. Random Selection of the Individual Respondent



Education and Training (Module 2)

The overall aim of this module is to obtain a full picture of the acquisition of skills throughout the respondent's lifetime. The module does this by asking questions related to formal education (defined as diplomas and degrees that are awarded by formal educational institutions and officially recognized), lifelong learning, and other types of training and certificates.

The survey includes questions on participation in early childhood programs, the highest grade and level of education attended and completed, the highest degree obtained, the kind of educational program (academic, career, technical, vocational), the fields of study, other degrees and diplomas obtained in different fields, formal and informal apprenticeship, and any training recently undertaken. The questions allow for a deeper

analysis of the contribution of education to human capital formation by showing: (i) the relative weight of academic versus vocational education, (ii) breakdowns by field of study, and (iii) incidence of (formal and informal) apprenticeships and training in the population.

The rationale behind the collection of data at this level of detail is well established, and the set of indicators produced by this module provides a full understanding of the educational attainment of respondents, which goes well beyond the usual measure of “years of education.” There is a long-standing debate on the desirable balance between vocational and general education at the national level. Understanding the current state of affairs is therefore important to inform the process by which national governments formulate education strategies.

Broadly speaking, academic education is understood to confer general human capital, which is adaptive in volatile economies with frequent job changing and requirements for flexibility (general cognitive skills). However, academic credentials often require longer schooling and are relatively expensive, compared with vocational education, which provides more directly job-relevant education, assuming quality is sufficiently high and graduates find jobs in their fields. But vocational school graduates receive few benefits from their education if they find themselves working outside their training field.

There is growing recognition that the distribution of specific fields of study brings additional insights in understanding the qualifications of the workforce. Education level alone remains too broad a concept. It is impossible to know the true state of skill utilization or mismatch if one cannot compare field of study with type of occupation. Data about the field of study can guide policy makers in judging whether their country produces too few doctors, engineers, and information technology workers than needed or more lawyers than desired, for example, and where to direct support. The household survey collects field of study for both secondary and post-secondary qualifications, as well as other non-formal qualifications. Without such data, vocational and higher education policy lack a full pictures of the skills produced.

Given the uneven coverage of formal education in some countries and the unclear link between formal education and the job market, many people acquire skills through other channels. Apprenticeships and stand-alone training courses (such as in typing, software, and foreign language) are two of the leading sources of human capital formation outside formal channels, particularly among nonprofessionals. The skills acquired represent a potentially large proportion of total skill stocks, yet they are typically unrecognized in official statistics. Clearly, a country with low formal educational attainment but relatively institutionalized apprenticeships in civil society is better positioned than similar societies lacking such structures. The household survey is designed to take into account such atypical pathways and cross-country variation in the provision and use of these channels.

The survey includes a comprehensive set of questions characterizing the schooling attainment of the population. Among these variables, we can highlight “the highest level of formal education completed,” “education level when currently attending a formal education program,” “highest level of formal education before dropping out,” “fields of subjects associated with highest qualification,” and “type of school or institution attended.”

In order to measure educational attainment as accurately as possible, the choice was made to let countries adapt the questions on education to fit their national education system. The purpose of this approach is ensure both that respondents have a full understanding of the questions and that the interviewers are not second-guessing which broad categories they should fit respondents’ answers into. Whenever education reforms have led to changes in education levels, answer options include past denominations of levels so that respondents can easily identify which category best describes their attainments. Since the questionnaire is adapted to each country’s circumstances, the teams involved agree on a mapping of the national educational system to the International Standard Classification of Education 1997 (ISCED 1997) for the purpose of international comparisons.

The STEP survey uses the ISCED 1997 in order to bring comparability across countries and other adult literacy surveys. The classification is as follows:

- **ISCED0: Pre-primary education.** Programs at this level are designed primarily to introduce very young children to a school-type environment, that is, to provide a bridge between the home and a school-based atmosphere.
- **ISCED1: Primary education.** First stage of basic education. Programs at level 1 are normally designed on a unit or project basis to give students a sound basic education in reading, writing, and mathematics.
- **ISCED2: Lower secondary education.** Second stage of basic education. The contents of education at this stage are typically designed to complete the provision of basic education, which began at level 1. The educational aim is to lay the foundation for lifelong learning and human development. The end of this level often coincides with the end of compulsory education where it exists.
- **ISCED3: (Upper) secondary education.** This level of education typically begins at the end of full-time compulsory education for those countries that have a system of compulsory education. The age of entrance to this level is typically 15 or 16 years. The educational programs included at this level typically require the completion of some nine years of full-time education (since the beginning of level 1) for admission or a combination of education and vocational or technical experience and, as minimum entrance requirements, the completion of level 2 or demonstrable ability to handle programs at this level.
- **ISCED4: Post-secondary non tertiary education.** This captures programs that straddle the boundary between upper-secondary and post-secondary education from an international point of view, even though they might clearly be considered as upper-secondary or post-secondary programs in a national context.

Considering their content, ISCED 4 programs cannot be regarded as tertiary programs.

- **ISCED5: First stage of tertiary education** (not leading directly to an advanced research qualification). This level consists of tertiary programs having an educational content more advanced than those offered at levels 3 and 4. All degrees and qualifications are cross-classified by type of programs, position in national degree or qualification structures (see below) and cumulative duration at tertiary. These are further classified as Bachelor's degree (5A) and Master's degree (5B).
- **ISCED6: Second stage of tertiary education** (leading to an advanced research qualification). This level is reserved for tertiary programs, which lead to the award of an advanced research qualification. The programs are therefore devoted to advanced study and original research and are not based on course-work only.

The information on education provides a unique opportunity to analyze the differences in “skill competency” levels across multiple educational breakdowns. The following is a list of the potential educational breakdowns:

- Average skills by highest level attended
- Average skills by highest level of education completed
- Differences in average skill levels: academic versus vocational degrees
- Differences in average skill levels by field of study
- Differences in average skill levels by institutional arrangements (cross-country comparisons)

The survey includes a set of retrospective questions on individual's schooling history that will be useful to analyze key issues in education, such as effects on skill levels and skill

formation, including: (a) age at enrollment in first grade, (b) different types of parental investments, and (c) school-to-work transition.

Health (Module 3)

The research on human capital has noted the importance of health status in influencing the acquisition of skills (**Walker et al., 2007**). Health indicators for people of all ages are important because health affects the ability to learn and work. This means that health status is both an important precondition for accumulating additional human capital and a form of human capital itself, particularly for economies in which manual labor accounts for a significant share of jobs. At the same time, the kind of work an individual does affects his or her health status. Moreover, the health status of the population of a country is an important development indicator in its own right. While health is not a principal focus of STEP, the survey includes information to show the levels of a number of key health indicators for the population (Table 1).

Employment (Module 4)

One of the main outcomes of interest in the household survey is the labor market performance of respondents. The survey therefore takes a comprehensive look at past and current labor market outcomes of the selected individuals.

The survey obtains basic employment information as would be found in any employment module. With this information, researchers can identify the labor force status of the population (employed, unemployed, or inactive), the reasons for not participating, the employment search methods that were used to find the current job or that are currently used by the unemployed, details on the first job obtained after finishing formal schooling, details on the latest job held by those who are currently inactive, as well as details about the labor force status of individuals immediately prior to their current status.

For those who work, the survey inquires in detail about their occupation, earnings, hours worked, contract status, and benefits, distinguishing between employees and self-employed

workers. Better understanding the labor market experience of the self-employed is crucial in developing countries, where a large proportion of the labor force is self-employed, underemployed, or holding low-productivity jobs. For the self-employed, both those with and those without paid workers, the survey therefore asks a series of specific questions that help determine the overall success of their businesses. To help gauge the extent to which such work is voluntary, the survey asks respondents their preference for wage jobs versus self-employment; it also asks all workers about previous self-employment experiences, if any.

Self-reported Cognitive Skills and Job-relevant Skills⁴ (Module 5)

Background. The STEP survey focuses on skill sets with direct job-relevance, whether respondents are currently using them in the labor market or not. An influential conceptualization of job skills requirements, in the United States as well as internationally, is the Dictionary of Occupational Titles (DOT) published by the United States Department of Labor. The DOT is an employment counseling tool based on expert job analysts' ratings on many dimensions. One important feature of this dictionary is the division of job-relevant skills according to their level of involvement with "Data, People, and Things." These categories correspond to cognitive, interpersonal (or interactive), and manual (or physical) skills. This scheme has been validated formally numerous times (e.g., Kohn and Schooler 1982; Peterson et al. 1999; Autor, Levy, and Murnane 2003; Autor and Handel 2009).

For illustration, the content of each category can be specified in greater detail as follows:

- *Cognitive skills:* required level of education, reading, writing, math, scientific/technical knowledge, general reasoning or problem-solving skills
- *Interpersonal skills:* customer service, team decision making, formal presentations
- *Manual skills:* levels and kinds of physical effort (such as standing, lifting, carrying); use of simple and complex tools, machinery, materials, and equipment

⁴ See Handel (2012).

Cognitive skills. Cognitive skills have been shown to directly affect certain labor market outcomes such as wages (Autor and Handel, 2009). The widely read reports from the U.S. Secretary's Commission on Achieving Necessary Skills (SCANS) identified certain types of skills (reading, writing, math, problem solving) as “foundational” cognitive skills that are critical outputs of the school system.⁵

Job-related interpersonal skills. Job-related interpersonal skills are much less well-theorized and well-measured than cognitive skills. Even at the most basic level, this domain is weakly conceptualized and there is low agreement on the elements properly included in the domain. The literature includes communication skills, courtesy and friendliness, service orientation, caring, empathy, counseling, selling skills, persuasion and negotiation, and, less commonly, assertiveness, aggressiveness, and even hostility, at least in adversarial dealings with organizational outsiders (e.g., for police and corrections officers, bill collectors, some lawyers and businessmen).⁶ If one were to include informal job demands that might arise in dealing with co-workers, the list would also include leadership, cooperation, teamwork skills, and mentoring skills.

These elements are qualitatively diverse, rather than falling along different levels of a single trait. Many of them could be considered ancillary job characteristics that, while often useful, are exercised at the discretion of the employee, rather than being job or employer requirements. Often it is not easy to separate interpersonal skills from more purely attitudinal and motivational aspects of work orientations.

On a practical level, survey questions about interpersonal demands produce very high rates of agreement and low variance if they do not distinguish relations with co-workers from those with organizational outsiders, such as customers and clients. Pretests on other surveys (see below), as well as STEP, showed that many people have a tendency to respond to questions reflexively, for example stating that working always requires a positive

⁵ United States Department of Labor. Secretary's Commission on Achieving Necessary Skills. 1991

⁶ See National Center for Occupational Information Network (O*NET) Development, sponsored by the US Department of Labor, Employment and Training Administration. 1991.

attitude, willingness to cooperate with others, and so on. For this reason, questions in the STEP survey were precise and concrete, introducing elements of context (for example: “Do you spend time cooperating or collaborating with coworkers?” and “Do you have any contact with people other than coworkers, for example customers, clients?”)

Manual job requirements. Manual job requirements are defined here as bodily activities that usually involve materials, tools, and equipment. Simple physical tasks include gross physical exertion (such as carrying heavy loads), elementary movements (such as sorting mail), use of simple tools or equipment, and machine tending. More complex physical tasks, which are widely found in craft jobs, require more training, experience, and background knowledge regarding the properties of physical materials, mechanical processes, and natural laws.⁷

Although there are elaborate taxonomies of physical job requirements, survey space limitations and manual skills’ weak or negative relationships with wages argued against their full incorporation into the STEP survey (Rotundo and Sackett, 2004; Autor and Handel, 2009). Insofar as some complex manual skills might command significant wage premiums, these are captured in STEP through extensive checklists of field of study or practice for those reporting an apprenticeship or technical/vocational education.

Item Selection, Reliability and Validity. The STEP items are drawn from the survey of Skills, Technology, and Management Practices (STAMP), a two-wave, nationally representative panel survey of U.S. wage and salary workers funded by the National Science Foundation (for details see Handel, 2008a). STAMP drew on an extensive literature in sociology, labor economics, industrial relations, human resource management, psychology, and education (including Cook et al., 1981; Milkovich and Newman, 1993). The DOT and its successor, the Occupational Information Network (O*NET), have a wealth of items and concepts. The United Kingdom Skills Survey was also consulted.

⁷ See National Center for Occupational Information Network (O*NET) Development, sponsored by the US Department of Labor, Employment and Training Administration. 1991.

Based on the STAMP methodology, STEP adopted an approach called explicit scaling, which favors measures that are as objective as possible and have absolute meanings for respondents. Questions are phrased in terms of facts, events, and behaviors, rather than attitudes, evaluations, and holistic judgments. Items are general enough to encompass the wide range of jobs within the economy, but sufficiently concrete that they have stable meanings across respondents. The response options aim to avoid floor and ceiling effects and use natural units when possible. Rating scales, vague quantifiers, and factor scores that have arbitrary metrics and lack specific or objective referents are used only as last resorts, whereas items such as those above are more interpretable than the alternatives and likely contain less measurement error.

Cognitive skills. Module 5 starts with interview questions about the foundational reading, writing, math, and problem-solving skills people use, both on their jobs and outside of work. These questions are modifications of similar items from the January 1991 supplement to the Current Population Survey (CPS), conducted by the U.S. Bureau of the Census. Eight items measure the frequency of various reading, writing, math, and computer tasks using a frequency scale (never, less than once per week, one or more times per week, every day). Very similar measures were used in Holzer's (1996) four-city survey of employers and in the National Adult Literacy Survey (Sum, 1999, pp.133ff.), which was conducted by the Educational Testing Service for the U.S. Department of Education and served as a model for the assessment portion of OECD's Programme for the International Assessment of Adult Competencies (PIAAC) study and STEP's assessment.

Like STAMP, the STEP survey assesses both the frequency and the level of complexity of reading, writing, and math use, whereas other surveys have only focused on frequency of use. This is important, because two jobs with the same frequency of use for math or reading can rely on vastly different *levels* of math or reading skills.

Questions on whether a lack of literacy skills has ever prevented the respondent from getting a job, promotion, or raise, addresses to both workers and non-workers, can provide some indication of mismatch and the unmet need for basic skills upgrading.

Of course, a wide range of unspecified skills that go beyond literacy/numeracy are imparted through education and any major field of study. A summary survey item that tries to capture much of this general cognitive skill that would otherwise go unmeasured is the question on complex problem solving on the job. Analyses of STAMP data show general problem-solving skills, derived from both school and general life experience, have high reliability and validity (Handel, 2008b).

Further indication of the usefulness of the cognitive skills measures described above is their growing use in other major surveys. The STAMP items on reading, writing, math, and problem solving have been used verbatim or in slightly modified form in the following surveys: the Princeton Data Improvement Initiative, the OECD's PIAAC, and the National Educational Panel Study in Germany. The growing use of these items across countries means there will be some basis for comparing STEP countries with high-income countries on these dimensions, along with the same kind of comparisons planned for countries participating in the assessment that overlaps with PIAAC.

Interpersonal skills. The STEP survey assesses the use of interpersonal skills at work through questions related to teamwork, supervision, contact with customers, as well as internal or external communication via presentations. In several instances, depth of involvement in using such skills is also captured to provide a richer set of data.

Manual skills. Questions ranging from whether respondents use physical strength at work to whether they operate certain types of heavy machinery or repair electronic equipment are also included in Module 5.

Diverse types of job-relevant skills are found in the labor market (such as administering intravenous drugs, calculating net present values, operating a pneumatic jackhammer).

Although these are critical skills, they are usually unmeasured in general labor force surveys because they are too numerous and apply individually to a small fraction of the workforce. Even if it were possible to include lengthy checklists of occupational skills, the information could not be used for analysis because the qualitative diversity of such skills prevents their conversion to common units on a scale. The usual alternative, and the one used by STEP, is to ask about prior job experience and length of time required to learn a job for the average person with the required education (job learning time). This helps to measure non-academic, job-relevant skill demands across jobs on common, absolute scales, in keeping with the principles of explicit scaling.

Exceptions are made to the preceding approach to identify job-relevant skills of moderate generality, such as the use of computer and other technology skills. Because these skills are relatively common, it is possible to include them in a general household survey, though naturally they will not apply to everyone. Like STEP, PIAAC also use some of these items, derived from STAMP, relating to computer use and skill adequacy. Analyses using STAMP data show that these items also have high reliability and validity (Handel 2008b).

Analyses based on the STAMP survey also show that the education required for a given job is a very strong predictor of wages, indeed a more powerful predictor than respondents' own personal education. These measures are also included in Module 5 and can help to establish a diagnostic of the overall level of job skill requirements and the extent of educational mismatch.

