STEPSKILLS MEASUREMENT

THE WORLD BANK
www.worldbank.org

A guide to understanding the literacy assessment of the STEP Skills Measurement Survey

November 2014

ETS

This paper was commissioned by the team responsible for managing the STEP Skills Measurement program. The team is composed by Alexandria Valerio, Maria Laura Sanchez Puerta, Tania Rajadel, Gaëlle Pierre and Sebastian Monroy.

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. One of these instruments is an assessment of reading literacy designed to identify respondent's levels of competence at accessing, identifying, integrating, interpreting, and evaluating information. The present document is a reference for readers seeking in depth information regarding the design of the literacy assessment of the STEP survey. The first chapter provides an overview of the assessment, including a definition of literacy, a summary of the assessment content, and descriptions of the proficiency levels of the scale. The second chapter describes the procedures followed for translating, adapting, administering and scoring the materials. The final section presents the guidelines followed for scaling and analyzing the literacy data using plausible values and sampling weights. Much of the overview and data scaling sections have been adapted from the PIAAC publications, namely the technical report (OECD, 2013) and development frameworks (OECD, 2012).

Chapter 1: An Overview of the STEP Literacy Assessment

Introduction

Through the administration of the literacy assessment module, the World Bank's STEP survey provides information about respondents' skills in reading literacy as well as reading components, which represent the basic set of skills that provide necessary preconditions for gaining meaning from written text. A primary goal for the design of the STEP literacy assessment was to be able to link to the Survey of Adult Skills, which was administered by the Organisation for Economic Co-operation and Development as part of its Programme for the International Assessment of Adult Competencies (PIAAC). This design had two major advantages: It capitalized on an established item pool developed for and successfully implemented in an international context, and it allowed for results to be reported on a common scale with established descriptions for interpreting the proficiency levels of the scale.

STEP was administered in a total of 12 countries: Armenia, Bolivia, Colombia, Georgia, Ghana, Kenya, Laos, Macedonia, Sri Lanka, the Ukraine, Vietnam, and Yunnan Province (People's Republic of China). ETS worked closely with the World Bank and the staff from the national survey firms, as well as staff from cApStAn (linguistic quality control) and the IEA Data Processing Center (data management) to make sure that key steps were taken to establish a link between the STEP literacy assessment and PIAAC in terms of instrumentation and survey operations.[1]

This section provides an overview of the literacy assessment administered in STEP, including a definition of the construct, a summary of the assessment content, and descriptions of the proficiency levels for the literacy scale.

Defining literacy

The literacy items selected for STEP were all developed based on the literacy frameworks developed for PIAAC, which defines literacy 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"* (OECD, 2012). 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—have been taken into consideration when selecting the texts and developing items included in the STEP literacy assessment.

[1] ETS has prepared a separate Methodology Note, which addressed the comparability of STEP with PIAAC titled "Comparability of the STEP Literacy Assessment with the OECD's Survey of Adult Skills."

Contexts for reading

For adults, reading normally is 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 comprised the following contexts: home and family, health and safety, community and citizenship, work and training, education, and leisure and recreation.

Cognitive processes with text

Both Adult Literacy and Lifeskills (ALL) survey/International Adult Literacy Survey (IALS) and PIAAC identified three broad types of tasks that readers were asked to carry out: those that require 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. The following three cognitive operations with text can be identified that are needed when working on items or tasks:

1) Access and identify information in the text
2) Integrate and interpret (relate parts of text to each other)
3) Evaluate and reflect (understanding of the text as a whole)

Reading components

As an extension of the main literacy assessment, STEP included an assessment of *reading components*. The Reading Components Assessment Framework built upon the basic principle that comprehension processes—that is, the "meaning construction" processes of reading—are built upon a foundation of component print skills that indicate the knowledge of how one's language is represented in one's writing system (the symbols and code system used when writing, such as the visual letters, words, and punctuation symbols) and the rules (e.g., decoding in alphabet writing systems) the individual needs to acquire to learn to read. This relationship may vary from language to language. In English, for example, the letters correspond to sounds in the language, though the sight-to-sound correspondence is not one-to-one (i.e., there are many sounds that correspond to single letters, and conversely many sounds may correspond to different letter or clusters of letters). Some written letters do not correspond to any sounds (i.e., silent letters as the letter "k" in the word "know") but may signal to the reader a change in the sound of other letters (i.e., in English, the "e" at the end of words like "cave," "save," and "pave" signal a different sound for the letter "a" than in the sound of "a" in "cat" and "hat.") In other words, to learn to read in a language requires learning the written symbols and the rules for how words and grammar are represented when the language is written down.

To capture interpretable chunks of this basic reading skill, the following reading components were identified for assessment:

- Word knowledge (vocabulary)
- Sentence processing

- Passage fluency (comprehension)

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 in tasks that examine a reader's ability and efficiency in processing the elements of the written language – letters/characters, words, sentences, and larger, continuous text segments.

Proficient readers are not only able to read words, sentences, and passages for meaning, they also do so relatively fluently, rapidly, and with ease; that is, with minimal exertion of conscious attention or effort. This is what we mean by efficiency or fluency—ease, accuracy, and speed of processing. For the skilled reader, reading is almost entirely an act of meaning construction. Decoding words, interpreting the punctuation, or parsing the syntax occurs, but it is not the reader's focus. It is like breathing—we are always doing so to stay alive but rarely attend to it consciously. Unfortunately for low skilled adults, the basic skills required to process the written word can be quite demanding of attention and effort, exhausting readers even as they try to construct basic meaning.

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. More detailed information about the reading components is given in the PIAAC literacy framework (OECD, 2012).

Assessment and instrument design

The design for the STEP literacy assessment had two primary goals: to provide items that target the lower end of the literacy scales (below Level 4) and to link to the literacy scale used in PIAAC. The selection of items for any assessment requires meeting certain constraints. In order to meet the psychometric linking requirements, the potential pool of items was limited to items used in PIAAC as well as some additional items from ALL and IALS.

STEP was a 45-minute assessment that allowed countries to be on the same reading literacy scale as used in PIAAC. The full module consists of a General Booklet and Exercise Booklets. Administering the General Booklet allowed us to get targeted information about the skills of individuals at the lower end of the literacy distribution, meeting an important goal of the STEP survey. It should also be noted that the selected items in the Exercise Booklets covered the full range of difficulty, allowing the survey to profile the full distribution of literacy skills in the adult populations of participating countries.