The STEP survey is the first attempt to gather this type of information and systematically measure job-relevant skills in developing countries. Items from the STAMP survey were chosen and adapted in order to ensure they were relevant in a developing country context. For example, the list of technology tools was restricted to items commonly used in such countries. Moreover, given the possibility that the STEP survey would be implemented in a rural context, job-relevant skills from the agricultural sector were also included.

Personality, Behavior, and Preferences (Module 6)

Background. Research by James Heckman and other economists in OECD countries in the past 15 years has conclusively demonstrated the importance of personality traits—such as conscientiousness, persistence, work motivation, extraversion, emotional resilience, ability to work with others, and willingness to bear risk—in determining labor market and other educational outcomes over an individual’s lifetime. These studies have benefitted from advances in psychology research in developing reliable measures of these traits and behaviors. The STEP program extends this line of inquiry to the study of the importance of these traits in developing countries. Its hypothesis is that individuals scoring high on pro-social attitudes and achievement motivation will not only exhibit more favorable economic and life-satisfaction outcomes, but will be better positioned to work in non-manual jobs, whose relative number is one measure of economic development.⁸

Personality traits are defined as patterns of thinking, feeling, and behaving that are relatively stable across time and situations. They have recently been recognized as important predictors of economic outcomes (Borghans et al., 2008; Paunonen, 2003). In particular, the *Big Five* taxonomy of personality traits is now widely accepted as the organizational structure of personality traits. This taxonomy has been replicated across cultures (John and Srivastava, 1999) and developmental stages of the life course (Soto et al., 2008).

⁸ Although these personality traits are a mixture of motivational and attitudinal traits, on the one hand, and what could be called interpersonal skills, on the other, they will be referred to under the umbrella term “skills” for convenience.

The Big Five taxonomy consists of

- Conscientiousness
- Openness to experience
- Neuroticism
- Agreeableness, and
- Extraversion.

Each of these five represents a family of more narrowly-defined, related yet distinct traits.

The STEP program engaged with several social scientists, including economists and psychologists, and discussed the rationale for including behavior and personality trait constructs in a skills measurement survey. For the purpose of the program, these constructs were defined as individual differences that are independent of cognitive ability. Drawing upon recent reviews (e.g. Almlund et al., 2011), the idea that such constructs have an incremental predictive validity above and beyond cognitive skills was validated. In particular, personality traits predict the same positive economic, social, and health outcomes as cognitive ability does. Likewise, there is evidence for the causal roles of biases (such as hostility attribution bias) and styles of decision-making (such as generating solution and considering future consequences) in determining the same outcomes.

Several social scientists proposed constructs for STEP's use, their rationale for inclusion, and specific items. Given the constraints imposed by the size of the survey, the principles guiding the ultimate selection of items included: the applicability and comprehension of the items in low-literacy cultures where people have little or no experience answering self-report questions; brevity; prior evidence of scale reliability and validity; and prior evidence of predictive validity for important outcomes, particularly in other large-scale surveys. Leading personality psychologists were consulted in this work. Each was asked to recommend constructs to be included, previously published scales corresponding to these

constructs, and response scales that would maximize variance while minimizing cognitive load and the possibility of misunderstanding. To formalize a set of social-emotional skills, leading developmental and social psychologists with expertise in assessment and intervention were also consulted.⁹

Once a final proposal was complete, pilots were conducted in several countries by the World Bank to identify problems of administration and suggest calibration of item wording. Feedback from these pilots led to several important adjustments, particularly in the rewording of items that had proven to be difficult for participants to understand and the general reframing of all items as questions (using a four-point frequency scale from *almost always* to *almost never*) instead of statements. Asking participants to answer questions rather than endorse statements seemed to be more “natural” in low-literacy populations. For example, participants felt more comfortable answering the question “When doing a task, are you very careful” than endorsing the statement “I see myself as someone who does a thorough job.”

To verify the convergent validity of the reworded items with their corresponding original versions, Angela Duckworth conducted an online validation study in which several hundred non-U.S. adults completed both sets of items. Internal reliability, discriminant validity, and convergent validity estimates from this study were considered when making final choices for items in the battery of questions assessing behavior and personality traits.

Instrument Design. The STEP survey measures socio-emotional skills through a series of items—Grit, Hostile Attribution Bias, and Decision Making—which are related to the Big Five personality trait factors described above. Each of the Big Five factors is assessed with three items in the short Big Five Inventory (BFI-S) originally developed by John and

⁹ Angela Duckworth (University of Pennsylvania) and Nancy Guerra (University of California) led the preparation of the items for STEP. The leading personality psychologists they consulted included Oliver John (University of California at Berkeley), Brent Roberts (University of Illinois at Champaign-Urbana), Gerard Saucier (University of Oregon), and Veronica Benet-Martinez (University of California at Riverside). The leading developmental and social psychologists with expertise in assessment and intervention included Kenneth Dodge (Duke University), Patrick Tolan (University of Virginia), and Roger Weissberg (University of Illinois at Chicago).

Srivastava (1999) and later validated in large-scale panel surveys (such as the GSOEP German panel survey; see Lang et al., 2011). Specifically, the domain of conscientiousness has been defined as “the propensity to follow socially prescribed norms for impulse control, to be goal directed, to plan, and to be able to delay gratification and to follow norms and rules” (Roberts et al., 2009, p. 369). This domain is assessed in the questionnaire with items such as “When doing a task, are you very careful?” Openness to experience refers to enjoyment of learning and new ideas (such as, “Do you come up with ideas other people haven’t thought of before?”). Neuroticism refers to the tendency to feel negative emotions (such as, “Do you worry a lot?”). Agreeableness refers to a pro-social, cooperative orientation to others (such as, “Do you forgive other people easily?”). Extraversion encompasses sociability and dominance in social situations (e.g., “Are you talkative?”).

The Big Five factors are *broad* families of personality traits, with component facets of varying relevance to particular outcomes. One motivation for investigating with higher-resolution measures through more narrowly specified facets is that they often demonstrate incremental predictive validity for relevant outcomes (Paunonen and Ashton, 2001). For example, the construct of *grit*, defined as trait-level perseverance for long-term goals, has been shown to provide incremental predictive validity over and beyond the Big Five factor of conscientiousness for objective measures of professional and educational achievement. To this end, the STEP survey includes three items assessing grit, from the Grit Scale (Duckworth et al., 2007). One of these, for example, is “Do you enjoy working on things that take a very long time (at least several months) to complete?”

The working group on behavior and personality traits considered a range of socio-emotional skills for inclusion in the study based on previous empirical studies and time allocated for the administration of this component of the survey. The initial scales included four items from the Rosenberg Self-Esteem scale, four items adapted from the Melbourne Decision-Making Scale (Mann et al., 1997), and six items selected by Kenneth Dodge (2003) from previous assessments of hostile attribution bias. Because of time constraints and difficulties in understanding items in the pilot sample, the self-esteem items were dropped and the

hostile attribution bias items were reduced from six items to two items. Two of the decision-making items tap alternative solution-thinking, a more controlled style of information processing that involves consideration of multiple options when making decisions. The other two items tap consequential thinking, particularly the extent to which individuals think about the future consequences of their decisions and actions on themselves and others. For example, “Do you think about how things you do will affect you in the future?”

The scoring for each of the scales (including the Big Five Conscientiousness, Openness to Experience, Neuroticism, Agreeableness, and Extraversion; Grit; Hostile Attribution Bias; and Decision-Making) is straightforward. For positively scored items, a score of 4 is assigned to signify “almost always,” 3 to signify “most of the time,” 2 to signify “some of the time,” and 1 to signify “almost never.” For negatively scored items (such as the Extraversion item “Do you like to keep your opinions to yourself?”) a score of 4 is assigned for “almost never,” and so on. These scores are averaged for each of the scales. Items also were revised to eliminate any double negatives (such as “people don’t like me” with “almost never,” meaning that “people like me”); although the double-negative strategy had been used to reduce response set (always answering “almost never”), the wording simply was too confusing for respondents to understand.

Reliability of the Scales. The internal reliability of each scale was assessed separately by calculating the Cronbach alpha statistic for the relevant items. Consistent with Lang et al. (2011), the STEM project team expected alphas to be in the range of .50 to .65 because of the breadth of conceptual territory covered by each scale and the relatively small number (three) of items. However, results from the first five countries involved in the study showed much lower alphas, ranging from 0.14 in Vietnam to 0.47 in Bolivia.

A number of checks were performed on the data. In particular, bivariate associations were examined. Consistent with prior research (e.g., Lang et al., 2011), these correlations revealed small to moderate ($r = 0.1$ to 0.4) correlations among the included scales, with

conscientiousness, agreeableness, and neuroticism (inversely) correlating more highly with one another than with other scales. Openness to experience should be more highly correlated with objectively measured cognitive ability than any other Big Five factor. Hostile attribution bias should correlate (inversely) with Big Five agreeableness.

Bivariate correlations with education and labor market variables also revealed results that were consistent with what could be expected.

All in all, after reviewing these findings as well as the relevant methodological literature, and given the fact that the scales used in the STEP surveys are short and cover broad constructs, Angela Duckworth concluded that the low alphas were collectively a function of three factors. The first is the low literacy of respondents and their unfamiliarity with such self-report, holistic "this is who I am" measures; this explanation would be consistent with the fact that coefficient alphas looked much better in Yunnan and Bolivia than in other countries, and dramatically better in countries like Germany and the United Kingdom. The second factor is the use of reverse-coded items. The third factor is the use of four response options (rather than the five of the original scale). However, the distributions of these variables were not deemed particularly problematic, and the pattern of predictive validities was overall considered satisfactory.

The team discussed potential revisions to the items. In order to keep consistency across the country datasets, the revisions consisted in rewording the reverse-coded items to be positively coded. Two items were therefore rewritten and added as new items. These changes were made to the questionnaires administered in a second wave of countries, namely Armenia, Georgia, Ghana, Kenya, and Macedonia.

Language and Family Background (Module 7)

The household survey collects information on languages, parental socio-economic status and resources, family size, early childhood development. A series of questions aims at obtaining a full picture of the languages that dominate in the household, and more

specifically that the respondents know, with a focus on the languages that they speak and write to a level that would enable them to use the language on the job. The module goes on to identify the children of the respondents who live in the household, if any; and the education level of the respondents' parents. Several questions ask the respondents to think back to when they were 12 and provide the number of siblings who lived in their household at the time – with details on seniority and gender, as well as the relatives who lived with them at that time. Finally the module asks respondents about the socio-economic status of their household when they were 15 years old, and whether any negative shocks impacted their household by that time; it also asks whether they worked before the age of 15 and at what age they first worked outside their home.

The variables that are found in this module can be analyzed as such to examine their link with skill formation and labor market outcomes, and they can also be used as instrumental variables.

Reading Literacy Assessment (Assessment Booklets)

The modules that have been described so far, which can be thought of as the “background questionnaire,” are all implemented through a face-to-face interview. This involves the active engagement of the interviewer, who is supposed to build a rapport with the respondent. The final part of the survey requires a change of pace; the respondent is now asked to sit alone and complete an assessment, without any help from the interviewer. This can be challenging, especially for adults, who are not used to taking tests anymore. This section describes this last module, a reading assessment that was developed by the Educational Testing Service (ETS).

Background. The relevance of large-scale literacy assessments has been growing in recent years as policymakers and other stakeholders have increasingly come to understand the critical role that foundational skills play in allowing individuals to maintain and enhance their ability to meet changing work conditions and societal demands. Findings from these assessments have contributed to informing policymakers by providing a wealth of

information about the distribution of foundational skills and their links with social and economic outcomes.

The STEP program is in line with a series of large-scale international surveys, which in addition to PIAAC include the International Adult Literacy Survey and the Adult Literacy and Life Skills Survey. These surveys contribute to an increasingly broader understanding of what it means to be “literate” in complex modern societies. Earlier work in the adult literacy assessment field had defined literacy by the attainment of certain scores on standardized academic tests of reading achievement. Such approaches are usually limited, since they tend to focus on school-age individuals and provide little information on individuals’ ability to navigate real-life and work-related materials. Assessments were improved through the use of competency-based tests, which employed non-school materials from adult contexts. Despite this improvement, these competency-based assessments still viewed literacy along a single continuum, defining individuals as “literate” or “functionally illiterate” based on where they fell along that continuum.

PIAAC, the International Adult Literacy Survey, and the Adult Literacy and Life Skills Survey all broadened the concept of literacy to reflect the diversity of tasks that adults encounter at work, home, school, and in their communities. Moreover, their reporting of findings has focused on types as well as levels of literacy. By design, STEP is directly linked to these surveys. As mentioned earlier, the STEP literacy assessment has been developed specifically for use in the context of developing countries, and it includes sets of questions taken from the three international surveys just mentioned. This overlap allows countries participating in the STEP program to compare their literacy results with those of over 30 other countries.

In developing the literacy assessment, ETS and the World Bank team had to account for a number of constraints due to the nature of the STEP program. Apart from having to adapt tools that had been used in the context of developed countries to the reality of developing countries, the primary constraint faced by the team was time: given the scale of the survey and the fact that it would be administered to populations with potentially low-level literacy,

it was important that the assessment could be administered in no more than 45 minutes. To accomplish this, some choices had to be made. Previous surveys have included multiple domains in their assessment; for example, PIAAC uses the domains of literacy, numeracy, and problem solving. While it is desirable from a policy standpoint to include a full range of adult competencies in a survey, given the time constraints for STEP a choice was necessary: between measuring one domain well versus measuring two or more domains with less precision than would be acceptable.

The team therefore decided to focus on the reading literacy domain. This domain was selected both because it has a strong relationship with a number of outcomes assessed in the rest of the STEP Skills Measurement Survey and because it is less dependent on formal education. Numeracy, for example, tends to be more dependent on the specific number and types of math courses that adults have taken. Moreover, reading literacy is the foundation that allows individuals to develop the full range of skills they need in order to meet today's rapidly changing workforce and societal demands.

The STEP conception of literacy is based on the same concept used in previous large-scale assessments, where it has been defined as “understanding, evaluating, using and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential” (PIAAC Literacy Framework). This definition gives us a broad understanding of the processes and goals of literacy as measured in STEP. The main aspects of the construct—contexts for reading and underlying cognitive processes required to complete the presented tasks—were taken into consideration when selecting the texts and developing items included in the assessment.

Contexts for Reading. For adults, reading is normally part of a social setting. Both the motivation to read and the interpretation of the content may be influenced by the context and the purpose for reading. As a result, a fair assessment must include material from a broad range of settings, so as to include some material that would be familiar to any participant. Therefore, the texts included in the STEP assessment comprise the following

contexts: home and family, health and safety, community and citizenship, work and training, education, and leisure and recreation.

Cognitive processes with text. STEP builds on three broad types of tasks that readers were asked to carry out in both PIAAC and the International Adult Literacy and Adult Literacy and Life Skills surveys: tasks that require the identification of pieces of information in the text, those that require connecting different parts of the text, and those that require some understanding of the text as a whole. Working on items or tasks can involve the following three cognitive operations:

- Access and identify information in the text
- Integrate and interpret (relate parts of text to each other), and
- Evaluate and reflect (understanding of the text as a whole).

As an extension of the core literacy assessment, STEP also includes an assessment of *reading components*. The Reading Components Assessment Framework builds on the basic principle that comprehension processes—that is, the “meaning construction” processes of reading—are built on a foundation of component print skills that indicate the knowledge of how one’s language is represented in one’s writing system. The following reading components were identified for the assessment:

- Word meaning (print vocabulary)
- Sentence processing, and
- Passage comprehension.

The assessment of reading components aims to provide information on the reading abilities of adults with poor skills in order to get a proper understanding of their difficulties. Evidence of an individual’s level of print skill can be captured through tasks that examine a reader’s ability and efficiency in processing the elements of the written language, namely

letters/characters, words, sentences, and larger, continuous text segments. Component efficiency is typically indexed by assessing speed or rate of processing as well as accuracy. In this assessment, speed or rate is approximated by the time it takes to complete certain tasks.