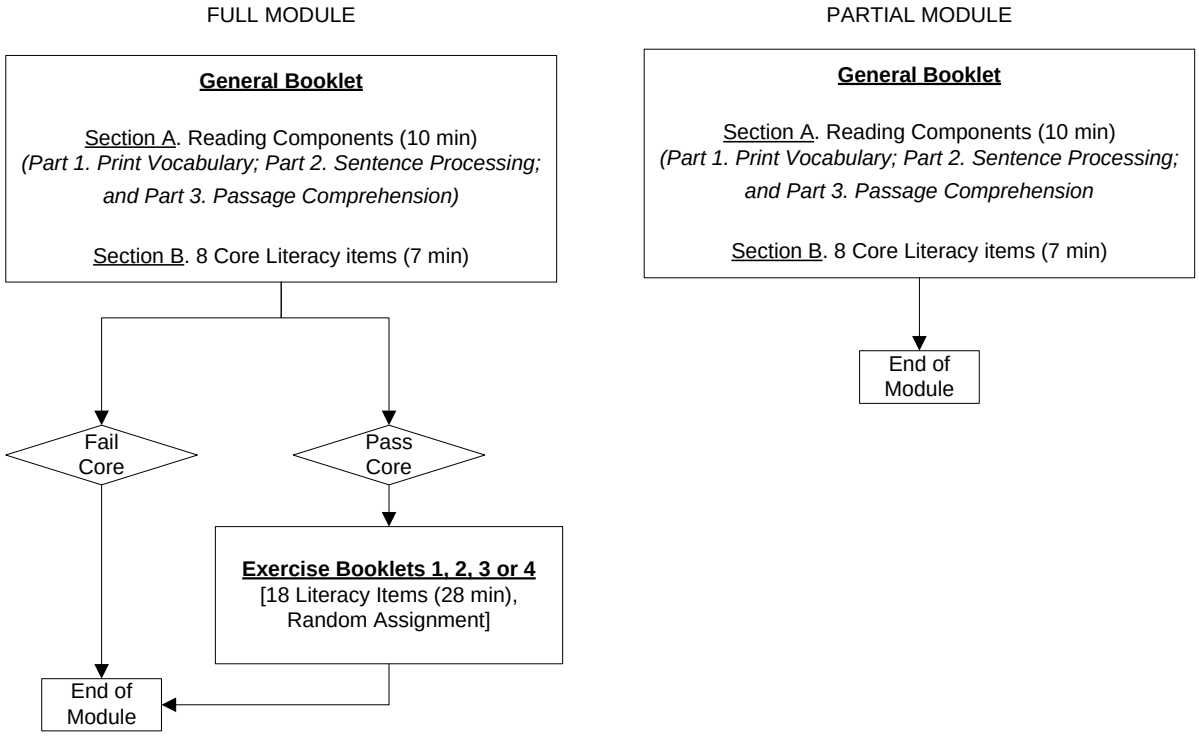
For reading literacy, stimulus materials were selected based on specifications provided in the framework. To the extent possible, stimuli for the PIAAC assessment were taken from real-world materials such as newspaper and magazine articles, advertisements, books, and forms that adults ages 16-65 would 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 1.1 illustrates the design of the literacy assessment and the administration workflow. For countries administering the full module, the General Booklet Part B was scored by the interviewer

during the interview and the score of that portion of the assessment determined whether the respondent proceeded to an Exercise Booklet.

Countries opting to administer the partial module included only the General Booklet in the literacy assessment and the interviewers were expected to score Section B during the interview.

Figure 1.1: Assessment designs for full and partial modules



General Booklet

The General Booklet consisted of two sections – Section A: Reading Components and Section B: Core cognitive literacy items.

Section A: Reading components item design

This section entailed 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. This section contained three parts – Part 1, Word Meaning (Print Vocabulary); Part 2, Sentence Processing; and Part 3, Passage Comprehension. The reading components took approximately 10 minutes to administer.

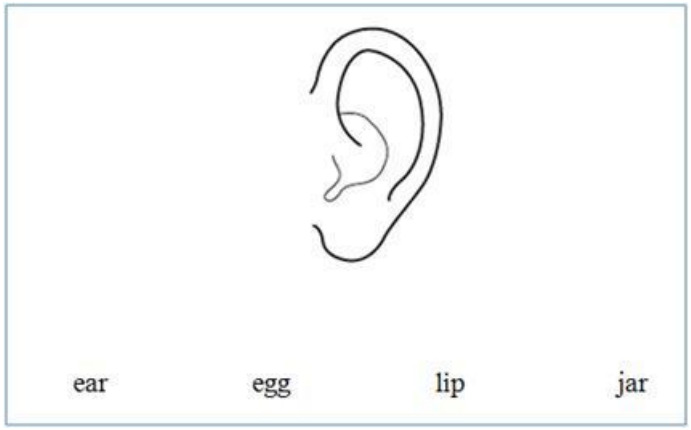
Word Meaning (Print Vocabulary)

Recognizing the printed symbols on the page as representing meaningful words is foundational to reading literacy. In the Word Meaning (Print Vocabulary) task set, the respondent must identify

everyday words that the average adult speakers of the language would understand if they heard the words spoken aloud. The items do not include specialized technical or academic words that would only be known by more educated individuals in the population. Instead, the words used are commonly known across country contexts (e.g., sun, triangle, foot).

The Word Meaning (Print Vocabulary) measure was, thus, useful to determine whether individuals could 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 language. Each item in this section presented an image and four word choices. The respondent had to circle the correct word choice that matched the picture. Target words were concrete, image-able nouns of common objects. Distractors were designed to tap similar semantic and/or orthographic features of the target word. This way, it was less likely that individuals could use only partial knowledge of spelling or visual symbols to guess the correct answer. For example, in the sample item below (see Figure 1.2), a reader might guess based on the first sound of the word "ear" that the spelling starts with "e". However, there are two choices that start with "e", making it more challenging to guess.

Figure 1.2: Sample word meaning item



Sentence Processing

The sentence is a natural "chunk" when reading continuous text (e.g., Kintsch, 1998). To build meaning from a sentence includes understanding all the words, parsing the syntactic structure, and encoding the propositions in memory. Depending on the specifics of a sentence, other operations might include making anaphoric inferences (e.g., relating pronouns to their referent), causal inferences, or knowledge-based inferences. Thus, each sentence requires some syntactic and semantic processing.

The Sentence Processing measure presented sentences of increasing difficulty (as indexed by length) and asked the respondent to make a sensibility 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 read the sentence and circled YES if the sentence makes sense or NO if the sentence did not make sense. This task demand is consistent with the "evaluation" goal of reading in the PIAAC reading literacy framework (OECD, 2012). Even at the most basic reading level, comprehension or

understanding may require evaluating text meaning against one's knowledge of the world to judge its veracity. One cannot always believe what one reads. Figure 1.3 below shows a set of sample Sentence Processing items.

Figure 1.3: Sample sentence processing 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

Passage Comprehension

Skilled reading is rapid, efficient, and fluent (silent or aloud). The Passage Comprehension task set targets silent reading for basic meaning comprehension in multiparagraph prose texts. As the adults read silently through a passage, they saw a word-choice item in selected sentences. They needed to circle the word among the alternatives that fit the meaning of the sentence. The incorrect choice was meant to be obviously wrong to a reader with some basic comprehension skills. The incorrect choice could be grammatically or semantically wrong.

The Passage Comprehension measure presented 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 used a forced-choice cloze paradigm—that is, a choice was given between a word that correctly completed a sentence in a passage and an option that was incorrect. The incorrect item was meant to be obviously wrong to a reader with some basic comprehension skills. The integration of decoding, word recognition, vocabulary, and sentence processing was required to construct the basic meaning of a short passage. The respondent was asked to read the passage and circle the word that makes the sentence made sense (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 1.4 with the options for selection underlined within the sentences.

Figure 1.4: Sample Passage Comprehension Items

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 items

This entailed a core set of eight literacy items that could be used to help sort the least literate from those with higher levels of skill. The Core took approximately seven minutes, on average, to administer. This core set of cognitive items was scored by the interviewer. Respondents who failed the Core were done with the interview, while those who passed the core proceeded to the Exercise Booklets (reading literacy).

Exercise Booklets

The assessment design for STEP specified a Core 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 assessed reading literacy, covering the full range of difficulty. Respondents who passed the Core were randomly assigned one of these four booklets.

The STEP survey includes four Exercise Booklets (Booklets 1, 2, 3, and 4). These booklets were assembled following the design provided in Figure 1.5 below. Each booklet had two blocks of nine literacy items, or 18 items in total. The booklets required 28 minutes, on average, for participants to complete.

Figure 1.5: 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

Reading components results: Accuracy and rate

The reading components measures were designed to provide information about what adults in Levels 1 and 2 *can do* with respect to selected building blocks of literacy proficiency.

- **Word Meaning (Print Vocabulary)** measured the extent to which participants can recognize the printed forms of common objects.
- **Sentence Processing** measured the extent to which participants can comprehend sentences of varying levels of complexity.
- **Basic Passage Comprehension** measured 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 answer were correct or incorrect).

Relationship of the reading components to one another and reading comprehension

People (adults or children) do not learn to read hierarchically, one component at a time. That is, one does not first learn how to read all the words in the language, then, only after those are learned, begin to learn how to process sentences, then once sentences are mastered, learn how to read passages. Instead, one learns very early on to build mental models of text meaning at a more global level. As each word of a text is recognized and its meaning is accessed in the mental lexicon of the reader, the word is integrated into propositions to construct higher order conceptual meaning (e.g., phrases like "he walks" "in the house," "fire and rain"). These propositions are compiled into larger meaningful structure and combined with background knowledge to form a schema or mental model of the situation described by the text. So if one is reading about going to a restaurant, all the knowledge about restaurants and procedures (e.g., ordering food, paying for it, etc.) are used to help structure and recall the meaning of the text.

The component skills are the building blocks for this more complex skill of reading literacy proficiency, but one would expect them all to be correlated to some extent. Once individuals know a few words, they can construct meaning from a sentence using those words. If they have schema knowledge about the content of a passage, then they will be able to use that knowledge, along with some word and sentence processing skill, to construct the gist of the passage.

The implications of this processing account of how components relate to reading comprehension ability are twofold. First, one would expect correlations among different ranges of the components to each other and to reading proficiency levels, except perhaps at the lowest extremes. That is, if adults cannot recognize any words, it seems logically unlikely that they can process many sentences comprising words they cannot recognize. On the other hand, it is feasible that an adult can recognize highly familiar words like those used in the word meaning components section but not know enough of the syntax or grammar of the language to score well on the sentence or passage tasks.

The literacy proficiency scale and proficiency levels

To adequately measure the skills of adults with differing educational backgrounds and life experiences, PIAAC included tasks that range 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 allowed 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

Reporting that one task falls at 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 requires. To provide a richer report of the PIAAC results, described proficiency scales were developed for each of the domains, describing what performance at various points along those scales means. To create these described proficiency scales, the expert groups in each domain met with psychometricians and test developers to review the PIAAC Main Study 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 below and one or more representative tasks are described to illustrate the key information-processing skills at each level.

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

- 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, noncentral 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.

STEP item distribution across proficiency levels

As mentioned earlier, the items selected for STEP were intended to focus on the lower levels of the PIAAC literacy scale. Figure 1.6 shows the distribution of the 43 STEP literacy items across the PIAAC literacy scale. Almost 90 percent of the items proposed for STEP fall within Levels 1 through 3, meeting the design requirements for the survey.

Figure 1.6: Difficulty distribution of STEP literacy items

PIAAC Literacy Scale	Distribution of Items in STEP
Level 1 and below (easiest)	**27%**
Level 2	**30%**
Level 3	**32%**
Level 4	**11%**
Level 5 (hardest)	

STEP sample items

Three sample literacy items used in the STEP literacy assessment are presented in the boxes below. The items represented 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.

Figure 1.7: 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?

_____

Figure 1.8: 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 Manhattan 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.

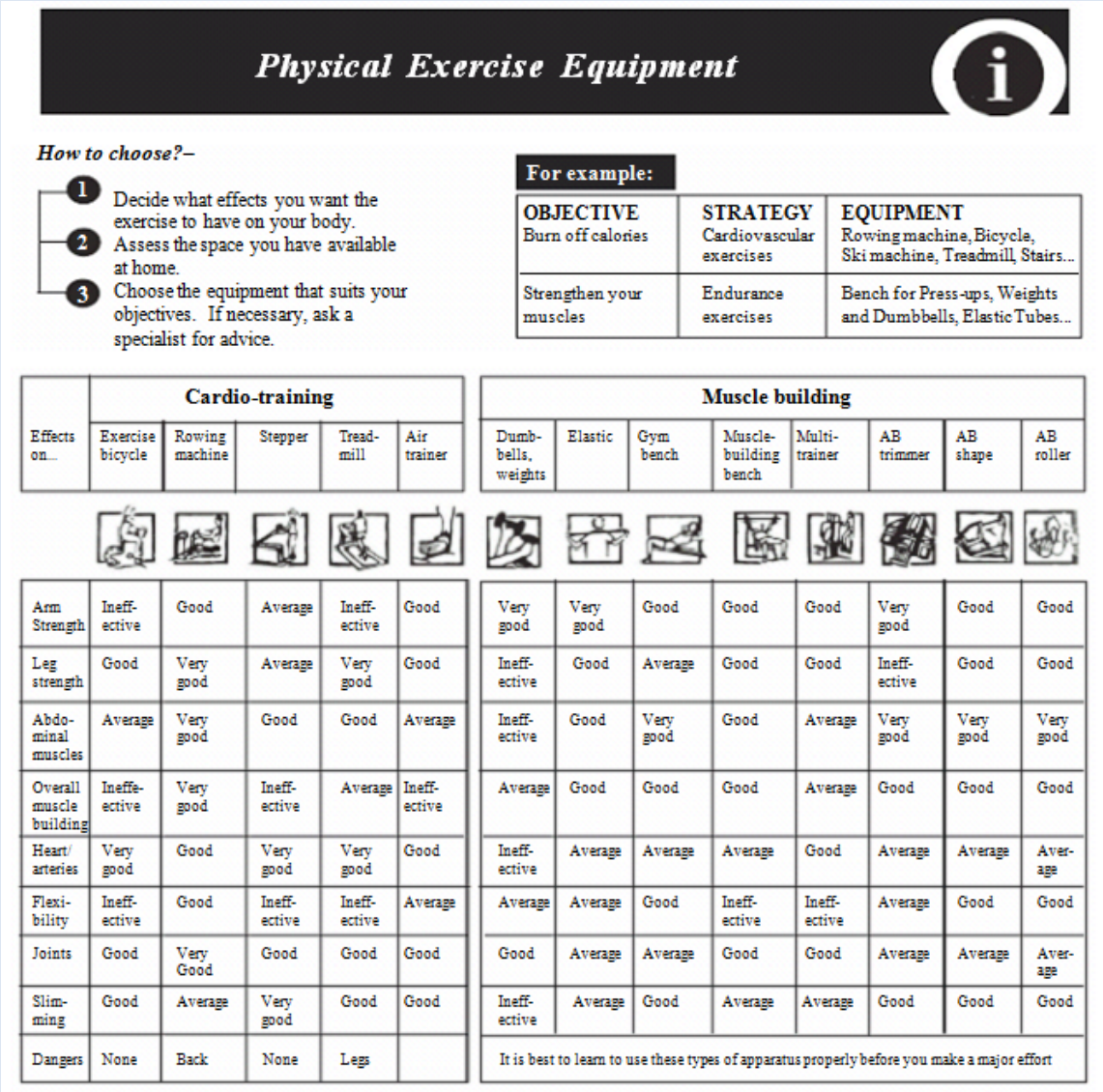
In July 1983, Julie Ridge became the first person to swim around Manhattan twice. With her three laps, Chanin came up just short of Diana Nyad's distance record, set on a Florida-to-Cuba swim.