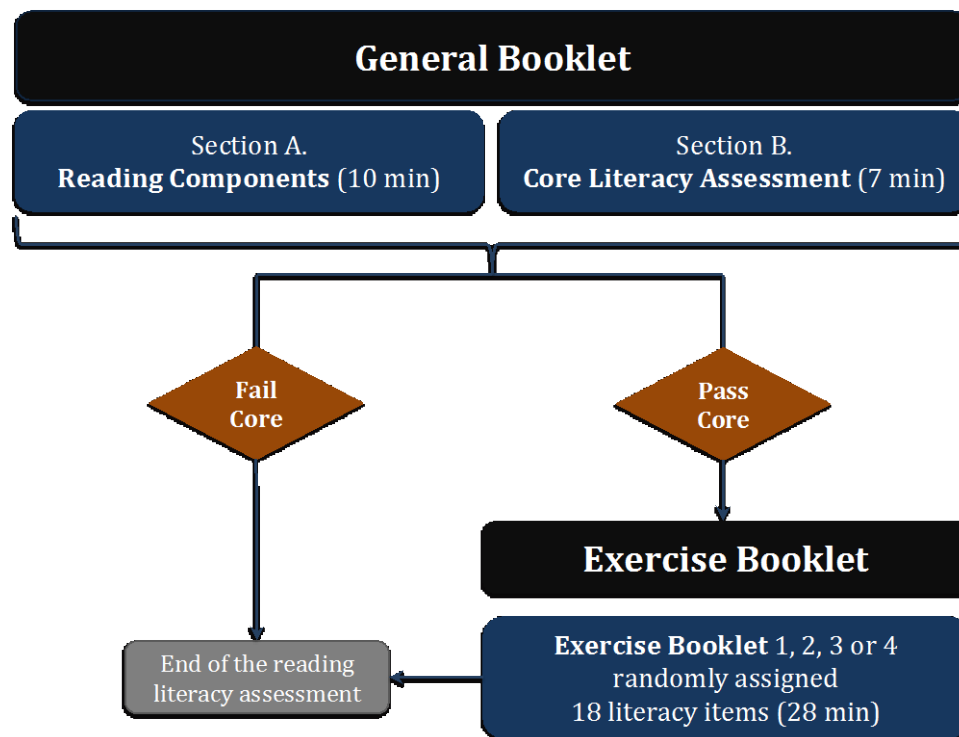
Assessment and Instrument Design. The STEP literacy assessment consists of two booklets: The General Booklet, which includes a Reading Components section and a Core Literacy Assessment section, and the Exercise Booklet. The Reading Components section focuses on foundational reading skills. The Core Literacy Assessment, which consists of eight basic literacy questions, is a screener, intended to sort the least literate from those with higher reading skill levels. Individuals who cannot successfully answer three out of the eight questions are not asked to go on and attempt the harder questions, in large part because such a requirement would likely put them in an uncomfortable or embarrassing position. Those who do pass this core assessment take one of the four Literacy Exercise Booklets developed for STEP, with each individual taking 18 items in total.¹⁰ The workflow for the STEP assessment is illustrated in Figure 3.

Each part of the assessment helps build a full picture of the level of literacy in the country. Administering the Reading Components to every respondent allows us to obtain targeted information about the skills of individuals at the lower end of the literacy distribution, meeting an important goal of the STEP survey. The Exercise Booklets, which include items covering the full range of difficulty, allow the survey to profile the full distribution of literacy skills in the adult populations of participating countries. Since the assessment is administered to adults who may not have been in an exam situation for a long time, most assessment materials were taken from real-world resources such as newspaper and magazine articles, advertisements, books, and forms that adults ages 15-64 would

¹⁰ Of the 12 countries participating in the STEP household survey so far, eight administered the full literacy assessment as it is represented in Figure 3. Four countries (Lao PDR, Macedonia, Sri Lanka and Yunnan Province) conducted what is called the “partial literacy assessment,” in which only the General Booklet (Reading Components and Core Literacy) was administered.

encounter in a range of everyday life contexts. Given the international context of the assessment, care was taken to select materials appropriate across cultures and languages.

Figure 3. Workflow for the STEP Reading Literacy Assessment



General Booklet. As mentioned, the General Booklet consists of two sections: Section A - Reading Components; and Section B - Core Literacy Assessment.

Section A: Reading Components. This section entails a set of reading component items aimed at providing countries with more detailed information about respondents who perform at the lower end of the reading literacy scale. It contains three parts: Part 1, Word Meaning (Print Vocabulary); Part 2, Sentence Processing; and Part 3, Passage Comprehension. Altogether, the reading components section takes approximately 10 minutes to administer.

World Meaning (Print Vocabulary). The Word Meaning (Print Vocabulary) measure is useful to determine whether individuals can identify, in print, words in the everyday listening lexicon of average adult speakers of the language. That is, the emphasis is on the everyday words of the spoken language. Each item in this section presents an image and four word choices. The respondent must circle the correct word choice that matches the picture. Target words are concrete, image-able nouns of common objects. “Distractors” were designed to tap similar semantic and/or orthographic features of the target word. Figure 4 provides a sample Word Meaning item.

Figure 4. Word Meaning: Sample Item

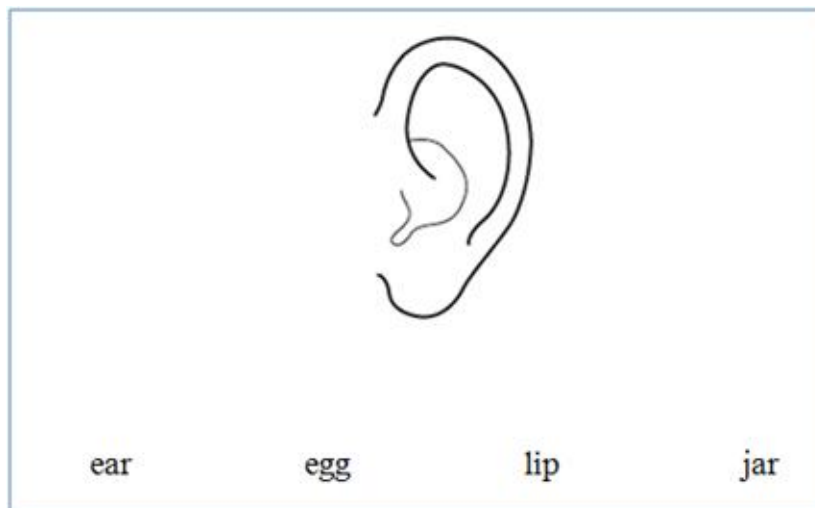


Figure 5. Sentence Processing: Sample Items

Three girls ate the song.	YES	NO
The man drove the green car.	YES	NO
The lightest balloon floated in the bright sky.	YES	NO
A comfortable pillow is soft and rocky.	YES	NO
A person who is twenty years old is older than a person who is thirty years old.	YES	NO

Sentence Processing. The Sentence Processing measure presents sentences of increasing difficulty (as indexed by length) and asks the respondent to make a sensible judgment about the sentence with respect to general knowledge about the world or about the internal logic of the sentence. For these items, the respondent reads the sentence and circles YES if the sentence makes sense or NO if the sentence does not make sense. Figure 5 shows a set of sample Sentence Processing items.

Passage Comprehension. The Passage Comprehension measure includes three passages each with embedded cloze items. Passages were constructed based on the kinds of text types that adults typically encounter: narrative, persuasive, and expository. The design uses a forced-choice cloze paradigm—that is, a choice is given between a word that correctly completes a sentence in a passage and an option that is incorrect. The incorrect item is meant to be obviously wrong to a reader with some basic comprehension skills. The integration of decoding, word recognition, vocabulary, and sentence processing is required to construct the basic meaning of a short passage. The respondent is asked to read the passage and circle the word that makes the sentence meaningful (in the context of the

passage). Fluent, efficient performance on such a basic, integrated reading task is a building block for handling longer, more complex literacy texts and tasks. A sample passage is shown in Figure 6 with the options for selection underlined within the sentences.

Figure 6. Reading Comprehension: Sample Passage

To the editor: Yesterday, it was announced that the cost of riding the bus will increase. The price will go up by twenty percent starting next wife / month. As someone who rides the bus every day, I am upset by this foot / increase. I understand that the cost of gasoline / student has risen. I also understand that riders have to pay a fair price / snake for bus service. I am willing to pay a little more because I rely on the bus to get to object / work. But an increase / uncle of twenty percent is too much.

This increase is especially difficult to accept when you see the city's plans to build a new sports stadium. The government will spend millions on this project even though we already have a science / stadium. If we delay the stadium, some of that money can be used to offset the increase in bus fares / views. Then, in a few years, we can decide if we really do need a new sports cloth / arena. Please let the city council know you care about this issue by attending the next public meeting / frames.

Section B: Core Literacy Assessment. This entails a core set of eight literacy items that can be used to help sort the least literate from those with higher levels of reading skill. The Core assessment takes approximately seven minutes, on average, to administer. In countries in which the full literacy assessment was administered, this core set of cognitive items was scored by the interviewer as soon as the respondent finished it. Respondents who failed the Core (those who failed to answer three or more items correctly) were done with the interview. Those who passed the Core proceeded to the reading Exercise Booklets. In

countries that opted for the partial literacy assessment, scoring of the eight literacy items in the Core was undertaken at a later stage by the survey firm, when it scored the rest of the General Booklet items.

Exercise Booklets. The assessment design for STEP specified a core literacy block consisting of the easiest items (administered in the General Booklet, Section B) and four additional blocks of literacy items. Similar to the items in the Core, the items in the Exercise Booklets assess reading literacy, covering the full range of difficulty. Respondents who passed the Core were randomly assigned one of these four booklets.

The four Exercise Booklets (Booklets 1, 2, 3, and 4) were assembled following the design provided in Figure 7. Each booklet has two blocks of nine literacy items, or 18 items in total. The booklets require 28 minutes, on average, for participants to complete. Sample items for the Exercise Booklets are presented in Appendix 1.

Figure 7. Exercise Booklet Design for STEP Literacy Items

	Block A	Block B	Block C	Block D
Booklet 1	x	x		
Booklet 2		x	x	
Booklet 3			x	x
Booklet 4	x			x

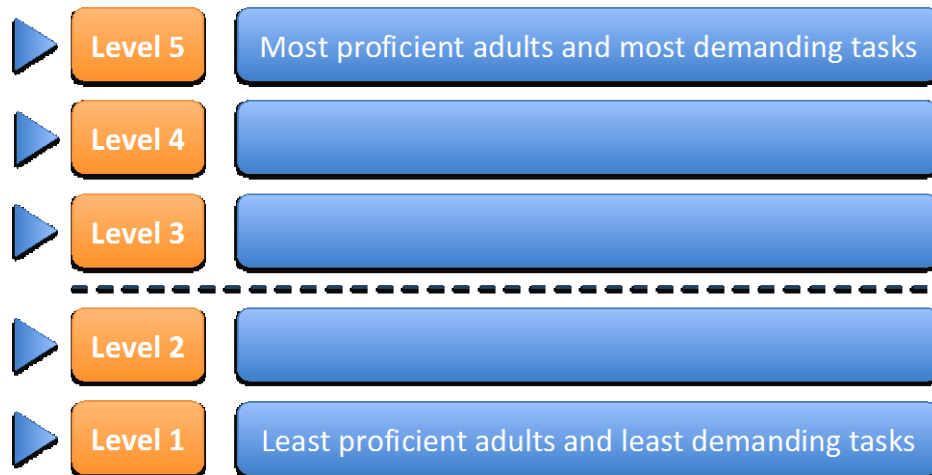
Since the assessment would be too long if every individual had to take the entire battery of items, the Exercise Booklet component was designed to divide the assessment into partially linked booklets. This method is common in large-scale assessment. This reduces the probability of some external factors interfering with the assessment (for instance, time of interview, interviewee burden and interruptions, among others).

In addition, to ensure that the pool of literacy items for STEP covered the full domain used in PIAAC, the design of the four exercise booklets used a variation of matrix sampling. As shown in Figure 7, in this methodology the domain of items is organized into blocks, and these are assembled into booklets so that each item is administered to a substantial number of respondents and each pairing of items is presented to some respondents. This allows for the computation of the correlation between any pair of items. In STEP, each of the four exercise booklets contains two blocks, each with nine items, for a total of 18 items per booklet.

The pool of items is then scaled using Item Response Theory (IRT). This methodology is fundamental to summarizing data in a meaningful way and is a preferred alternative to computing the percent of items answered correctly. Under certain assumptions, it also allows results for survey respondents who were given different subsets of items (or assessment booklets) to be placed on a common scale. It also provides a basis for comparing subgroups within a country and for linking across surveys such as STEP and PIAAC.

Data Analysis and Scaling. Data from the STEP reading literacy assessment include pass/fail information for the Core assessment, the Reading Component score(s) and timing data, and information on the target population's reading literacy level, which is provided on the same five-level scale as PIAAC (Figure 8; also see Section IV for a more detailed description of each level). The countries that implemented the partial assessment have all these indicators, except the performance on the literacy scale.

Figure 8. Reading Literacy Performance / Scale from 1 - 5



The STEP reading literacy assessment is designed to assess cognitive skills based on the PIAAC literacy scale. Several steps were taken to assure comparability of the literacy scale in STEP to the PIAAC literacy scale in terms of instrumentation, target populations, and survey operations.

Before data could be used for analyses, the quality of the data had to be evaluated. This was done by reviewing the item responses to determine whether each respondent received the items and booklets as planned in the design (completion), reviewing item analyses (percent of correct responses per item) within and across countries to detect potential errors in translation or scoring, and reviewing scorer agreement to evaluate whether the scoring was accurate (reliability). Quality checks were also undertaken to evaluate the handling and patterns of missing values (that is, whether values were missing by design or omitted by the respondent).

The matrix design of the Exercise Booklets, described earlier, enables the project to reduce the response burden for an individual while the item pool can be expanded to represent the framework as completely as possible. However, the use of this design makes it

inappropriate to use any statistics based on the number of correct responses to describe or compare the skills of respondents.

This limitation can be overcome by using the IRT scaling approach, also described earlier. When a set of items requires a given skill, the response patterns should show regularities that can be modeled using the underlying commonalities among the items. These regularities can be used to characterize respondents by estimating so-called person or ability parameters through IRT models. They can also be used to characterize items in terms of a common scale, even if not all respondents answer identical sets of items; this too is done by estimating certain item parameters, such as item difficulty, through IRT models. In other words, if an item pool is used to measure a certain skill uni-dimensionally, that is, when only one skill is necessary to solve the items, respondents can be compared with each other even if they responded to different sets of items from this item pool (given that this item pool was scaled using a certain IRT model and showed appropriate model fit). This makes it possible to describe distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables.

In order to link STEP and PIAAC with a common scale, the appropriateness of using the item parameters estimated in the PIAAC 2012 main study was evaluated against STEP data for every item, by country. Using essentially the same IRT item parameters ensures that the scale linkage of STEP to PIAAC can be established and that inference structures will remain intact from PIAAC. To achieve this, the majority of item parameters in STEP should be the same as in PIAAC (common item parameters). Only a few items need unique item parameters in certain countries (these are newly estimated item parameters) in case they show no fit to the common item parameters obtained in PIAAC. Once item parameters were evaluated or established for each country, a latent regression model (population/latent regression model) was applied to an optimized set of background variables, separately for each country, to STEP item parameters, in order to produce plausible values of literacy proficiency within each country.

(j) Interviewer Impressions (Modules 8-11)

The survey collects detailed information about the interviewer's impressions on the conditions in which the interview took place. Upon completing the interview, interviewers go through two sets of questions: one set refers to the background questionnaire modules, and the second set of questions refers to the direct assessment of literacy. As mentioned above, these two parts of the survey require very different implementation processes. These questions relate to circumstances that could affect the way individuals answered the questions. For example, respondents may be reluctant to give personal information in the presence of neighbors or friends; they may be constantly interrupted during the interview to take care of their business, household chores, or young children. Some questions are specific to the literacy assessment: for example, it is important to know the conditions in which respondents were completing the assessment: whether they had table, chairs, and enough light. More generally, interviewers are asked whether they perceived respondents to be serious and truthful in their answers, and if they seemed to have difficulty understanding any question. The interviewer impression data can be very useful in checking the reliability of the data. In the case of the direct assessment of literacy, ETS will use the information collected in this module in their analysis of the data and their final measures of literacy.

III. EMPLOYER SURVEY

1. Overview

The Employer Survey is designed to help us better understand the demand for skills emanating from the private sector. Employers want productive workers, and in this context, the outcomes of interest to be generated from the survey are the employers' preferences for particular skills and workers' characteristics. Specifically, the survey seeks to assess the structure of the labor force; the skills currently being used; the skills that employers look for when hiring new workers; the propensity of firms to provide training; the link between skills and compensation and promotion; and relevant firms' background characteristics. The information generated may be used to establish and identify relevant traits (characteristics relatively stable over time), skills (capacity to perform a task), behavior (actions in response to stimulation) and beliefs.

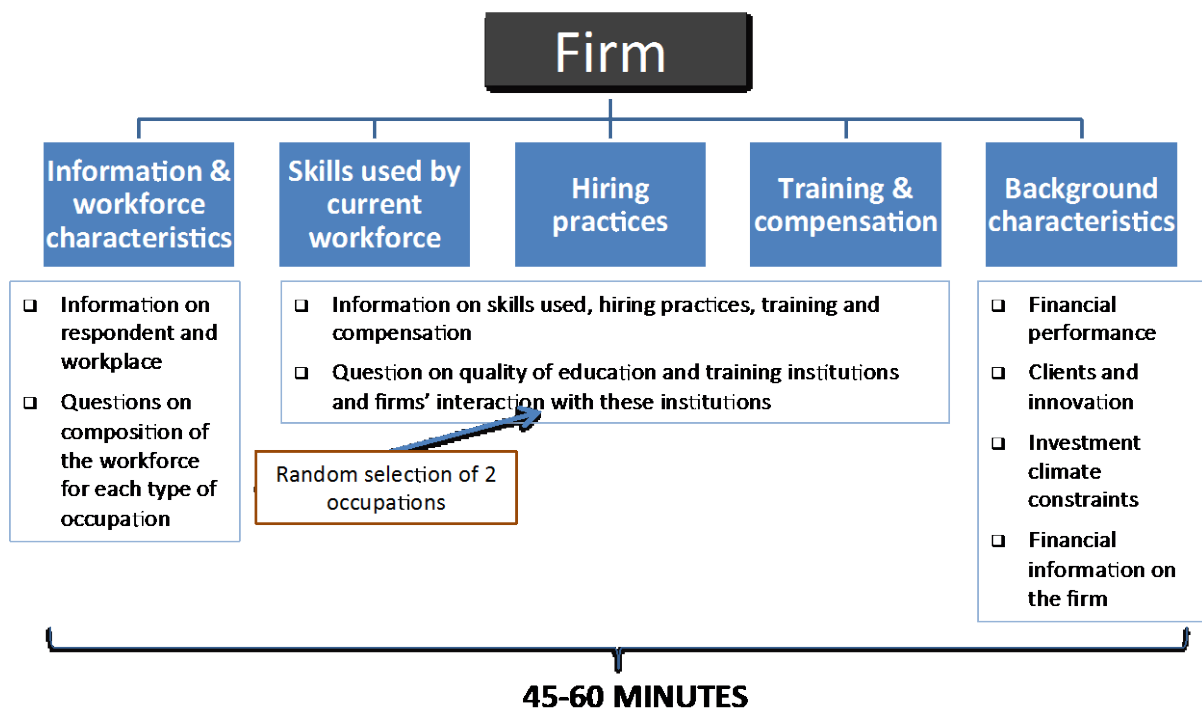
Two important challenges were taken into account when designing the employer survey: keeping non-response to a minimum, and including firms from the informal sector. To address the first challenge, the entire survey was conceived so it could be administered in less than 45 minutes. Employers are generally busy people and rarely have spare time to allocate for surveyors; this is why non-response rates in employer surveys are often quite high. The second challenge meant that the questionnaire had to be relevant for potentially very small firms.

The employer survey adopts several innovative approaches. First, information about skills usage, skills demand, and training and remuneration is gathered with respect to two randomly selected types of worker from among all the types of workers the firm employs. Second, the information on skills usage is derived from questions asked about the regular activities these reference workers engage in. Third, the instrument is designed to measure, as the household survey does, three main skills domains: cognitive skills, behavior and personality traits, and job-relevant skills. Fourth, questions on accounting and workforce data were simplified so employers who may have lower capacities could answer them.

In addition, the survey pays particular attention to the potential differences in skill levels among experienced workers and recent (new) hires. Specifically, the data include different sets of questions characterizing the skill levels of the workforce, distinguishing between these two groups. As in the case of the household survey, the information is organized so that for most of the sub-domains researchers will retrieve comparable information for both reference groups.

The employer survey covers organizational background characteristics (size, legal form, full-time vs. non-standard employment, industry, occupational breakdown), performance (revenues, wages and other costs, profits, scope of market), key labor market challenges and their ranking relative to other challenges (including satisfaction with education, training, and levels of specific skills), and job skill requirements, training, and recruitment issues associated with two randomly chosen occupations represented in the establishment. The final structure of the survey is shown in Figure 9.

Figure 9. STEP Employer Survey Structure



2. Description of the Questions Module by Module

The Employer Survey is divided into five modules. Two modules (1 and 5) collect background information on the firm; Module 2 asks about skills used; Module 3 inquires about recently hired workers; and Module 4 asks about training, compensations and promotions (see Table 2 for a detailed description).

Information and Workforce Characteristics (Module 1)

This module helps interviewers establish whether or not they are talking to a person knowledgeable about recruiting practices and skills development processes in the firm. It seeks information on the firm in order to determine the type of firm surveyed and in which economic sector it operates. The module then goes into detail about the workforce of the firm, in particular getting details on the share of females, the share of foreign workers, and changes in the workforce. All this information is gathered for each 1-digit occupation, as categorized by the International Standard Classification of Occupations (ISCO).¹¹

Skills Used by the Current Workforce (Module 2)

This module starts with the random selection of two occupations, which the questionnaire will then focus on. Hiring and compensation practices, as well as skills requirements, depend strongly on the occupation, so it is necessary to know what type of workers respondents are referring to when answering such questions. The questionnaire is organized in such a way that the respondent¹² is asked to select one occupation from a list of three alternatives (managers, professionals, and technicians), and another type of worker from a list of seven alternatives (clerical support workers, service workers, sales workers, skilled agricultural, craft and related trades workers, plant and machine operators, and elementary occupations).

¹¹ ISCO classifies occupations in the following groups: (i) managers, (ii) professionals, (iii) technicians and associate professionals, (iv) clerical support workers, (v) service workers, (vi) sales workers, (vii) skilled agricultural, forestry and fishery workers, (viii) construction, craft and related trades workers, (ix) plant and machine operators and assemblers and drivers, and (x) elementary occupations.

¹² Depending on the firm, the respondent could be the human resource manager, owner, president/vice-president/CEO, partner, director, general manager, financial officer, manager or other.

Table 2. STEP Employer Survey: Topical Content, by Module

(a) Basic Information and Workforce (Module 1)

- Job title of respondent, multi-site firm, year firm founded, establishment function and industry, legal status, owners, economic activity
- Number of employees (full-time, part-time, casual, with/without benefits, men/women) One-digit occupational employment breakdown (current, 1 year ago, 1 year from now, female, foreign)
- Information on hiring history in the previous 12 months per occupation

(b) Skill Used by the Current Workforce (Module 2)

Based on the selection of two occupations present in the firm's workforce, for each occupation type:

- Skills use (reading, writing, mathematics, problem solving, speaking a foreign language, making presentation, interacting with co-workers, computer use, punctuality)
- Compensation, promotion, education level, vocational degree holding

(c) Information on New Hires (Module 3)

Again for each type of occupation:

- Importance to the firm when deciding to retain a worker after probation period (i.e. rank),
- Personal characteristics (age, appearance, gender, networks)
- Job-related skills (reading, writing, mathematics, English, foreign language, technical skills, communication skills, leadership skills, team work skills, creative and critical thinking, problem solving, ability to work independently, time management)
- Personality traits (Conscientiousness, Emotional stability, Agreeableness, Extraversion, Openness to experience)
- Ranking of three groups of skills in order of importance
- Sources of new recruits, time to fill vacancies, number of offers per filled position, education of most recent hire, whether most recent hire holds a vocational degree, negotiability of salary at time of hiring, use of contractors to fill skills shortages

(d) Training and Compensation (Module 4)

- Contacts with education and training institutions (existence and purpose)
- Percentage of workers fully qualified for the job
- Training on premises (share of workers for different types of training, average number of days per type of training)
- Training outside workplace (share of workers per type of training, spending)
- Opinion of technical and vocational education system(s)
- Opinion of general education system
- Remuneration

(e) Firm Background (Module 5)

- Firm performance (last fiscal year, coming 3 years)
 - Main buyer, international business contacts, recent innovations
 - Opinion of labor market constraints (EPL, labor availability, education, training, experience, turnover, payroll taxes and benefit contributions, wage levels, minimum wage level)
 - Opinion of other investment-climate constraints compared with labor constraints (utilities, transportation, land, other taxes, customs and trade regulations, access to and cost of credit, uncertain government policy or macroeconomic conditions, corruption, crime/disorder)
 - Personnel function present (y/n)
 - Financial report and registration status
-

The procedure is the following: (i) the interviewer establishes which occupation types are present in the firm; (ii) the interviewer refers to a pre-given sticker on the front of the questionnaire, which lists the occupations under the two broad types, and selects the first occupation that appears on the list and is present in the firm. The stickers are generated by an Excel macro, which is provided by the STEP team to the survey firm. Each sticker lists the 10 occupation types in a random order.

The module goes on to ask a series of questions about specific skills that workers of these two types of occupation are currently using. The aim is to understand skills use within firms.

Skills Used by the Current Workforce (Module 2)

This module starts with the random selection of two occupations, which the questionnaire will then focus on. Hiring and compensation practices, as well as skills requirements, depend strongly on the occupation, so it is necessary to know what type of workers respondents are referring to when answering such questions. The questionnaire is organized in such a way that the respondent¹³ is asked to select one occupation from a list of three alternatives (managers, professionals, and technicians), and another type of worker from a list of seven alternatives (clerical support workers, service workers, sales workers, skilled agricultural, craft and related trades workers, plant and machine operators, and elementary occupations).

The procedure is the following: (i) the interviewer establishes which occupation types are present in the firm; (ii) the interviewer refers to a pre-given sticker on the front of the questionnaire, which lists the occupations under the two broad types, and selects the first occupation that appears on the list and is present in the firm. The stickers are generated by an Excel macro, which is provided by the STEP team to the survey firm. Each sticker lists the 10 occupation types in a random order.

¹³ Depending on the firm, the respondent could be the human resource manager, owner, president/vice-president/CEO, partner, director, general manager, financial officer, manager or other.

The module goes on to ask a series of questions about specific skills that workers of these two types of occupation are currently using. The aim is to understand skills use within firms.

Information on New Hires (Module 3)

In order to understand firms' skills needs, this module focuses on new hires in the two selected occupation types. It asks respondents to rank skills in order of importance when deciding to keep a worker beyond the probation period. Skills are grouped under three categories: (i) personal characteristics (age, appearance, gender, networks); (ii) cognitive and job-relevant skills (reading, writing, mathematics, English, foreign language, technical skills, communication skills, leadership skills, team work skills, creative and critical thinking, problem solving, ability to work independently, time management); and (iii) personality traits (Conscientiousness, Emotional stability, Agreeableness, Extraversion, Openness to experience). In order to ascertain the relative importance of these three types of skills, it then asks respondents to rank them. The module also asks for the education level of the most recent hire.

The skills that are listed here are directly comparable with those collected in the household survey. This enables researcher to assess the gap between the skills of the labor force and those sought by firms.

In addition to obtaining information on the skills that are valued by the firm, the module seeks to obtain information on its recruitment strategies and any difficulties it might be experiencing when hiring. Subsequent questions collect more information on how the firm recruits: the places from which firms recruit, the time and number of offers it takes to fill a vacancy, whether salary is negotiable at entry, and the remuneration offered to the most recent hire.

Table 3 presents a detailed breakdown of the different constructs included in the Employer Survey. It includes the number of specific questions associated with each of the sub-

domains as well as their associated scales (range of values) and the characteristics of the group of reference.

Table 3. Cognitive, Socio-emotional, and Job-relevant Skills

Type of Skill	Domain	Reference Group	Total Number of Questions	Range of Values
Cognitive	Reading	Experienced workers/new hires	2	[0,1], [1-5] or missing
	Writing	Experienced workers/new hires	2	[0,1], [1-5] or missing
	Numeracy	Experienced workers/new hires	2	[0,1], [1-5] or missing
Behavior and Personality Traits (Socio-emotional)	Big Five Inventory (Conscientiousness, Openness to Experience, Neuroticism, Agreeableness, and Extraversion)	Experienced workers/new hires	6	[0,1], [1-5] or missing
Job-relevant Skills (Skills at Work)	Interpersonal skills	Experienced workers/new hires	5	[0,1], [1-5] or missing
	Use of technology	Experienced workers	1	[0,1]
	Job-specific skills	Experienced workers/new hires	2	[0,1], [1-5] or missing
	Language	Experienced workers/new hires	2	[0,1], [1-5] or missing
	Autonomy	New hires	1	[0,1], [1-5] or missing
	Solving & learning	New hires	1	[0,1], [1-5] or missing

Training and Compensation (Module 4)

Training provides another avenue to improve staff's skills sets. This module collects information on training provided by the firm, both in the workplace and outside the workplace, for the two selected occupation types. It asks for the share of workers who benefited from such training programs, the number of days these involved, as well as the costs and type of training provided.

The module also contains questions to assess the overall quality of the formal education system (technical and vocational as well as general). It asks firms a series of questions aimed at determining whether they feel that the education system provides adequate skills and whether they have direct contacts with the education system and why. These questions are broad-reaching and no longer focus on the two occupations types selected earlier on.

Firm Background (Module 5)

Given the length of the survey, and the usual difficulties in collecting financial information from firms, the last module features two simplified questions about the firm's accounts. Questions were designed with small firms in mind and are phrased in a simple manner. It also asks subjective questions about the current financial performance of the company and prospects in the coming three years. Finally, the module includes detailed questions about the various ways in which labor market constraints could impair business, and questions that situate labor market constraints with respect to the other investment climate constraints.

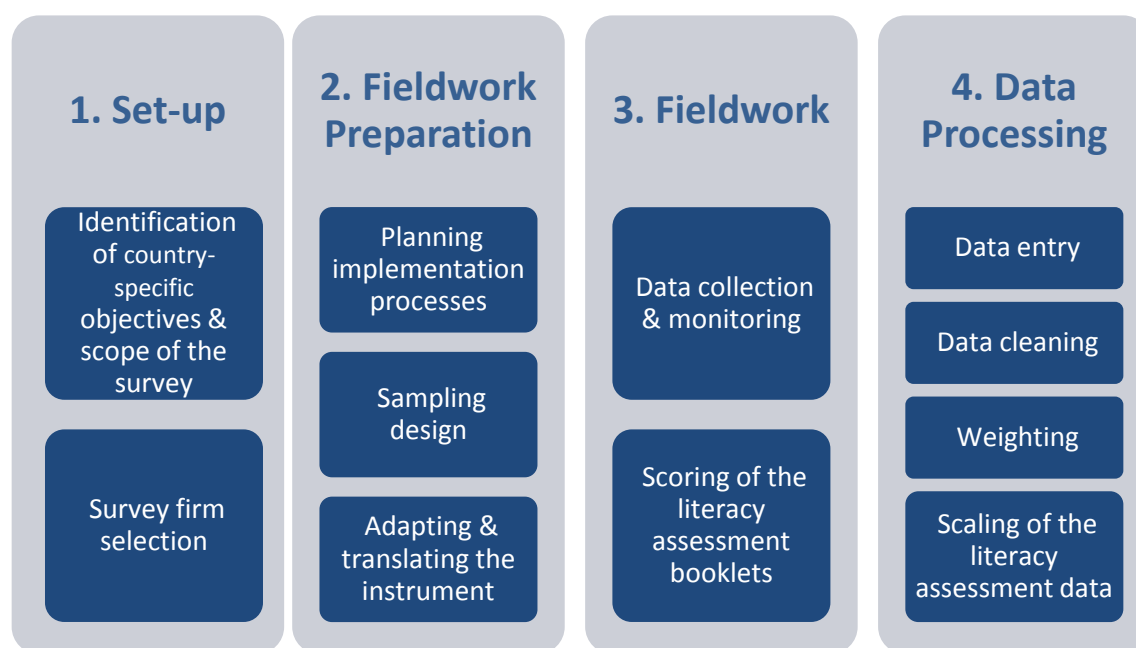
IV. STANDARDIZED IMPLEMENTATION

One of the main innovative features of the STEP project is that the same surveys are undertaken following the same technical standards in all countries. In order to ensure a homogenous implementation in all countries, the choice was made to centralize the coordination and supervision of the country survey firms.

1. Implementation of the STEP Household Survey

The implementation of the direct reading literacy assessment required particular care to ensure proper administration as well as data quality and comparability. Compliance with the STEP technical standards was essential at each stage of the survey implementation. Figure 10 describes the main stages of the process.

Figure 10. STEP Household Survey: Implementation Stages



Coordination and supervision were centralized so that survey instruments could be administered in a standardized and fully consistent way across all participating countries. All survey firms benefited from the STEP team's technical assistance throughout the implementation process. The STEP team and ETS also held several group training sessions to prepare survey firms to implement the STEP household survey. The STEP team provided the survey firms with all training materials required to train their staff.