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

_____

Figure 1.9: Literacy Level 2

"**Swimmer Completes**" is a relatively easy item that at least 50% of respondents with scale scores in the Level 1 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...

Effects on...	Cardio-training					Muscle building							
	Exercise bicycle	Rowing machine	Stepper	Tread-mill	Air trainer	Dumb-bells, weights	Elastic	Gym bench	Muscle-building bench	Multi-trainer	AB trimmer	AB shape	AB roller
Arm Strength	Ineff-ective	Good	Average	Ineff-ective	Good	Very good	Very good	Good	Good	Good	Very good	Good	Good
Leg strength	Good	Very good	Average	Very good	Good	Ineff-ective	Good	Average	Good	Good	Ineff-ective	Good	Good
Abdo-minal muscles	Average	Very good	Good	Good	Average	Ineff-ective	Good	Very good	Good	Average	Very good	Very good	Very good
Overall muscle building	Ineffe-ective	Very good	Ineff-ective	Average	Ineff-ective	Average	Good	Good	Good	Average	Good	Good	Good
Heart/ arteries	Very good	Good	Very good	Very good	Good	Ineff-ective	Average	Average	Average	Good	Average	Average	Aver-age
Flexi-bility	Ineff-ective	Good	Ineff-ective	Ineff-ective	Average	Average	Average	Good	Ineff-ective	Ineff-ective	Average	Good	Good
Joints	Good	Very Good	Good	Good	Good	Good	Average	Average	Good	Good	Average	Average	Aver-age
Slim-ming	Good	Average	Very good	Good	Good	Ineff-ective	Average	Good	Average	Average	Good	Good	Good
Dangers	None	Back	None	Legs		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?

_____

Chapter 2: Instrument and Data Preparations[2]

Introduction

This chapter describes the guidelines for translating and/or adapting the STEP literacy assessment materials, administering the STEP literacy booklets and scoring the literacy items. Although the World Bank staff had responsibility for organizing and training staff responsible for the STEP household survey overall, ETS provided the survey firms with specific guidelines and training for the administration and scoring of the STEP literacy booklets, as well as data entry and quality assurance. The guidelines and procedures and scoring training materials were adapted from the PIAAC publications to maintain comparability between PIAAC and STEP.

Translation and/or adaptation of the STEP literacy assessment materials

The STEP literacy assessment instruments, comprising cognitive test units, were prepared for administration to participating adults in 12 countries in 15 language versions. Localization (translation, adaptation for local use, and independent validation) of these instruments was a key aspect. The localization process was a complex operation involving staff from various organizations and components that followed different processes.

The process included the following activities:

- cApStAn Linguistic Quality Control, in close cooperation with ETS, developed the localization design and was responsible for implementing linguistic quality assurance (LQA) and linguistic quality control (LQC) processes.
- World Bank survey firms were responsible for appointing two professional translators to translate or adapt the materials from the English source files according to the Translation and Adaptation Guidelines provided with the materials.
- The two independent translations were reconciled into a draft national version of the materials and all national adaptations were documented for review by an independent translation verification company, cApStAn.
- cApStAn independently verified materials translated by the survey teams and proposed changes before finalization of the assessment booklets.

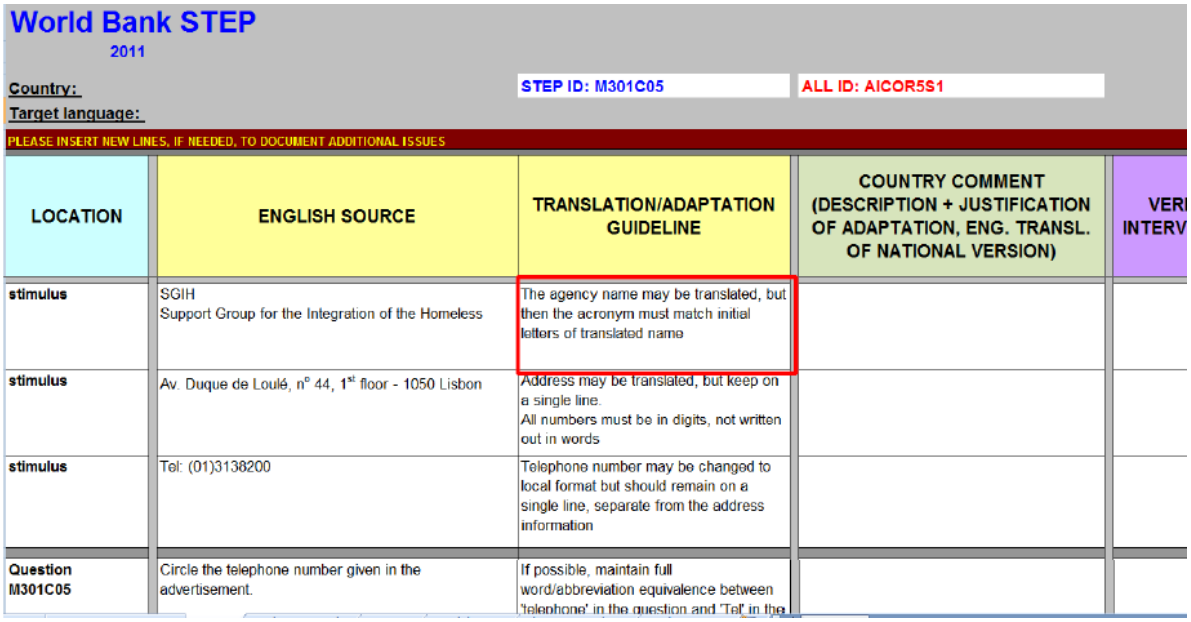
STEP literacy assessment localized instruments were produced for the following participating countries: Armenia, Bolivia, Colombia, Georgia, Ghana, Kenya, Laos, Macedonia, Sri Lanka, the Ukraine, Vietnam, and Yunnan Province (People's Republic of China). A total of 15 national versions were produced in 11 languages: Albanian, Chinese, English, Georgian, Lao, Macedonian, Sinhalese, Spanish, Tamil, Ukrainian, and Vietnamese.