Set-up

Target Population

The STEP target population is the urban population ages 15 to 64. However, countries may broaden the scope of the survey or oversample particular subgroups to serve country-specific objectives. The surveys in Lao PDR and Sri Lanka, for instance, included rural areas as well as urban centers. In Vietnam and Colombia, the target population focused on major urban centers (see country weighting documentation).

Country-specific Questions

Collecting strictly comparable data required using the STEP household questionnaire. However, countries were also offered the possibility of including up to five questions to gather more detailed information on a specific issue. In Ukraine, for example, the STEP and Ukraine Living Standard Measurement surveys were intertwined, and great care was taken to ensure that all STEP technical requirements were complied with. Other countries added specific questions aimed at better understanding the educational system.

Fieldwork Preparation

Implementation Planning

Each survey firm prepared, in close consultation with the STEP team, a National Survey Design Planning Report (NSDPR) describing its implementation plan. Each NSDPR was closely reviewed by the STEP team for compliance with STEP standards. The STEP technical

standards required survey firms to comply with specific norms regarding key areas such as project management, fieldwork team composition, data collection and fieldwork monitoring processes, scoring of the reading literacy booklets, and data entry.¹⁴

Sampling Design

The sampling strategy was designed to ensure that the target population represents at least 95 percent of the urban working-age population (ages 15 to 64) in each country. To allow comparability of the data collected with other country surveys and to account for country contexts, the STEP surveys used each country's official definition of "urban." This was also essential to the quality of the sample frames. To ensure consistency of the sampling strategies across all countries, all survey firms designed their sampling strategies in close cooperation with the STEP survey methodologist, who approved all sampling plans and drew the sample files used in each country.

Adaptation and Translation of the Survey Instruments

Great attention was provided during the adaptation and translation stages to ensure data quality and comparability. The household questionnaire and reading literacy assessment were translated separately by two independent translators before a third translator reconciled, and documented, any discrepancies. The STEP team and ETS checked the translations and worked closely with the survey firms to finalize the instruments. In English-speaking countries, the instruments were adapted to reflect local idioms.¹⁵

¹⁴ For more detailed information, see country National Survey Design Planning Reports and STEP Technical Standards in the STEP Skills Measurement Program Collection on the World Bank's Microdata Catalog (<http://microdata.worldbank.org/index.php/catalog/step/about>).

¹⁵ In addition, in Kenya the section of the questionnaire assessing behavior and personality traits (Module 6) was translated into Swahili to adapt to respondents' language preferences, so that the respondent could choose to answer in either English or Swahili.

Data Collection

Data Collection and Monitoring

The household questionnaire was administered through paper and pencil in all countries except Colombia and Kenya, where computer-assisted personal interviews were carried out. The reading literacy assessment was systematically administered through paper and pencil. Survey firms submitted regular fieldwork reports to the STEP team, which monitored data collection progress as well as non-response rates and provided guidance to the teams whenever required. In most countries, fieldwork spanned three to four months.

Scoring of the Reading Literacy Assessment Booklets

In countries implementing the full reading literacy assessment, that is, both the General Booklet and the Exercise Booklets, interviewers scored the Core assessment during the interview to determine whether or not a respondent should take the Exercise Booklet items. Close supervision was required at this stage, and overall very few errors were made. In all countries, survey firms scored the reading literacy assessment following STEP guidelines. A workshop was organized by ETS to ensure the survey firms would be able to score the booklets adequately.

Data Processing

Data Entry

All survey firms were requested to enter the data through a double-data entry process. For the household questionnaire, the survey firms could either use the STEP data entry program or develop their own. If they developed their own program, the STEP team tested it as part of its quality assurance process. For the reading literacy assessment data, survey firms were all requested to use the data entry program provided by ETS.

Data Cleaning

The STEP team checked the household questionnaire data submitted by the survey firms to identify data entry errors and compliance with fieldwork procedures (such as random selection of the individual respondent). ETS checked the reading literacy assessment data. The process consisted of reviewing the item responses to determine whether each respondent received the items and booklets as planned in the design (completion), reviewing item analyses (percent of correct responses per item) within and across countries to detect potential errors in translation or scoring, and reviewing scorer agreement to evaluate whether the scoring is accurate (reliability). Quality checks were also carried out to evaluate the handling and pattern of the missing values (that is, whether missing by design or omitted by the respondent).

Weighting

The data weighting was undertaken by the STEP survey methodologist to ensure consistency across sampling strategies. Whenever recent population counts were available, the weights were adjusted against benchmark variables (such as gender and age). A weighting documentation was produced for each country and provides country-specific information on the sample design and weighting process.

Scaling the Reading Literacy Data

Once the data had been cleaned and weighted, ETS undertook the Item Response Theory scaling of the reading literacy data to provide the estimation of item parameters and the proficiency distribution of the population. The latter was then used to calculate *a posteriori* distribution together with the household questionnaire variables, using latent regressions. From this distribution, plausible values (which are multiple imputations) were obtained to provide a more accurate and reliable proficiency estimation than the proficiency estimation of the Item Response Theory scaling alone. Similar to the approach used by PIAAC, STEP used the two-parameter logistic model for dichotomously scored responses.

STEP and PIAAC Comparability

Several steps were taken to ensure comparability of the literacy scale in STEP with the PIAAC literacy scale in terms of instrumentation, target populations, and survey operations. Items selected for STEP belong to the PIAAC literacy framework: they were either (i) previously administered through the PIAAC paper-based assessment or were adapted from the PIAAC computer-based instruments or (ii) had been administered in other large-scale adult literacy assessments that had been previously linked to the PIAAC literacy scale. The characteristics of the target population for STEP were a subset of the adult population, ages 16-65, included in the total population of PIAAC national samples. Both PIAAC and STEP are assessed by an interviewer face-to-face at home or at a place most convenient for the respondent. The systems of test administration, scoring, and the evaluation of scoring accuracies employed for STEP were comparable to those used in the paper-based PIAAC assessment. The analysis, methods, and procedures for STEP were based on identical psychometric principles used for PIAAC.

Sample Size and Response Rates

Sample sizes varied from country to country, from 2,989 observations in Sri Lanka to 4,009 observations in Macedonia. Response rates generally ranged from 60 percent in Sri Lanka to 98 percent in the Yunnan Province. (In Bolivia and Colombia, however, response rates were markedly lower; 43 percent and 46 percent, respectively). Sample sizes were determined based on the scope of the survey and literacy rates to ensure that a sufficient number of reading literacy booklets would be completed. Countries in which the survey was to be administered in two languages were also required to increase their sample size in consequence (Table 4).

Table 4. STEP Household Surveys: Sample Sizes and Response Rates

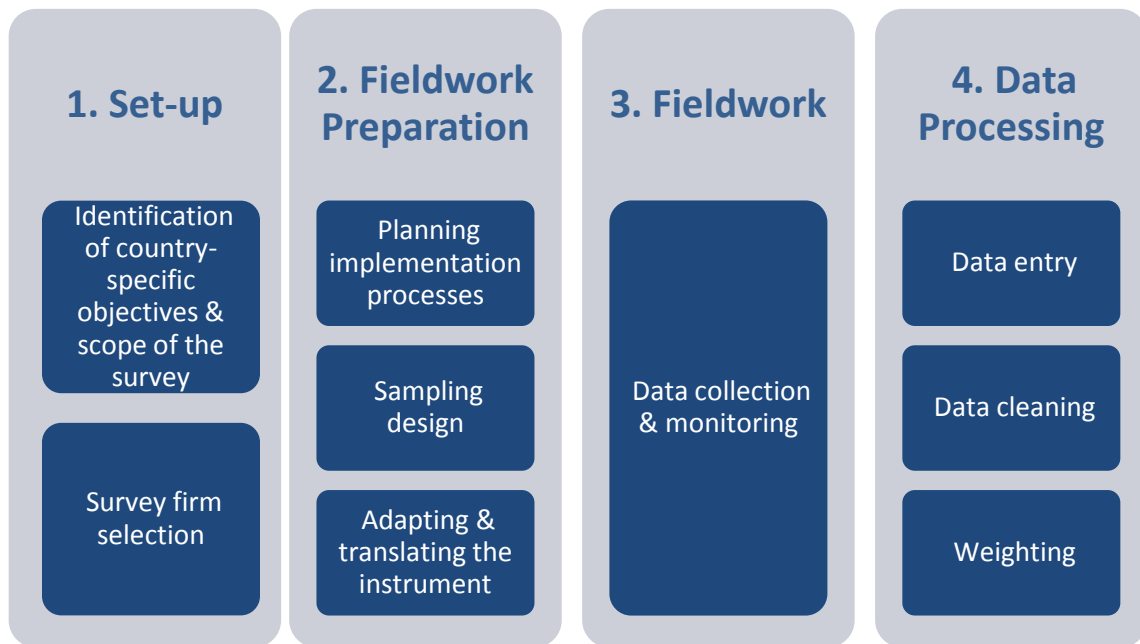
	Armenia (urban)	Bolivia (urban)	Colombia (urban)	Georgia (urban)	Ghana (urban)	Kenya (urban)
Sample Size	2,992	2,435	2,617	2,996	2,987	3,894
Response Rate	50%	43%	48%	63%	83%	92%

	Lao PDR (urban and rural)	Macedonia (urban)	Sri Lanka (urban and rural)	Ukraine (urban)	Vietnam (urban)	Yunnan Province (urban)
Sample Size	2,845	4,009	2,989	2,389	3,405	2,017
Response Rate	95%	67%	63%	61%	62%	98%

2. Implementation of the STEP Employer Survey

The STEP employer survey was developed by the STEP team, which also provided technical assistance to implementing agencies whenever requested. Technical standards were set to ensure proper administering of the survey instrument and data quality. The STEP team also held a training session to prepare survey firms to implement the survey. All training materials were provided to the survey firms so they could train their staff.

Figure 11. STEP Employer Survey: Implementation Stages



Set-up

Scope and Target Population

Participating countries were offered some flexibility regarding the scope of the employer survey. In some countries, specific economic sectors were selected, whereas others sampled a wider range of sectors. The target unit of observation was the workplace so that the information gathered on workers' actual skills and potential mismatches or gaps could be as precise as possible.

Country-specific Questions

Collecting strictly comparable data required using the STEP employer questionnaire. However, countries were offered the possibility of adding questions to gather more detailed information on a particular issue, such as youth employment.

Fieldwork Preparation

Implementation Planning

Survey firms assisted by the STEP team prepared an Employer Survey Design Planning Report (ESDPR) describing their implementation plan. Each ESDPR was closely reviewed for compliance with STEP standards, which required survey firms to comply with specific norms regarding key areas from project management to data entry.

Sampling Design

An issue of primary concern for the survey was always the sampling strategy. Since the overarching objective of the survey was to quantify the demand for skills by employers, capturing a representative sample of employers was essential. Four sampling strategies were discussed: drawing from firm register data, drawing from establishment census data, building a sampling frame from responses in the household survey and drawing from those employers, and door-to-door sampling. Each approach has advantages and disadvantages, as seen in Table 5. The preferred sampling strategy is to use either the door-to-door method or the household survey data as a sample frame, since these approaches include a broad range of employers.¹⁶

¹⁶ So far, the STEP employer surveys have used firm registries because of time constraints or because they focused on specific economic sectors.

Table 5. STEP Employer Survey Sampling Options

Sampling Approach	Advantages	Disadvantages
Register Data	<ul style="list-style-type: none"> • Easy to find firms • Easy to establish a representative sample of registered firms • Less expensive 	<ul style="list-style-type: none"> • Can miss a large share of firms and employment in countries with large informal sectors • Leads to biased results (for the full economy)
Establishment Census Data	<ul style="list-style-type: none"> • Covers formal- and informal-sector establishments • Provides a complete sample frame • Less expensive 	<ul style="list-style-type: none"> • Undertaken infrequently <ul style="list-style-type: none"> ◦ Samples successful establishments disproportionately ◦ Can lead to many bad addresses • Difficult to administer well
Door-to-Door	<ul style="list-style-type: none"> • Captures both formal and informal sector • Captures new establishments 	<ul style="list-style-type: none"> • Expensive • Time-consuming
Sample Frame from Household-based Survey	<ul style="list-style-type: none"> • Able to match workers to their employers • Captures formal- and informal-sector employers in proportion to the share of employment they represent 	<ul style="list-style-type: none"> • Respondents may be hesitant to answer or may provide inaccurate information • Sample is representative of the stock of employment, not stock of firms (may be an advantage) • Cannot be launched until household survey is completed (cannot be run in parallel)

Using the sample frame from the household survey presents two main advantages, as it provides (i) a direct match with the household survey and (ii) a picture reflecting actual labor market realities, by including both formal and informal workplaces. So far, due in great part to time constraints, all participating countries have opted for the approach using a firm registry. This sampling strategy was also chosen because country teams wished to focus the analysis on the needs of fast-growing or innovative sectors, which were expected to drive future growth. On a more technical note, since firm registries usually provide more information on firm characteristics, stratification and weighting processes were eased, which helped to ensure data quality.

Adaptation and Translation of the Survey Instruments

The employer questionnaire was adapted and translated in close collaboration with the STEP team. It was finalized after being pre-tested in the field by senior members of the survey firm in each country.

Data Collection

Data Collection and Monitoring

The employer questionnaire was administered through paper and pencil. Survey firms submitted regular fieldwork reports to the STEP team, which monitored data collection progress as well as non-response rates and provided guidance to the teams whenever required. In most countries fieldwork spanned four to five months.

Data Processing

Data Entry

All survey firms were requested to enter the data through a double-data entry process. Survey firms could either use the STEP data entry program or develop their own. In these instances, the STEP team tested the program as part of its quality assurance process.

Data Cleaning

The STEP team checked the employer questionnaire data submitted by the survey firms to identify data entry errors and compliance with fieldwork procedures (such as random selection of the worker types).

Weighting

The data weighting was undertaken either by the STEP survey methodologist or by the survey firms themselves.¹⁷ A weighting documentation was produced for each country

¹⁷ The STEP survey methodologist weighted the employer survey data for Armenia, Azerbaijan, and Georgia.

weighted by the STEP team. It provides country-specific information on the sample design and weighting process.

Sample Size and Response Rates

Sample sizes for the STEP employer survey vary from about 300 to 500 workplaces. As is usual for such surveys, response rates are low. Employers lack time but are also wary of providing potentially sensitive information about their business to outsiders. (Table 6.)

Table 6. STEP Employer Survey: Sample Sizes and Response Rates for Selected Surveys

	Armenia	Azerbaijan	Georgia
Sample Size	384	314	354
Response Rate	42%	38%	51%

V. USING THE STEP DATA IN ANALYSES

1. Skills Aggregation Methodology

Aggregation Principles

As described above, in order to have accurate measures of the three types of skills, the STEP Skills Surveys measure most of the cognitive, behavior, personality trait, and job-relevant sub-domains through multiple items. The first step of any analysis of these data is to combine this information in order to form an indicator for each of the sub-domains. This section describes the general principles that should be the basis for aggregating the measures into meaningful skills sub-domains.

The first principle is that researchers should not innovate when the sub-domain has been empirically validated. The set of sub-domains included in the STEP Skills Surveys are the result of a long and detailed analysis of the best short-instrument measuring the prototypical components of cognitive, behavior, personality trait, and job-relevant skill measures available. In this context, research should follow the literature and form the sub-domains accordingly. Consider, for example, “Extraversion,” one of the components of the Big Five personality scale (Table 7).

Table 7. How to Generate the Sub-domain “Extraversion”

How do you see yourself?	Possible answers			
	Almost always	Most of the time	Some of the time	Almost never
Are you talkative?	1	2	3	4
Do you like to keep your opinions to yourself? Do you prefer to keep quiet when you have an opinion?	1	2	3	4
Are you outgoing and sociable; for example, do you make friends very easily?	1	2	3	4

Notice that the second question measures the opposite of “Extraversion,” and consequently the score must be reversed. Researchers must modify the score before using this particular item. They must recode it so that “4” is associated with “Almost always”, “3” with “Most of the time”, “2” with “Some of the time” and “1” with “Almost never.” After doing this, the individual score is obtained as the simple average across the different items. Note that a low (high) score should be interpreted as a high (low) level of “Extraversion.” In most of the cases, researchers can follow a similar strategy to construct the sub-domains from the respective items. When constructing sub-domains, researchers should use only non-missing items. This means that for all skills sub-domains that are based on the aggregation of multiple variables, if one of these variables is missing, then the final aggregate variable should be coded as a missing value. For the purpose of aggregating, all “Don’t know” answers should be recoded as missing values.

The second principle is that simple scales should be preferred. For sub-domains without a well-established scoring scale, researchers should prefer simple and intuitive scales. Most of the skill measures collected under the STEP surveys can be scored using simple algorithms (simple averages across questions will work in most of the cases).

Identifying Relevant Sub-domains of Skills

Rather than simply aggregating the sub-domains into a single measure, researchers might want to identify a limited set of relevant sub-domains from the full battery based on a particular (scientific) criterion. Here we can identify three different empirical strategies, as follows.

Selection Based on Association

The results obtained from the set of sub-domains might be interpreted as direct manifestations of individual's true skills. For example, one could interpret the results from self-reported literacy and numeracy as proxies of the underlying individual's true cognitive skill. In this context, sub-domains can be interpreted as (noisy) measures of an underlying domain, and their correlations might inform us about its relative importance. This intuition has motivated an entire literature which, based on statistical methods, has identified underlying skills from batteries of tests (sub-domains). Formally, statistical methods such as principal components (a descriptive statistical technique) and factor analysis (based on regression modeling techniques to test hypothesis producing error terms) have been used to extract what is usually interpreted as the underlying skill explaining an individual's performance/answer in each of the sub-domains. These techniques are commonly applied in the case of cognitive skills (Carroll, 1993; Jensen, 1998; Heckman, Stixrud, and Urzua, 2006; Diaz et. al., 2011) as well as behavior and personality traits (Digman, 1990; Heckman, Stixrud, and Urzua, 2006).

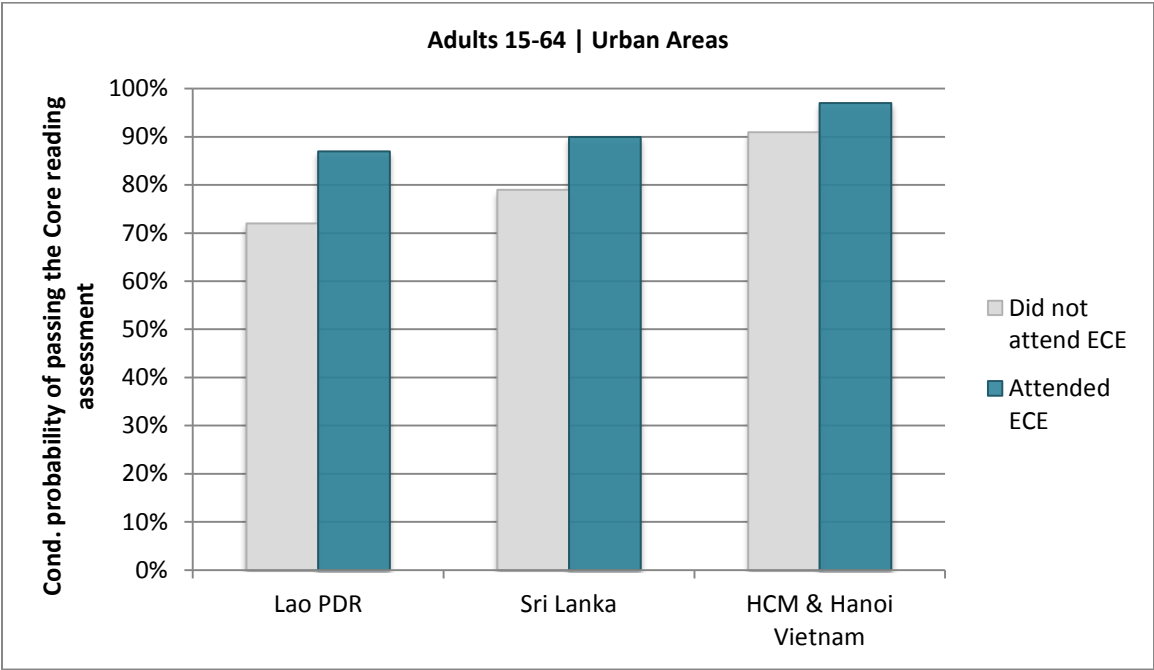
Selection Based on Malleability

A second possible approach to identify relevant sub-domains could emerge from the empirical capacity to determine to what extent they can be explained by individual-specific variables. Consider, for example, a set of different cognitive sub-domains, and suppose researchers are interested in understanding what variables explain them. This could be motivated by the need to implement public policies specifically designed to improve the cognitive levels of the population. Let X denote the set of candidate variables explaining the cognitive sub-domains ($Skill(1,c), \dots, Skill(N,c)$). Therefore, the researcher could select the cognitive sub-domain for which the vector X has the maximum explanatory power. This approach could be easily implemented in the context of linear regression models for the conditional mean of each $Skill(c)$ as a function of X .

A better understanding of the variables explaining specific sub-domains could help to identify critical dimensions over which public policies could intervene.

The example illustrated in Figure 12 concerns the link between early childhood education and reading proficiency in selected countries. The data show that adults who participated in an early childhood education program were more likely than those who did not to pass the Core Reading Literacy Assessment in Lao PDR, Sri Lanka, and Vietnam, controlling for age, gender and mother’s educational attainment.

Figure 12. Early Childhood Education and Reading Literacy in Urban Lao, Sri Lanka, and Vietnam



Notes: Conditional probability controlling for age, gender, and mother’s education. Results are statistically significant. ECE stands for early childhood education.

Selection Based on Predictability

Ultimately, researchers might want to identify those sub-domains that have the maximum predictive power for outcomes. In other words, the association between skills and labor, education or health outcomes might motivate the selection of the relevant sub-domains. When considering cognitive skills, behavior and personality traits, or job-relevant skills, researchers might prefer sub-domains that better explain earnings. In summary, researchers could select those sub-domains that have the highest correlation with the outcome of interest Y. A similar strategy could be developed for the set of behavioral and job-specific sub-domains.

Translating Complex Scoring Scales into Interpretable Objects

Scores obtained for the different questions measuring individuals' skills can be transformed into interpretable objects. This not only facilitates the analysis but also secures the intuitive communication of the main results to a general audience. Depending on the characteristics of the specific domain, researchers can develop the following two strategies to construct easily interpretable scales.

For Binary Sub-domains

In some cases, the scoring scales are based on binary answers. Consider for example question 2.05 in the STEP employer survey: "Does their job regularly involve reading?" The associated answers are "YES" or "NO". In this case, researchers can simply label "YES" as "the individual possesses the skill" and "NO" as "the individual does not possess the skill."

For Continuous or Multi-valued Scoring Scales

In general, sub-domains in the STEP surveys will be measured using continuous or multi-valued scoring scales. For a fraction of them, the scoring scale will be simple, allowing an intuitive interpretation. This is the case with many behavioral and personality trait

measures that have already been used and validated in previous studies. In such cases, researchers will not need to modify the original scales.

Good examples are the scoring scales associated with the individual items forming the Big Five personality scale. In other cases, the scoring scale might be complex (multiple values without a clear interpretation). The complexity of the scoring scales should not discourage researchers from developing more interpretable alternatives. Consider, for example, the following case.

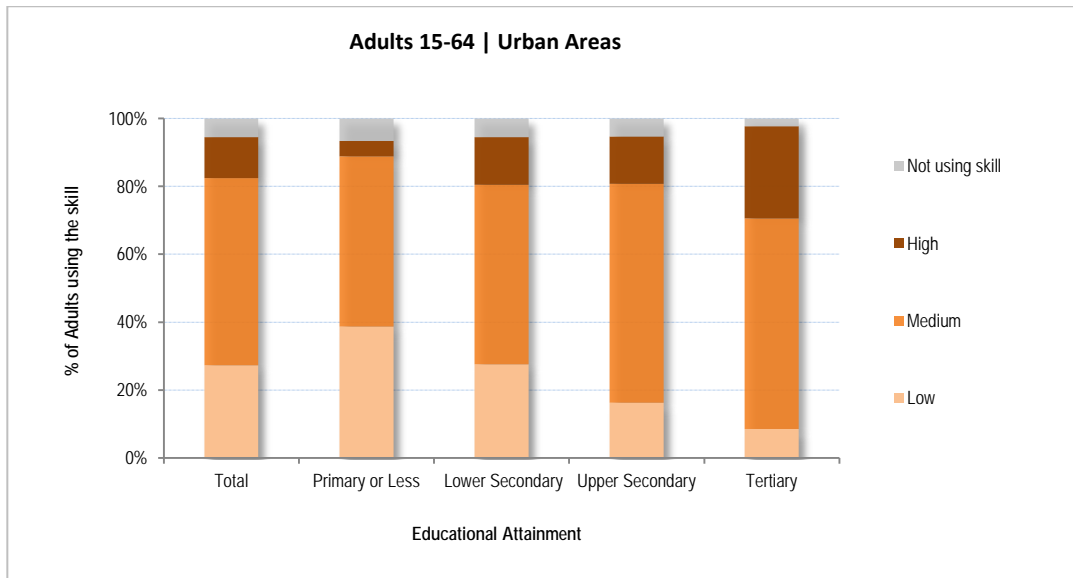
Let $T(j)$ be the results associated with the sub-domain j . Furthermore, assume that the scoring scale is complex, and researchers cannot easily determine the specific value defining a “low,” “medium,” or “high” use or proficiency on the sub-domain. With individual-level data on $T(j)$, researchers could generate the following ranking:

$$R(j) = \begin{cases} 1 & \text{if Low Use} \\ 2 & \text{if Medium Use} \\ 3 & \text{if High Use} \end{cases}$$

where $R(j) = 1$ for those individuals with scores in the bottom 25 percent of the distribution, $R(j) = 2$ for individuals with scores between 25 and 75 percent of the distribution, and $R(j) = 3$ for those with scores in the top 25 percent of the distribution.

For example, using a three-valued scale for literacy, researchers could generate a simple figure depicting the association between literacy levels and education (see Figure 13).

Figure 13. Reading Skills Use and Educational Attainment in Urban Ghana



Aggregating Self-reported Cognitive and Job-relevant Skills

Self-reported Cognitive Skills

The STEP survey asks respondents to report on their use of cognitive skills in daily life and at work, namely if they read, write, or use mathematics. For each skill, a score ranging from 0 to 3 was computed. When a respondent reports not using a given skill, the score is set at 0. For respondents who do use a given skill, intensity or complexity of use is defined based on the criteria presented in Table 8.

Table 8. Self-reported Cognitive Skills

Use of Reading and Writing Skills		Intensity of Use	Level
Does not read/write	=	Does not use	0
Reads/writes documents of 5 pages or less	=	Low	1
Reads/writes documents of 6 to 25 pages	=	Medium	2
Reads/writes documents of more than 25 pages	=	High	3
Use of Numeracy Skills		Complexity of Use	Level
Does no math	=	Does not use	0
Measures or estimates sizes, weights, distances; calculates prices or costs; performs any other multiplication or division	=	Low	1
Uses or calculates fractions, decimals or percentages	=	Medium	2
Uses more advanced math such as algebra, geometry, trigonometry	=	High	3

Job-relevant Skills

Job-relevant skills are task-related and build on a combination of cognitive and socio-emotional skills. The STEP survey asks respondents about their use of such skills on the job, including, among others computer use, repair and maintenance of electronic equipment, operation of heavy machinery, client contact, supervision. For each skill, a score ranging from 0 to 3 was computed. When a respondent reports not using a given skill, the score is set at 0. For respondents who do use a given skill, intensity or complexity of use is defined based on the criteria presented in Table 9.

Table 9. Selected Job-relevant Skills

Computer Use	Intensity of Use	Level
“As a part of your work do you use a computer?”		
“As a part of your life [outside work] have you used a computer in the past 3 months?”		
Does not use a computer/ almost never uses a computer	Does not use	0
Uses computer less than three times per week	Low	1
Uses computer three times or more per week	Medium	2
Uses computer every day	High	3

(Table 9. Cont'd)

Contact with Clients	Intensity of Use	Level
“As part of this work, do you have any contact with people other than co-workers, for example customers, clients, students, or the public?” *		
Does not have any contact with clients	= Does not use	0
Involvement scale ranges from 1 to 4	= Low	1
Involvement scale ranges from 5 to 7	= Medium	2
Involvement scale ranges from 8 to 10	= High	3

* Scale ranges from 1 to 10, where 1 is little involvement and 10 means much of the work involves meeting or interacting with people other than co-workers

Solving and Learning at Work	Intensity of Use	Level
Average of 2 items	Item 1. “Some tasks are pretty easy and can be done right away or after getting a little help from others. Other tasks require more thinking to figure out how they should be done. As part of this work, how often do you have to undertake tasks that require at least 30 minutes of thinking?”	
	Never	= Does not use 0
	Less than once per month	= Low 1
	Less than once a week but at least once a month	= Medium 2
	OR at least once a week but not every day	
	Every day	= High 3
	Item 2. “How often does (did) this work involve learning new things?”	
	Rarely	= Does not use 0
	At least every 2-3 months or at least once a month	= Low 1
	At least once a week	= Medium 2
	Every day	= High 3

(Table 9, cont'd.)

Autonomy and Repetitiveness		Intensity of Use	Level
Average of 2 items	Item 1. “Still thinking of your work, how much freedom do you have to decide how to do your work in your own way, rather than following a fixed procedure or a supervisor's instructions? Use any number from 1 to 10 where 1 is no freedom and 10 is complete freedom.”		
	Decision freedom scale from 1 to 2	= Close to none	0
	Decision freedom scale from 3 to 6	= Low	1
	Decision freedom scale from 7 to 9	= Medium	2
	Decision freedom scale 10	= High	3
	Item 2. “How often does (did) this work involve carrying out short, repetitive tasks?”		
	Almost all the time	= Close to none	3
	More than half the time	= Low	2
	Less than half the time	= Medium	1
	Almost never	= High	0

Aggregating Behavioral and Personality Trait Measures

As discussed in Section II, the STEP survey builds on the Big Five personality traits: conscientiousness, openness, neuroticism (or its opposite: emotional stability), agreeableness, and extraversion. Measures of grit, which has been shown to have an impact in life outcomes, and of hostile attribution bias were also included, as well as questions pertaining to how individuals make important decisions. Response categories ranged from 1, “almost never,” to 4, “almost always.”

The aggregation process, also discussed in Section II, was based on a simple average across items. Other aggregation methods may be explored. Negatively scored items were recoded prior to the aggregation. Table 10 indicates which items from Module G are to be mapped to each behavior or personality trait.

Table 10. Behavioral and Personality Trait Measures

Behavior & Personality Trait	Question in Module G	Items
Openness	Q.1.03	Do you come up with ideas other people haven't thought of before?
	Q.1.11	Are you very interested in learning new things?
	Q.1.14	Do you enjoy beautiful things, like nature, art and music?
Conscientiousness	Q.1.02	When doing a task, are you very careful?
	Q.1.12	Do you prefer relaxation more than hard work?
	Q.1.17	Do you work very well and quickly?
Extraversion	Q.1.01	Are you talkative?
	Q.1.04 *	Do you like to keep your opinions to yourself? Do you prefer to keep quiet when you have an opinion? *
	Q.1.20	Are you outgoing and sociable, for example, do you make friends very easily?
Agreeableness	Q.1.09	Do you forgive other people easily?
	Q.1.16	Are you very polite to other people?
	Q.1.19	Are you generous to other people with your time or money?
Emotional Stability (Neuroticism) *	Q.1.05 *	Are you relaxed during stressful situations? *
	Q.1.10	Do you tend to worry?
	Q.1.18	Do you get nervous easily?
Grit	Q.1.06	Do you finish whatever you begin?
	Q.1.08	Do you work very hard? For example, do you keep working when others stop to take a break?
	Q.1.13	Do you enjoy working on things that take a very long time (at least several months) to complete?
Hostile Attribution Bias	Q.1.07	Do people take advantage of you?
	Q.1.22	Are people mean/not nice to you?
Decision-making	Q.1.15	Do you think about how the things you do will affect you in the future?
	Q.1.21	Do you think carefully before you make an important decision?
	Q.1.23	Do you ask for help when you don't understand something?
	Q.1.24	Do you think about how the things you will do will affect others?

*Note: In the Wave 2 household questionnaire, two additional questions were added: Q.1.25, "Do you like to share your thoughts and opinions with other people, even if you don't know them very well?" can be used instead of Q.1.04; and Q.1.26, "Do you get very upset in stressful situations?" can be used instead of Q.1.05.

In the Wave 2 version of the household questionnaire, two additional questions were added to enhance the measurement of Extraversion (question 1.25) and Emotional Stability (question 1.26). This was done because results from the first wave of countries suggested that reverse items might slightly decrease the performance of the measures, especially in the context of the reduced number of items used in STEP. Users can test the two different ways of aggregating by (i) using the initial 24 items (Q.1.01 to Q.1.24), or (ii) replacing Q.1.04 and Q.1.05 with Q.1.25 and Q.1.26, respectively. For comparisons with Wave 1 countries, the initial 24 items should of course be used.