[2] Developed by Educational Testing Service and cApStAn

The STEP localization design was based on the one used for PIAAC, which was in turn based on the design used for PISA (Programme for International Student Assessment). The processes implemented by cApStAn in cooperation with ETS included:

- Early identification of potential localization issues, via preliminary scrutiny of source assessment materials, to anticipate adaptation issues, ambiguities, cultural issues, or item translatability problems, with suggestions for either rewording or adding item-specific translation/adaptation guidelines.
- An adapted version of the PIAAC Translation/Adaptation Guidelines, a key document setting out requirements and roles, identifying linguistic difficulties, psychometric traps, cultural adaptations, and so on.
- Preparation of a tool called the Verification Follow-up Form (VFF) for documenting and monitoring the successive localization activities for each country (see Figure 2.1 for a sample worksheet from a VFF). This tool conveniently provided detailed item-specific translation and adaptation guidelines for the attention of all parties involved, including advice on adaptations that were mandatory, desirable or ruled out; advice on terminology problems and idiomatic expressions, literal or synonymous matches (e.g., between stimuli and items to be echoed, patterns in response options to be echoed, formatting issues).

Figure 2.1: Sample VFF showing item-specific guidelines



World Bank STEP
2011

Country:
Target language:

STEP ID: M301C05 | ALL ID: AICOR5S1

PLEASE INSERT NEW LINES, IF NEEDED, TO DOCUMENT ADDITIONAL ISSUES

LOCATION	ENGLISH SOURCE	TRANSLATION/ADAPTATION GUIDELINE	COUNTRY COMMENT (DESCRIPTION + JUSTIFICATION OF ADAPTATION, ENG. TRANSL. OF NATIONAL VERSION)	VERI INTERV
stimulus	SGIH Support Group for the Integration of the Homeless	The agency name may be translated, but then the acronym must match initial letters of translated name		
stimulus	Av. Duque de Loulé, n° 44, 1st floor - 1050 Lisbon	Address may be translated, but keep on a single line. All numbers must be in digits, not written out in words		
stimulus	Tel: (01)3138200	Telephone number may be changed to local format but should remain on a single line, separate from the address information		
Question M301C05	Circle the telephone number given in the advertisement.	If possible, maintain full word/abbreviation equivalence between 'telephone' in the question and 'Tel' in the		

- Participation in preparation and delivery of training sessions for the survey firms, held at the World Bank in Washington, DC in September 2011 for survey firms in seven countries administering STEP in 2012 and again in December 2012 for the remaining survey firms administering STEP in 2013.
- Continued phone and email assistance to national survey firms throughout the localization process.

The implemented LQC processes included:

- Verification by cApStAn verifiers of translated versions submitted by national survey firms against the source versions. Verification involved sentence-by-sentence comparison versus the source versions with reporting of residual errors and expert advice where corrective action was required.
- Analysis and selective implementation of edits after representatives from participating countries reviewed instruments and suggested changes, with reporting and follow-up of residual errors and/or unresolved issues.
- A final check procedure after national survey firms carried out their post-verification revision of instruments and ETS staff assembled and prepared booklets. This step was primarily to verify layout of the booklets prior to printing.

Administration of the STEP literacy assessment

Administration of the STEP literacy assessment focused on a set of guidelines for supervising and documenting the session during which the interviewers handed the pre-assigned literacy booklets to the respondents and guided the respondents through the sections within the booklets. In all countries, the literacy assessment began with the General Booklet, consisting of Section A: Reading Components and Section B: Core cognitive literacy items. As the time spent on each reading components item in Section A was an important part of the construct being measured, the interviewers were asked to record the time, in seconds, that a respondent took to answer each item.

Interviewers for countries participating in the full assessment, which included the literacy exercise booklets, received instructions in scoring the Core literacy items in Section B. Interviewers then calculated the total score across the Core items and either terminated the administration if a respondent did not receive a passing score of at least three correct responses or handed the respondent a pre-assigned exercise booklet to continue if a respondent passed the Core. At the end of the session, the interviewer collected the booklets from the respondents and completed all necessary documentation of the session.

Accuracy in the reporting of STEP results begins with scoring activities. Therefore, the scoring of the assessment items to determine whether respondents have answered the questions correctly needed to be carried out accurately and consistently among scorers within countries, as well as across all scorers in the participating countries. Using the PIAAC guidelines and procedures for scoring, as well as scoring training materials, the national survey firms followed prescribed steps for creating and training scoring teams, organizing booklets, and monitoring the scoring activities. A scoring supervisor was responsible within each scoring team to monitor the extent to which the scorers were adhering to the scoring rubrics for each item. This was done not only through extensive training exercises with the members of the scoring team, but also through monitoring the agreement in the “double scoring” of a portion of the booklets that were independently scored by two scorers.

Data entry and quality assurance

The IEA Data Processing Center (DPC) designed the software and procedures for survey firms to use for entering the information from the STEP literacy booklets, including the scores for the literacy items. The software allowed for the standardization of the underlying structure across the national databases and incorporated tools for survey firms to monitor the consistency of data entry. Documentation and training was provided to instruct data managers within the survey firms how to conduct checks to verify the quality of the data. The survey firms submitted all national data files to the IEA DPC for final data cleaning and processing prior to data analysis. Once the IEA DPC completed its final checks, the data were sent to ETS and the World Bank to be merged with the STEP background questionnaire data and be prepared for data analysis.

Chapter 3: Data Analysis and Scaling

Data Quality, IRT Analyses, and Population Modeling

Introduction

This chapter contains a complete description of the data analysis and scaling procedures for the STEP literacy assessment, including country-specific information and results.

STEP was designed to assess the cognitive skills of a sampled adult population based on the PIAAC literacy scale (including reading components). 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.

Items selected for STEP represent the literacy framework of PIAAC; the items were either administered in the PIAAC paper-based assessment, adapted from the survey's computer-based instruments, or administered in other large-scale adult literacy scales that have been previously linked to the PIAAC literacy scale. The target population for STEP was a subset of the total adult population (ages 16-65) of the PIAAC national samples. Both PIAAC and STEP were assessed by an interviewer face-to-face at home or a place convenient for the respondent. The systems of test administration, scoring, and the evaluation of scoring accuracies employed for STEP were comparable to those for the paper-based PIAAC assessment. The analysis methods and procedures for STEP were based on identical psychometric principles used for PIAAC.

The STEP design was based on matrix sampling, a variant of a sampling design most common to the major large-scale surveys, where each respondent was administered a subset of items from a larger pool, resulting in different groups of respondents answering different sets of items. The design enabled reducing the response burden for an individual while allowing the item pool to be expanded to represent the framework as completely as possible.

As a result, it was inappropriate to use any statistic based solely on the number of correct responses in reporting the survey results. But the limitations of conventional scoring were overcome by using Item Response Theory (IRT) scaling. 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) as well as items (by estimating certain item parameters through IRT models, e.g., item difficulty) in terms of a common scale, even if not all respondents take identical sets of items. In other words, if an item pool is used to measure a certain skill unidimensionally (i.e., only one skill is necessary to solve the items), respondents can be compared with one another even if they responded to different sets of items from this pool (given that the pool was scaled using a certain IRT model and showed appropriate model fit). IRT scaling thus makes it possible to describe distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables.

Before it 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 done to evaluate the handling and pattern of the missing values (i.e., missing by design, omitted by the respondent).

In order to link STEP and PIAAC in terms of 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 assured that the scale linkage of STEP to PIAAC could be established and inference structures preserved. To achieve this, the majority of item parameters in STEP were the same as in PIAAC (common item parameters); only a few items needed unique item parameters in certain countries (newly estimated item parameters in case they showed 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 to produce plausible values of literacy proficiency within each country.

In the following sections, the data evaluation process and the population model used for STEP scaling (IRT analysis, latent regression model, and computation of plausible values) are described.