2. Direct Reading Literacy Assessment Data

Reading Components Results: Accuracy and Rate

The reading components measures are designed to provide information about how adults with low levels of reading proficiency fare with respect to selected building blocks of literacy proficiency. These building blocks are the following:

- *Word Meaning (Print Vocabulary)* measures the extent to which participants can recognize the printed forms of common objects.
- *Sentence Processing* measures the extent to which participants can comprehend sentences of varying levels of complexity.
- *Basic Passage Comprehension* measures the extent to which participants can comprehend the literal meaning of connected text.

Results for each of the three reading components can be interpreted in terms of accuracy (how many items were answered correctly) and rate (how quickly the tasks were completed, whether the answers were correct or incorrect). Variables such as the number of correct answers and the time taken to answer correctly in each of these three sub-domains can be of use in analyses.

Core Assessment

The Core assessment was designed to identify respondents to whom the Exercise Booklets could be administered. In countries in which a significant proportion of the sampled population is at the lower end of the reading proficiency distribution, a variable indicating whether a respondent passed or failed the Core can be used to identify individuals who have very low literacy levels.

Exercise Booklets: the Literacy Proficiency Scale & Proficiency Levels

To adequately measure the skills of adults with differing educational backgrounds and life experiences, STEP included tasks ranging from very easy to very challenging. Results from the literacy assessment were reported along a proficiency scale ranging from 0 to 500 with tasks at the lower end of the scale being easier than those at the higher end. The scaling analysis for STEP allows us to place the STEP literacy items on the PIAAC literacy scale. This means that the STEP scale scores have the same range (0–500) and the description of the underlying skills along the scale are the same as for PIAAC.

Defining the Proficiency Levels

The scores by themselves are of limited interest: reporting that one task gets a score of 215 on a scale while another falls at 345 provides some information – namely that the first task is easier than the second – but it does not tell us much about the underlying skills and knowledge each task requires. To provide a richer report, “described proficiency scales” for each of the domains were defined. These described proficiency scales explain what performance means at various points along those scales. To create these described proficiency scales, the expert groups in each domain met with psychometricians and test developers to review the data, look at the tasks as they were distributed along the 500-point scales, and articulate how the requisite skills and knowledge to complete those tasks progressively increased along the scale.

The purpose of described proficiency scales is to facilitate the interpretation of the scores assigned to respondents. That is, respondents at a particular level not only demonstrate knowledge and skills associated with that level but also the proficiencies required at lower levels. Thus, respondents scoring at Level 2 are also proficient at Level 1, with all respondents expected to answer at least half of the items at that level correctly.

Each of the six literacy scale proficiency levels is defined in Table 11, where one or more representative tasks are described to illustrate the key information-processing skills at each level.

Using Plausible Values

Because the assessment would be too long if every individual had to take the entire battery of items, the Exercise Booklets' assessment design divides the test into partially linked booklets. As mentioned earlier, this method is common in large-scale assessments. This reduces the probability of some external factors interfering with the assessment (for instance, time of interview, interviewee burden, and interruptions, among others). However, these advantages come at the cost of a loss in individual-level accuracy, since only a portion of the entire battery of items is administered to a single individual.

As was explained above, the Exercise Booklets' assessment design divides the test into partially linked booklets. Reporting the reading literacy score for the Exercise Booklets is thus based on multiple plausible values. Because of the relatively small number of items taken by each respondent, the accuracy of measurement for a single respondent is considerably lower than is common with assessments designed for individual reporting. To address this issue, plausible values methodology has been used in all large-scale assessments since the 1980s. Research has shown that this is the most reliable and valid approach for estimating performance on these data sets. Plausible values methodology was used to estimate respondents' literacy proficiencies based on their performance on the literacy tasks and responses to the background questionnaire. Using this methodology, ETS provided 10 plausible values for each respondent that must be used to estimate proficiency

for the target population as well as selected subgroups. Note that the plausible values cannot be used for individual reporting, only for group-level reporting.

Analyses such as multiple regressions need to incorporate all 10 plausible values. This requires systematically replicating the analyses using each plausible value in order to accurately reflect measurement errors. Failure to follow such procedures would result in an underestimation of the error associated with the regression coefficient and an overestimation of the effect size, in other words, both the significance and the size of estimated effects would be incorrect.

It is important to understand that international surveys such as PIAAC, STEP, PISA, TIMSS and PIRLS¹⁸ are designed to assess the knowledge and skills of a population rather than individuals. So the goal of reducing error around inferences about the target population is more important than the goal of reducing error around an estimate of ability or proficiency for an individual. STEP was designed using the same assumptions and goals as other international surveys, including PIAAC and PISA. And, as in other large-scale national and international surveys, the challenge is to employ a methodology that allows for adequate coverage of the construct or domain of interest in a relatively short administration time.

It is important to recognize that plausible values are not the same as individual assessment scores. Rather, they are random samples drawn from the posterior distribution of each respondent in the STEP survey that accurately reflect the measurement error. Through a process known as marginal estimation, the empirically based prior distribution of the type of group to which each respondent belongs can be calculated. These groups are defined by some combination of background variables such as gender, age, and education. The posterior distribution for an individual is a function of both the likelihood function derived from each respondent's answers to the set of cognitive questions they received and their values on a set of predictor variables obtained from the background questionnaire.

¹⁸ PIAAC – Programme for the International Assessment of Adults Competencies. PISA – Program for International Student Assessment. TIMSS - Trends in International Mathematics and Science Study. PIRLS - Progress in International Reading Literacy Study.

Table 11. Reading Proficiency Levels

Reading Proficiency Levels	
Literacy Below Level 1	0 to 175
The tasks at this level require the respondent to read brief texts on familiar topics to locate a single piece of specific information. Only basic vocabulary knowledge is required, and the reader is not required to understand the structure of sentences or paragraphs or make use of other text features. There is seldom any competing information in the text and the requested information is identical in form to information in the question or directive. While the texts can be continuous, the information can be located as if the text were noncontinuous. Tasks below Level 1 do not make use of any features specific to digital texts.	
Literacy Level 1	176 to 225
Most of the tasks at this level require the respondent to read relatively short digital or print continuous, noncontinuous or mixed texts to locate a single piece of information which is identical to or synonymous with the information given in the question or directive. Some tasks may require the respondent to enter personal information into a document, in the case of some noncontinuous texts. Little, if any, competing information is present. Some tasks may require simple cycling through more than one piece of information. Knowledge and skill in recognizing basic vocabulary, evaluating the meaning of sentences, and reading of paragraph text is expected.	
Literacy Level 2	226 to 275
At this level, the complexity of text increases. The medium of texts may be digital or printed, and texts may comprise continuous, noncontinuous or mixed types. Tasks in this level require respondents to make matches between the text and information, and may require paraphrase or low-level inferences. Some competing pieces of information may be present. Some tasks require the respondent to <ul style="list-style-type: none"> • cycle through or integrate two or more pieces of information based on criteria, • compare and contrast or reason about information requested in the question, or • navigate within digital texts to access and identify information from various parts of a document. 	
Literacy Level 3	276 to 325
Texts at this level are often dense or lengthy, including continuous, noncontinuous, mixed or multiple pages. Understanding text and rhetorical structures become more central to successfully completing tasks, especially in navigation of complex digital texts. Tasks require the respondent to identify, interpret or evaluate one or more pieces of information and often require varying levels of inferencing. Many tasks require the respondent construct meaning across larger chunks of text or perform multistep operations in order to identify and formulate responses. Often tasks also demand that the respondent disregard irrelevant or inappropriate text content to answer accurately. Competing information is often present, but it is not more prominent than the correct information.	
Literacy Level 4	326 to 375
Tasks at this level often require respondents to perform multiple-step operations to integrate, interpret, or synthesize information from complex or lengthy continuous, noncontinuous, mixed, or multiple type texts. Complex inferences and application of background knowledge may be needed to perform successfully. Many tasks require identifying and understanding one or more specific, non-central ideas in the text in order to interpret or evaluate subtle evidence claim or persuasive discourse relationships. Conditional information is frequently present in tasks at this level and must be taken into consideration by the respondent. Competing information is present and sometimes seemingly as prominent as correct information.	
Literacy Level 5	376 to 500
At this level, tasks may require the respondent to search for and integrate information across multiple, dense texts; construct syntheses of similar and contrasting ideas or points of view; or evaluate evidence-based arguments. Application and evaluation of logical and conceptual models of ideas may be required to accomplish tasks. Evaluating reliability of evidentiary sources and selecting key information is frequently a key requirement. Tasks often require respondents to be aware of subtle, rhetorical cues and to make high-level inferences or use specialized background knowledge.	

Plausible values are not assessment scores for individuals in the usual sense, but rather are imputed values, drawn from a conditional distribution, that are used to estimate population characteristics correctly. When used appropriately, they provide consistent estimates of population statistics with appropriate measurement errors. Plausible values can be used as the basis for typical statistical estimates like means, standard deviations, and percentages above or below selected cut points.

Stata STEP Module

A Stata module was tailored to meet the STEP estimation procedures presented above in order to facilitate the analysis and help with results reporting for the STEP Skills Survey. Several specificities linked to the STEP survey are embedded in the STEP module in Stata. Users opting for other software or programs (such as the ETS modules for PIAAC in SPSS) must take the specificities into account when undertaking estimations and reporting results. The STEP Skills Survey uses basic population weights. The sampling weights have been estimated to represent the urban population of the countries surveyed. Thus, one must always check the type of variance estimation used in order to avoid using other techniques that involve replicate weights (for example, Jackknife and BRR, among others) in other software. Second, it is important to use *all* the plausible values marked from 1 to 10 in the estimation. Finally, country-specific stratification must be taken into account when relevant. Additional information on the STEP Stata Module is presented in Appendix 2.

3. Matching Skills from the STEP Household and Employer Surveys

The STEP household and employer surveys were designed jointly with the view for researchers to be able to match skills individuals reported having with skills employers reported looking for. An effort was made to use consistent definitions of skills in both surveys, although it was not always possible to reflect the full range of skills in both surveys. For example, while teamwork was a very important job-relevant skill in the employer survey, asking specific questions about aspects of teamwork was limited in the household survey. This was due

principally to the dual constraints of keeping the surveys to a manageable length and keeping questions as concrete, clear, and simple as possible. Moreover, questions are often not phrased in the same way in both surveys. This was done to obtain the maximum level of accuracy of answers in the context of each survey. Table 12 provides a correspondence between the skills measured in the two surveys.

By combining information in the employer and household surveys, one can get a better picture of labor market mismatch:¹⁹

- Does the workforce provide the right skills for employers?
- Compare the distribution of skills present in the workforce to the distribution of skills demanded and the distribution of skills being used.
- A single measure can quantify mismatch: the mean squared difference between the shares of skills supplied and demanded:

$$\sum_{s \in \{Skills\}} \frac{(Supply_s - Demand_s)^2}{N_{Skills}}$$

where $Supply_s$ is the share of the workforce possessing skill s , $Demand_s$ is the share of employers demanding skill s (this can use multiple definitions, such as whether a skill is being used, whether it is in the top half of skills sought, in the top 5, and so on), and N_{Skills} is the number of different skills categories considered.

- Mismatch can also be broken down by skill type (cognitive, non-cognitive and technical) or individual skill set.
- Are individual educational decisions reflective of firm beliefs about the performance of the educational/training system?

¹⁹ This analysis is limited to the coverage of the employer survey.

Table 12. Skill Domains and Matching Questions in the STEP Household and Employer Surveys

Type of Skill	Domain	STEP Employer Survey Question	STEP Household Survey Question
Cognitive	Reading	Q.2.05 Q.3.02	Q.5a.4-10 Q.5a.21-23
	Writing	Q.2.06 Q.3.02	Q.5a.11-17 Q.5a.24-26
	Numeracy	Q.2.07 Q.3.02	Q.5a.18-19 Q.5a.27
Behavior and Personality Traits (Socio-emotional)	Big Five Inventory (Conscientiousness, Openness to experience, Neuroticism, Agreeableness, Extraversion)	Q.3.03	Q.6.1.01-1.26
Job-relevant Skills (Skills at Work)	Interpersonal skills	Q.2.11 Q.3.02	Q.5b.4-6 Q.5b.13
	Use of Technology	Q.2.12	Q.5b.15 Q.5b.17-22 Q.5b.27 Q.5b.30-31
	Job Specific Skills	Q.2.10 Q.3.02	Q.5b.7-9 Q.5b.12 Q.5b.28-29
	Language	Q.2.09 Q.3.02	Q.5b.11
	Autonomy	Q.3.02	Q.5b.16
	Solving & Learning	Q.2.08	Q.5b.10

VI. RESOURCES AND WAY FORWARD

1. Resources

The STEP Skills Measurement data files will be available in the World Bank's Microdata Catalog ([link](#)) starting on July 15, 2014. The *STEP Skills Measurement Program Collection* will provide a unique source of information on the STEP Skills Measurement program. It will be updated regularly to share new findings and materials related to the STEP program with a wide audience.

Country Reports

The following reports have already been published:

- **SNAPSHOT 2014** • *STEP Skills Measurement Study: Snapshot 2014*, by Alexandria Valerio, Maria Laura Sanchez Puerta, Pierre Gaëlle, Tania Rajadel, and Sebastian Monroy Taborda (Washington, DC : World Bank Group, 2014) ([link](#))²⁰
- **LAO PDR** • *Lao PDR—Skills for Quality Jobs And Development in Lao PDR: A Technical Assessment of the Current Context*, by Ximena Del Carpio, Yuki Ikeda, and Michele Zini. (Washington, DC : World Bank Group, 2013) ([link](#))²¹
- **SRI LANKA** • *Building the Skills for Economic Growth and Competitiveness in Sri Lanka* (Directions in Development, Human Development Series) by Halil Dunder, Benoit Millot, Yevgeniya Savchenko, Harsha Aturupane, and Tilkaratne A. Piyasiri (Washington, DC: World Bank Group, 2014) ([link](#))²²

²⁰http://www.worldbank.org/content/dam/Worldbank/Feature%20Story/Education/STEP%20Snapshot%202014_Revised_June%20202014%20%28final%29.pdf

²¹ <http://documents.worldbank.org/curated/en/2013/11/19300212/lao-pdr-skills-quality-jobs-development-lao-pdr-technical-assessment-current-context>

²² <http://documents.worldbank.org/curated/en/2014/01/19556815/building-skills-economic-growth-competitiveness-sri-lanka>

- **VIETNAM** • *Skilling up Vietnam: Preparing the Workforce for a Modern Market Economy* (Vietnam Development Report 2014) by Christian Bodewig and Reena Badiani-Magnusson (Washington, DC: World Bank Group, 2014) ([link](#))²³
- **YUNNAN PROVINCE (CHINA)** • *Developing Skills for Economic Transformation and Social Harmony in China: Study of Yunnan Province* (Directions in Development, Human Development Series) by Xiaoyan Liang and Shuang Chen (Washington, DC: World Bank, 2013) ([link](#))²⁴

Materials

The *STEP Skills Measurement Data Collection* will feature all STEP household survey datasets and background documentation required to understand how skills are measured in the STEP surveys and how to use the data. In addition to the present methodology note, the *STEP Skills Measurement Data Collection* will provide country-by-country datasets and household questionnaires.

Technical documentation will also be made available to users seeking particular information on specialized topics (such as sampling and weighting, operations, and technical standards).

2. Way Forward

The STEP surveys, used in conjunction with other tools such as SABER, promise to enrich tremendously the policy dialogue that the World Bank leads with client countries. At the same time, the richness of the data means that they will be used for years to come by researchers.

Additional countries have expressed interest in administering the STEP surveys. Doing so will bring additional internationally comparable data.