Data handling and evaluation: Missing values, completion, item analysis, and scoring reliability

The assurance of the data quality was an important step prior to the IRT scaling and population modeling. Only if the analyses were based on correct data could reasonable and meaningful results be provided. Procedures for evaluating scoring and handling of missing data, data completion, and item analyses are described below.

Scoring and handling of missing data

STEP followed the same scoring guidelines and procedures as those applied in PIAAC for the paper-and-pencil administration. The literacy and reading components items were dichotomously scored: correct responses were scored as 1, and incorrect responses were scored as 0. The two kinds of missing values were scored differently: items that were administered but omitted by the respondent were scored as 8, and items that were not administered by design were scored as 9.

Regarding the handling of missing data, the STEP design followed a similar procedure to that used in PIAAC in order to maintain comparability between the two studies. STEP data have a characteristic structure of missing responses that is derived from the matrix sampling design and the instituted accommodation for respondents with very low literacy skills through core items. This structure is characterized by data missing completely at random due to the test design (random assignment of booklets) and data missing due to omitted responses. More specifically, there are different types of missing values within the cognitive part of STEP:

1) Missing by design: Items that were not presented to each respondent due to the matrix sampling design used in STEP. Accordingly, these structural missing data, unrelated to respondents' literacy skills, were ignored when calculating respondent proficiencies.
2) Omitted responses: Missing responses that occurred when respondents chose not to perform one or more presented items, either because they were unable or for some other reason. Any missing response followed by a valid response (whether correct or incorrect) was defined as an omitted response. Omitted responses were treated as wrong, because a random response to an open-ended item would almost certainly result in a wrong answer.
3) Not reached or not attempted responses: Missing responses at the end of a booklet were treated as if they were not presented due to the difficulty of determining if the respondent was unable to finish these items or simply abandoned them.

Cases where respondents did not answer a sufficient number of background questionnaire (BQ) questions (< 5 items) were considered as incomplete and not used in the latent regression. They also were not included in computing plausible values.

For respondents who answered a sufficient number of BQ questions but may not have been able to respond to the cognitive items or were unwilling to do so, the interviewers were required to document the extent to which the background questions and cognitive items were answered and to ascertain the reason for missing responses. These reasons could be categorized as:

1) Nonresponse due to refusal to participate, thus unrelated to literacy skills
2) Unable to respond due to a language difficulty or cognitive skill-related disability, thus indicating a deficiency of literacy skills
3) Inability to provide a written response due to a physical disability
4) Other unspecified reasons

Only the missing responses of nonrespondents in the second category were imputed as incorrect. The rest of the missing responses were considered unrelated to cognitive skills and thus ignored.

Respondents who correctly solved fewer than 3 of the 8 core items (administered after the BQ and before the cognitive assessment) were not required to continue with an additional task booklet of cognitive items; their missing responses were considered incorrect for the proficiency estimation. This decision was based on the findings in the PIAAC 2012 Field Test, which showed that respondents who correctly answered fewer than 3 were not likely to provide a correct answer to more than 8% of items.

Data completion – treatment of respondents with fewer than 5 cognitive item responses

A separate issue involved respondents who provided background information but did not completely respond to the cognitive items. A minimum of 5 completed items per domain was necessary to assure sufficient information about the proficiency of respondents.

In some cases a sampled individual decided to stop the assessment. The reasons for stopping could be classified into two groups: those unable to respond to the cognitive items (i.e., for cognitive-related reasons) and those unwilling to respond (i.e., for noncognitive-related reasons).

STEP followed the PIAAC procedure with respect to cases with responses to fewer than 5 cognitive items per domain. All consecutively missing responses at the end of a block of items were treated as incorrect if the reason for not responding to the cognitive items was related to literacy skills. Otherwise, all consecutively missing responses were treated as "not reached" and coded with the value 9.

This scoring method is important with regard to the latent regression population model. The treatment of nonresponding examinees due to noncognitive-related reasons has no impact on the likelihood function of proficiency. But there is an impact associated with the treatment for nonresponding cases due to cognitive-related reasons. With this scoring procedure, summary statistics can be produced for the entire population, including those who responded to cognitive items correctly in various degrees, as well as those who were not able to respond to cognitive items.

In general, the data were prepared as described above and the evaluation showed that the data were reliable in most cases; for cases where this was not true, the data were fixed and cleaned to be used for subsequent analyses (item analyses, scoring reliability, IRT scaling, population modeling).

Item analyses

Once the data were prepared, item analyses were conducted separately for each country for the literacy and reading component items. The purpose of the item analyses was to identify outliers or unexpected patterns that might signify issues with translations of items or scoring guides, or issues related to a misinterpretation of scoring guides during scoring training. ETS provided the World Bank with an item analysis report including the following statistics for each item:

Summary Statistics:

- Statistics for the computation of the alpha reliability coefficient and standard error of measurement for the test.
- Summary statistics for the block scores for the literacy Exercise Booklets. The block score is the sum of correct responses for each respondent.
- Summary statistics for the criterion score across all subjects

Response Categories (Columns) of Each Item within the Block:

NOT RCH	Subjects who did not respond to or omitted the question and did not respond to any subsequent question.
OMIT	Subjects who did not respond to or omitted the question but did respond to at least one subsequent question in the block.
1*	Subjects who responded correctly.
7	Subjects who responded incorrectly. (Note: In the IRT scaling these responses were coded as 0.)
TOTAL	The aggregation of subjects who either omitted the item or had valid response codes. These statistics do not include the subjects who did not reach the item.

Item Statistics

R BIS	The Rbis (R-biserial) indicates the correlation between students' performance on an individual question and their performance on the criterion score. It is a measure of a question's power to discriminate among students of different abilities. A relatively high R-biserial indicates that students who scored higher on the criterion score were more likely than students who scored lower to get that individual question correct. The r biserial estimates the product moment correlation that would be obtained from two continuous distributions if the dichotomized variable were normally distributed. In special cases it can take on a value greater than 1, and it is actually unbounded in both directions.
PT BIS	The point biserial is the Pearson product moment correlation coefficient between the dichotomous item score (0, 1) and the continuous criterion score. Its range is (-1, 1).
P+	P+ is the percent of students who reached the question and selected the correct answer.
Delta	Delta index is the inverse-normal transformation of proportions correct to describe item difficulty with the mean of 13.0 and the standard deviation of 4. Smaller delta index indicates easier item and larger number indicates difficult item. The index can vary often between 1 and 25.

Scoring reliability

Accurate and reliable scoring of items, especially for open-ended items coded by human scorers, are key components of quality control and are necessary for ensuring valid assessment results. The scales on which the statistical framework of PIAAC and STEP were built are only as good as the scores that constitute them. Items were scored to classify responses into pre-defined response categories and to determine whether respondents answered the questions correctly. Some items were double scored as a measure of quality assurance to determine whether the scoring rubrics were being applied consistently across scorers. Comparing results among scorers as well as across countries gives information whether the application of scoring rules by all personnel responsible for human scoring of item responses both within and across countries are consistent.

Cohen's Kappa (Cohen, 1960) is a statistical measure of scorer agreement analyses that is used in many applications of scoring reliability studies. Kappa provides a correction for agreement by chance. Kappa, as well as the percent of direct scoring agreement, was calculated for every item and over all items per country. Moreover, scoring reliability was computed within and across countries. This information was used to evaluate scoring procedures, scorers, the scoring guides, and items based on the consistency of scoring. Scoring issues might be the underlying reason for item misfit or country-by-item interactions in the IRT scaling and were identified and fixed before scaling wherever possible.