²³ <http://www.worldbank.org/en/country/vietnam/publication/vietnam-development-report2014-skilling-up-vietnam-preparing-the-workforce-for-a-modern-market-economy>

²⁴ <https://openknowledge.worldbank.org/handle/10986/16197>

VII. REFERENCES

- Almlund, Mathilde, Angela L. Duckworth, James J. Heckman, and Tim D. Kautz. 2011. *Personality Psychology and Economics*. Working Paper no. 16822. Cambridge, MA: National Bureau of Economic Research.
- Autor, David H., and Michael J. Handel. 2009. *Putting Tasks to the Test: Human Capital, Job Tasks and Wages*. Working Paper no. 15116. Cambridge, Mass.: National Bureau of Economic Research.
- Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Investigation." *Quarterly Journal of Economics* 118 (3) (November): 1279–1333.
- Borghans, Lex, Angela L. Duckworth, James J. Heckman, and Bas Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43 (4): 972–1059.
- Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.
- Cook, John D., Sue J. Hepworth, Toby D. Wall, and Peter B. Warr. 1981. *The Experience of Work: A Compendium and Review of 249 Measures and Their Use*. New York: Academic Press.
- Diaz, Juan J., Omar Arias, and David V. Tudela. 2012. *Does Perseverance Pay as Much as Being Smart? The Returns to Cognitive and Non-Cognitive Skills in Urban Peru*. Mimeo. Washington, DC: The World Bank. <http://econweb.umd.edu/~Urzúa/DiazAriasTudela.pdf>
- Digman, John M. 1990. "Personality Structure: Emergence of the Five-Factor Model." *Annual Review of Psychology* 41: 417–40.
- Dodge, Kenneth A. 2003. "Do Social Information Processing Patterns Mediate Aggressive Behavior?" In *Causes of Conduct Disorder and Juvenile Delinquency*, edited by B.B. Lahey, T.E. Moffitt, and A. Caspi. New York: Guilford Press.
- Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92 (6): 1087–1101.
- Handel, Michael J. 2008a. *Measuring Job Content: Skills, Technology, and Management Practices*. Institute for Research on Poverty Discussion Paper No. 1357-08. Madison, WI: University of Wisconsin.

- _____. 2008b. *What Do People Do at Work? A Profile of U.S. Jobs from the Survey of Workplace Skills, Technology, and Management Practices (STAMP)*. Paper presented at the Labor Seminar, Wharton School of Management, University of Pennsylvania.
- _____. 2012. *Understanding Skills Measurement / Job Skills Requirement*. Technical Note Washington, DC: World Bank.
- Härdle, Wolfgang, and Leopold Simar. 2003. *Applied Multivariate Statistical Analysis*. Springer.
- Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–482.
- Holzer, Harry J. 1996. *What Employers Want: Job Prospects for Less Educated Workers*. New York: Russell Sage.
- Jensen, Arthur R. 1998. *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.
- John, Oliver P., and Sanjay Srivastava. 1999. "The Big Five Trait Taxonomy: History, Measurement and Theoretical Perspectives." In *Handbook of Personality: Theory and Research*, edited by L.A. Pervin and O.P. John. New York: The Guilford Press.
- Kaiser, Henry F. 1985. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika* 23: 187–20.
- Kohn, Melvin L., and Carmi Schooler. 1982. "Job conditions and Personality: A Longitudinal Assessment of Their Reciprocal Effects." *American Journal of Sociology* 87: 1257–86.
- Lang, Frieder R., Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G. Wagner. 2011. "Short Assessment of the Big Five: Robust Across Survey Methods Except Telephone Interviewing." *Behavior Research Methods* 43 (2): 548–67. doi: 10.3758/s13428-011-0066-z
- Macdonald, Kevin. 2014. *PV: Stata module to perform estimation with plausible values*. Statistical Software Components from Boston College Department of Economics <http://EconPapers.repec.org/RePEc:boc:bocode:s456951>
- Mann, Leon, Paul Burnett, Mark Radford, and Steve Ford. 1997. *The Melbourne Decision Making Questionnaire: An instrument for measuring patterns for coping with decisional conflict*. *Journal of Behavioral Decision Making*, 10(1), 1-19. doi:10.1002/(SICI)1099-0771(199703)10:1<1::AID-BDM242>3.0.CO;2-X
- Milkovich, George T., and Jerry M. Newman. 1993. *Compensation* (4th ed.). Homewood, IL: Irwin.

- Neisser, Ulric, Gwyneth Boodoo, Thomas Jr Bouchard, Wade A. Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, R. Perloff, Robert J. Sternberg, and S. Urbina. 1996. "Intelligence: Knowns and unknowns." *American Psychologist* 51, 77-101.
- Paunonen, Sampo V. 2003. "Big Five Factors of Personality and Replicated Predictions of Behavior." *Journal of Personality and Social Psychology* 84: 411–22.
- Paunonen, Sampo V., and Michael C. Ashton. 2001. "Big Five Factors and Facets and the Prediction of Behavior." *Journal of Personality and Social Psychology* 81 (3): 524–39.
- Peterson, Norman G., Michael D. Mumford, Walter C. Borman, P. Richard Jeanneret, and Edwin A. Fleishman. 1999. *An Occupational Information System for the 21st Century: The Development of O*NET*. Washington, DC: American Psychological Association.
- Roberts, Brent W., Joshua J. Jackson, Jennifer V. Fayard, Grant Edmonds, and Jenna Meints. 2009. "Conscientiousness." In *Handbook of Individual Differences in Social Behavior*, edited by M. Leary and R. Hoyle. New York, NY: Guilford.
- Rotundo, Maria, and Paul R. Sackett. 2004. *Specific versus general skills and abilities: A job level examination of relationships with wage*. *Journal of Occupational and Organizational Psychology*, 77, 127-148. doi: 10.1348/096317904774202108.
- Soto, Christopher J., Oliver P. John, Samuel D. Gosling, and Jeff Potter. 2008. "The Developmental Psychometrics of Big Five Self-Reports: Acquiescence, Factor Structure, Coherence, and Differentiation From Ages 10 to 20." *Journal of Personality and Social Psychology* 94: 718–37.
- Sum, Andrew. 1999. *Literacy in the Labor Force: Results from the National Adult Literacy Survey*. NCES 1999–470. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- United States Department of Labor. Secretary's Commission on Achieving Necessary Skills. 1991. "What work requires of schools. A SCANS report for America 2000". <http://wdr.doleta.gov/SCANS/whatwork/whatwork.pdf>. Washington, DC: U.S. Department of Labor
- Walker, Susan P., Theodore D. Wachs, Julie Meeks Gardner, Betsy Lozoff, Gail A. Wasserwan, Ernesto Pollitt, and Julie A. Carter. 2007. "Child Development: Risk Factors for Adverse Outcomes in Developing Countries." *The Lancet* 369 (9556): 145–57.

VIII. APPENDIX

Appendix 1. STEP Reading Literacy Assessment | Sample Items

Three sample literacy items used in the STEP literacy assessment are presented in the boxes below. The items represent the range of literacy tasks included in the assessment across the proficiency levels. For each item, respondents were given the directions to use the information provided about each topic to answer the question.

Sample Item 1: Below Literacy Level 1

“**Preschool Rules**” represents an easy item that at least 50% of respondents with scale scores in the Below Level 1 range (0-175) would be expected to answer correctly.

Preschool Rules

Welcome to our Preschool! We are looking forward to a great year of fun, learning and getting to know each other. Please take a moment to review our preschool rules.

- **Please have your child here by 9:00 am.**
- **Bring a small blanket or pillow and/or a small soft toy for naptime.**
- **Dress your child comfortably and bring a change of clothing.**
- **Please no jewelry or candy. If your child has a birthday please talk to your child’s teacher about a special snack for the children.**
- **Please bring your child fully dressed, no pajamas.**
- **Please sign in with your full signature. This is a licensing regulation. Thank you.**
- **Breakfast will be served until 7:30 am.**
- **Medications have to be in original, labeled containers and must be signed into the medication sheet located in each classroom.**
- **If you have any questions, please talk to your classroom teacher or to Ms. Marlene or Ms. Tree.**

8. What is the latest time that children should arrive at preschool?

Sample Item 2: Literacy Level 1

“Swimmer Completes” is a relatively easy item that at least 50% of respondents with scale scores in the Level 1 range (176-225) would be expected to answer correctly.

Swimmer completes Manhattan marathon

The Associated Press

NEW YORK — University of Maryland senior Stacy Chanin on Wednesday became the first person to swim three 28-mile laps around Manhattan.

Chanin, 23, of Virginia, climbed out of the East River at 96th Street at 9:30 p.m. She began the swim at noon.

A spokesman for the swimmer, Roy Brunett, said Chanin had kept up her strength with “banana and honey sandwiches, hot chocolate, lots of water and granola bars.

Chanin has twice circled Man-

hattan before and trained for the new feat by swimming about 28.4 miles a week. The Yonkers native has competed as a swimmer since she was 15 and hopes to persuade Olympic authorities to add a long-distance swimming event.

The Leukemia Society of America solicited pledges for each mile she swam.

In July 1983, Julie Ridge became the first person to swim around Manhattan twice. With her three laps, Chanin came up just short of Diana Nyad’s distance record, set on a Florida-to-Cuba swim.

5. At what age did Ms. Chanin begin to swim competitively?

Sample Item 3: Literacy Level 2

“Physical Exercise Equipment” is a relatively easy item that at least 50% of respondents with scale scores in the Level 2 range (226-275) would be expected to answer correctly.

Physical Exercise Equipment

How to choose?—

- 1 Decide what effects you want the exercise to have on your body.
- 2 Assess the space you have available at home.
- 3 Choose the equipment that suits your objectives. If necessary, ask a specialist for advice.

For example:

OBJECTIVE	STRATEGY	EQUIPMENT
Burn off calories	Cardiovascular exercises	Rowing machine, Bicycle, Ski machine, Treadmill, Stairs...
Strengthen your muscles	Endurance exercises	Bench for Press-ups, Weights and Dumbbells, Elastic Tubes...

	Cardio-training				
Effects on...	Exercise bicycle	Rowing machine	Stepper	Treadmill	Air trainer
Arm Strength	Ineff-ective	Good	Average	Ineff-ective	Good
Leg strength	Good	Very good	Average	Very good	Good
Abdominal muscles	Average	Very good	Good	Good	Average
Overall muscle building	Ineff-ective	Very good	Ineff-ective	Average	Ineff-ective
Heart/arteries	Very good	Good	Very good	Very good	Good
Flexibility	Ineff-ective	Good	Ineff-ective	Ineff-ective	Average
Joints	Good	Very Good	Good	Good	Good
Slimming	Good	Average	Very good	Good	Good
Dangers	None	Back	None	Legs	

	Muscle building							
	Dumb-bells, weights	Elastic	Gym bench	Muscle-building bench	Multi-trainer	AB trimmer	AB shape	AB roller
Arm Strength	Very good	Very good	Good	Good	Good	Very good	Good	Good
Leg strength	Ineff-ective	Good	Average	Good	Good	Ineff-ective	Good	Good
Abdominal muscles	Ineff-ective	Good	Very good	Good	Average	Very good	Very good	Very good
Overall muscle building	Average	Good	Good	Good	Average	Good	Good	Good
Heart/arteries	Ineff-ective	Average	Average	Average	Good	Average	Average	Average
Flexibility	Average	Average	Good	Ineff-ective	Ineff-ective	Average	Good	Good
Joints	Good	Average	Average	Good	Good	Average	Average	Average
Slimming	Ineff-ective	Average	Good	Average	Average	Good	Good	Good
It is best to learn to use these types of apparatus properly before you make a major effort								

3. Which muscles will benefit most if you use the gym bench?

Appendix 2. STEP Stata Module

The first step to use the Stata module is to set up the program and the data.²⁵ If the Stata module has not been installed, please use the following command to do so: *ssc install pv, replace*. This should be done regularly to check for updated versions. After opening the pertinent data set, then set up the weights, using the *svyset* command to enable the *svy* prefix.

```
. u "$in/Primer_STEP_data.dta", replace

. svyset cluster [pw=W_FinSPwt]

      pweight: W_FinSPwt
      VCE: linearized
Single unit: missing
  Strata 1: <one>
    SU 1: cluster
    FPC 1: <zero>
```

Or for a country with stratified data,

```
. u "$in/Primer_STEP_data_stratified.dta", replace

. svyset cluster [pw=W_FinSPwt], strata(strata) singleunit(centered)

      pweight: W_FinSPwt
      VCE: linearized
Single unit: centered
  Strata 1: strata
    SU 1: cluster
    FPC 1: <zero>
```

The Stata module developed by Kevin Macdonald, with support from the STEP Core Team and tailored to work with the STEP Skills Survey data, works as a prefix to any required command. The basic syntax for the STEP Skills Survey data is as follows:²⁶

pv, pv(varlist) [options]: svy, subpop(varname): command 1 ||| command 2

²⁵ For more information on how to use plausible values with the STEP survey, see STEP Skills Measurement Survey | Analyzing the Literacy Proficiency Scores, 2014

²⁶ Adapted from Macdonald (2014).

Where *varlist* is the place to list the plausible values (either the score or the levels), *options* are the options for the *pv* command, *varname* refers to the variable that the estimation is being conditioned to (i.e., female, self-employed) and *command 1* and *command 2* are the placeholders for the desired command (i.e., mean, regression, margins).²⁷

Descriptive Statistics

The Stata Module can be used for various types of estimations. This subsection presents two examples of descriptive statistics to demonstrate the use of the mentioned module.

Example 1: Mean literacy proficiency by age group

The STEP Skills Survey includes a large number of background variables in addition to the literacy proficiency scores. A possible estimation could be the mean literacy proficiency score by age group. A variable was generated with values 1, 2 and 3 for the following age cohorts: 15-24, 25-54 and 55-65.

```
. pv, pv(PVLIT*): svy: mean @pv, over(age_group)
```

(output omitted)

```
Number of observations: 3405
Average R-Squared: .
```

	Coef	Std Err	t	t Param	P> t
PVLIT10:_subpop_1	277.55433	2.2097933	125.60194	115.79108	1.32e-125
PVLIT10:_subpop_2	242.07176	2.2198465	109.04887	229.63429	8.94e-200
PVLIT10:_subpop_3	227.705	3.3563456	67.843133	234.81189	2.93e-156

Note the use of the *@pv* to denote the place where the plausible values are within an estimation. In this example, results suggest that the younger cohorts do better (_subpop_1) than the older cohorts (_subpop_3). The names of the categories are not always visible

²⁷ More detailed information can be found in the Help File in Macdonald (2014)

when using the *mean* command. If needed, the following command provides labels / category names.

```
. svy: mean PVLIT1, over(age_group)
(running mean on estimation sample)
```

Survey: Mean estimation

Number of strata =	1	Number of obs =	3405
Number of PSUs =	227	Population size =	6290361
		Design df =	226

```
_subpop_1: age_group = 15-24
_subpop_2: age_group = 25-54
_subpop_3: age_group = 55-64
```

Example 2: Hourly earnings by level of literacy proficiency

Breaking down key variables by the different levels of plausible values (for instance, hourly earnings) may also provide useful descriptive statistics. In this case, each of the 10 plausible values must be divided into the six proficiency levels. The number of observations must also be checked. Sometimes, the number of observations is very low in the top or bottom levels. In such instances, it is suggested to lump the two bottom or top levels together.

```
. pv, pv(1_PVLIT*p): svy: mean earnings_h, over(@pv)
```

(output omitted)

	Coef	Std Err	t	t Param	P> t
earnings_h:_subpop_1	5008.7505	582.92425	8.5924553	136.94316	1.706e-14
earnings_h:_subpop_2	7100.9705	1568.9365	4.5259769	202.24958	.00001025
earnings_h:_subpop_3	8798.7639	2075.9697	4.2383874	46.95619	.00010446
earnings_h:_subpop_4	8852.7004	2277.3264	3.8873218	43.970168	.00033824

In this case, hourly earnings seem to increase with literacy proficiency. However, they appear similar for the last two levels.

Regression Analysis

This subsection presents two types of regression analysis that can be done using the Stata Module. This is not an exhaustive list of possible analyses. They are chosen to demonstrate features of the basic set-up. Examples selected below show that the Stata module allows the plausible values to be treated either as an independent or dependent variable.

Example 3: Mincer regression

The first example is the classic mincer equation and it adds the literacy proficiency score as right-hand variable. For this example, the sample includes only males and the estimation is done only for wage workers. The hourly earnings variable has been transformed to log hourly earnings and it has been trimmed on the corners. Also, *experience* and *experience2* are the potential experience (=Age-years of education-6) and potential experience squared, respectively.

```
. pv, pv(PVLIIT*): svy, subpop(wage_workers): reg log_hourly_earnings experience experience2 yea  
> rs_educ @pv
```

(output omitted)

```
Number of observations: 768  
Average R-Squared: .2536754920292164
```

	Coef	Std Err	t	t Param	P> t
experience	.03270909	.01127205	2.9017882	144.45973	.00429198
experience2	-.00034902	.0002493	-1.3999982	144.63319	.16365526
years_educ	.08460669	.01824233	4.6379319	148.32683	7.664e-06
pv	.00133954	.00078026	1.7168	129.5909	.08840486
_cons	.61848626	.2076284	2.9788134	152.25459	.00336856

Note that in this case the average R-squared is reported. In this example, an increase of 1 point on the literacy score, holding everything else constant, would result in a 0.1 log points increase in hourly earnings.