Results in STEP showed high scoring reliability in most cases, so scoring could be assumed as accurate and reliable, and data could be used for the further IRT scaling and population modeling. The scorer reliability file provided to the World Bank includes within-country and across-country (anchor scoring) Kappa coefficients and percent of scorer agreement per item.

IRT scaling: Estimation of item parameters

The IRT scaling was the first step of the population modeling and provided the estimations of item parameters and the proficiency distribution of the population. The latter was then used to calculate

a posteriori distribution together with the BQ variables using latent regressions. From this posteriori distribution, plausible values (which are multiple imputations) were obtained to provide a more accurate and reliable proficiency estimation than the proficiency estimation of the IRT scaling alone. Similar to PIAAC, STEP used the two-parameter logistic model (2PL; Birnbaum, 1968) for dichotomously scored responses. For more details about the models and IRT scaling process see the PIAAC technical report (OECD, 2013, ch. 17, pp. 2–6).

From the *cognitive assessment,* there were 44 literacy items as well as 81 items measuring reading component skills in word recognition (vocabulary), sentence processing, and passage comprehension. While the IRT scaling was conducted for the literacy items, the reading component items were not included (only item analyses were conducted for the reading components).

The *2PL model* is a mathematical model for the probability that an individual will respond correctly to a particular item from a single domain of items. The probability of solving an item (i) depends only on the ability or proficiency (θ_j) of the respondent (j) and two item parameters characterizing the properties of the item (item difficulty β_i and item discrimination α_i).

A central assumption of IRT is conditional independence (sometimes also called local independence). In other words, item response probabilities depend only on the respondent's ability and the specified item parameters—there is no dependence on any demographic characteristics of the examinees, or responses to any other items presented in a test, or the survey administration conditions. Moreover, the 2PL model assumes unidimensionality, that is, a single latent variable, the ability or proficiency θ, accounts for performance on a set of items.

As STEP used the literacy items from PIAAC with the aim to provide a link between these two surveys, it was analyzed whether the PIAAC items did, in fact, work similarly in STEP. For this, the item parameters in the IRT scaling of the STEP data were fixed to the values of the item parameters obtained in PIAAC (fixed item parameter linking). It was assumed that the common data (including the data from all participating countries) were comparable for all items in the assessment. The IRT scaling and population modeling was done for 8 countries in 3 subsequent waves based on when the data were provided by the countries. The first wave of data included Bolivia, Colombia, and Vietnam; the second wave included Armenia, Georgia, and Ghana; the third wave included Ukraine and Kenya. For a more reliable estimation of parameters in the IRT scaling, it was taken into account whether a respondent failed or passed the core items (respondents who passed the core items received one of four booklets including the literacy items. Therefore, each country was divided into two groups (failed versus passed core) and parameters were estimated for each group separately using equality constraints at the beginning and allowing for unique item parameters in subsequent steps of the analyses based on item fit statistics (root mean square deviation, or RMSD, and mean deviation, or MD; see below). Table 2.2 gives an overview of the participating countries and sample sizes.

Table 2.2: Sample sizes of participating countries in STEP for the domain literacy

Country	Country Code in Data File	Core	n per Booklet Assignment	n per Country
Wave 1				
1 Bolivia	BOL	passed	1,937	2,343
		failed	406	
2 Columbia	COL	passed	2,420	2,560
		failed	140	
3 Vietnam	VNM	passed	3,100	
		failed	228	3,328
Wave 2				
4 Armenia	ARM	passed	2,717	2,737
		failed	20	
5 Georgia	GEO	passed	2,542	2,643
		failed	101	
6 Ghana	GHA	passed	1,194	2,179
		failed	985	
Wave 3				
7 Ukraine	UKR	passed	2,349	2,362
		failed	13	
8 Kenya	KEN	passed	2,621	3,200
		failed	579	

Standardized sample weights were used in the STEP IRT scaling. Items in the scaling that showed deviations from the PIAAC item parameters were assumed to work differently in STEP than in PIAAC, meaning they would harm the link to PIAAC. To examine deviations in the IRT scaling, so-called item fit statistics were used to test the fit of the model for each single item. Like PIAAC, STEP used the RMSD and the MD. With these item fit statistics, it was examined whether the item characteristic curves (ICC)—which illustrate the relationship between the respondent's ability and the item parameters—for each single item within a country found in the empirical data showed deviations from the expected ICC of the model for this item.

Both the RMSD and the MD are measures to quantify the magnitude and direction of the shift of the observed data from the estimated ICC for each single item. While MD measure is most sensitive to the deviations of observed item difficulty parameters from the estimated ICC, the RMSD measure is sensitive to the deviations of both the observed item difficulty parameters and item slope parameters.

Poorly fitting items or item characteristic curves in STEP were generally revealed using a RMSD > 0.15 criterion as well as an MD > 0.15 and < -0.15 criterion where a value of 0 indicates no discrepancy (in other words, a perfect fit of the model); for Ghana and Kenya, a less strict criterion of RMSD > 0.20 as well as an MD > 0.20 and < -0.20 was used. For such items, it was assumed that the common item parameters from PIAAC were not appropriate (common meaning that item

parameters were equal to all or most countries of an assessment) and country-specific unique item parameters were estimated in a second step (unique meaning that item parameters were unique for one country or a small group of countries).

In this subsequent step, unique item parameters were estimated in order to account for national deviations for a small subset of items. This involved a close monitoring of the IRT scaling for item-by-country interactions and allowing country-specific item parameters only in instances where substantial deviations were identified. This procedure takes into account that some items work differently in certain countries due to language or cultural differences or due to translation issues. The common and unique item parameters were estimated using a mathematical algorithm[3] that still allowed us to estimate all item parameters in relation to one another and, thus, common and unique item parameters are on the same latent scale. As long as only a few item parameters are unique, the link to PIAAC is not harmed. Thus, STEP (like PIAAC) allowed for different sets of item parameters to improve model fit and optimize the comparability of countries.

The scaling procedure also needed to take into account the possibility of any systematic interaction between the samples and the items that were used to produce estimates of the item parameters and sample distributions. For this reason, the 2PL model was estimated as a multiple-group IRT model using a mixture of normal population distributions (one for each sample) where item parameters were generally constrained to be equal across countries with a unique mean and variance for each country. The moments of these distributions were updated at each iteration during IRT calibration.

In the STEP analyses in most cases, the item responses across countries were accurately described by the common PIAAC item parameters. The deviation pattern of these items from the common PIAAC item parameters was not consistent for any one particular country. Table 2.3 provides the number of unique item parameters per language group (see appendix for detailed information about group-item interactions).

In wave 1, 12.8% unique (country-specific) item parameters were estimated due to item-by-group interactions. For the first part of wave 2, 12.5% unique parameters were estimated, and for the second part of wave 2, 9.7% unique item parameters had to be used.

[3] The software *mdltm* (von Davier, 2005) was used for the IRT calibration, which provides marginal maximum likelihood estimates obtained using customary expectation-maximization methods, with optional acceleration.

Table 2.3: Number of unique item parameters for each country for the scale literacy

Country	Data Wave	Number of Country-Specific Item Parameters Literacy (44 items)
Bolivia	1	7
Colombia	1	8
Vietnam	1	7
Armenia	2-pt.1	7
Georgia	2-pt.1	13
Ghana	2-pt.1	7
Ukraine	2-pt.2	4
Kenya	2-pt.2	12

Given this estimation and optimization approach, no item was dropped from the analysis in STEP. The final item parameters were then used to estimate the respondents' proficiency needed for the population model.

The following section describes how the proficiency distribution obtained from the IRT scaling was used to build the population model, with BQ variables integrated as well, with the aim of providing plausible values.

Population modeling using BQ variables

Most tests that measure cognitive skills are concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. The accuracy of these measurements can be improved, meaning a reduction in the amount of measurement error, by increasing the number of items administered to the individual. Thus, it is common for achievement tests to be composed of more than 70 items.

In international large-scale assessments such as PIAAC or STEP, test forms are kept relatively short to minimize individuals' response burden. At the same time, the aim is to achieve broad coverage of the tested constructs. The full set of items is organized into different but linked assessment booklets; each individual receives only one booklet. Thus, the survey solicits relatively few responses from each respondent while maintaining a wide range of content representation when responses are aggregated. The advantage of estimating population characteristics more efficiently is offset by the inability to reliably measure and make precise statements about individuals' performance.

By applying the population model, the problem of low reliability or uncertainty can be overcome by providing plausible values. The population model, a latent regression item response model, combines the proficiency estimates from the IRT scaling with additional information of the

respondents from the BQ variables with the aim of providing plausible values as a more accurate and reliable estimation of the literacy proficiency distribution for a group of respondents. The "bridge" between the proficiency estimate from the IRT scaling and the BQ variables is a latent regression model (latent regressions).

In the latent regression model, the distribution of the proficiency variable (θ) is assumed to depend not only on the cognitive item responses X but also on a number of predictors Y, which are variables obtained from the BQ (e.g., gender, country of birth, education, occupation, employment status, reading practices, etc.). Both the item parameters from the calibration stage and the estimates from the regression analysis are needed to generate plausible values.

In STEP (as in PIAAC), a considerable number of background variables (predictors) were collected, including demographic information, educational experiences, occupational experiences and skill use, among others. All variables in the BQ were contrast coded before they were processed further in the population model. Contrast coding allows the inclusion of codes for refused responses as well as codes for responses that were not collected by means of routing and avoiding the necessity of linear coding. The increased number of variables obtained through contrast coding is substantial. To capture most of the common variance in the contrast-coded background questions with a reduced set of variables, a principal component analysis was conducted. Because each population can have unique associations among the background variables, a single set of principal components was not sufficient for all countries included in STEP. Therefore, the extraction of principal components was carried out separately by country. Each set of principal components or conditioning variables (y^C) was selected to include 80% of the variance, with the aim of explaining as much variance as possible while at the same time avoiding overparameterization. The use of principal components also serves to retain information for examinees with missing responses to one or more background variables.

After the principal components were obtained, a regression of these variables on the proficiency variable was calculated (in STEP, only parameters from the group that passed core items were used to estimate the proficiency variable for the domain literacy). Thereby, the latent regression parameters were estimated conditional on the item parameter estimates determined in the IRT scaling. The latent regression model was set to maximize the variance explained by the conditioning variables and to minimize measurement error at the same time. Through this latent regression IRT model or population model, a posterior distribution was obtained that provided a combination of IRT proficiency estimates and BQ variables (as principal components). In a final step, 10 plausible values for each respondent *j* were drawn from this conditional posteriori distribution.[4] These plausible values were included in the STEP country datasets.

[4] The software DGROUP (Rogers et al., 2010) was used to estimate the latent regression model using an expectation maximization (EM) algorithm (cf. Mislevy, 1985) and generate plausible values. A multidimensional variant of the latent regression model was used that is based on Laplace approximation (Thomas, 1993).

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997.

Organisation of Economic Co-operation and Development (2013). *Technical report of the Survey of Adult Skills (PIAAC)*. Retrieved from http://www.oecd.org/site/piaac/publications.htm.

Rogers, A., Tang, C., Lin, M.-J., & Kandathil, M. (2006). DGROUP (computer software). Princeton, NJ: Educational Testing Service.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309–322.

von Davier, M. (2005). *A general diagnostic model applied to language testing data.* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

Appendix

Table A1: Literacy items per wave and language group that received group-specific item parameters in the IRT scaling

	Wave 1						Wave 2						Wave 3			
STEP Item-ID	**BOL – passed-core**	**BOL – failed-core**	**COL – passed-core**	**COL – failed core**	**VNM – passed-core**	**VNM – failed core**	**ARM - passed-core**	**ARM - failed-core**	**GEO - passed-core**	**GEO - failed-core**	**GHA - passed-core**	**GHA - failed-core**	**UKR - passed-core**	**UKR - failed-core**	**KEN- passed-core**	**KEN - failed-core**
M301C05	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M300C02	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N000C01	*	*	*	*	*	*	*	*	*	*	**O**	*	*	*	*	*
P330001	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
P330002	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N302C02	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N311701	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N311703	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N307401	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N307402	*	*	**X**	*	*	*	*	*	*	*	**O**	*	*	*	**W**	*
M313410	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*
M313411	*	*	*	*	*	*	**O**	*	*	*	*	*	**W**	*	**W**	**W**
M313412	*	*	*	*	*	*	*	*	*	*	**O**	*	*	*	*	*
M313413	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*
M313414	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N306110	**X**	*	*	*	*	*	*	*	*	*	**O**	*	*	*	*	*
N306111	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	**W**	*
M000423	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M000425	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*

Table A1 (continued)

	Wave 1						Wave 2						Wave 3			
STEP Item-ID	**BOL – passed-core**	**BOL – failed-core**	**COL – passed-core**	**COL – failed core**	**VNM – passed-core**	**VNM – failed core**	**ARM - passed-core**	**ARM - failed-core**	**GEO - passed-core**	**GEO - failed-core**	**GHA - passed-core**	**GHA - failed-core**	**UKR - passed-core**	**UKR - failed-core**	**KEN- passed-core**	**KEN - failed-core**
M312315	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M312316	*	*	*	*	*	*	**O**	*	**V**	*	*	*	*	*	*	*
M312318	*	*	*	*	*	*	*	*	**O**	*	*	*	*	*	*	*
M000124	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*
M000126	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*
P324002	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*	**Z**	*
P324003	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*
M310406	**X**	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*
M310407	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N314101	**X**	**X**	**X**	**X**	**X**	**X**	**O**	**O**	**O**	**O**	**O**	**O**	*	*	*	*
N314102	*	*	*	*	**X**	*	*	*	**O**	*	*	*	*	*	*	*
M308116	**X**	**X**	**X**	**X**	**X**	**X**	*	*	**O**	*	*	*	*	*	*	*
M308117	**X**	**X**	**X**	**X**	**X**	**X**	**O**	*	**O**	*	**V**	*	*	*	*	*
M308118	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	**W**	*
M308119	**X**	**X**	**X**	**X**	**X**	**X**	**O**	*	**O**	*	**V**	*	**W**	*	**Z**	*
M308120	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M309319	*	*	*	*	*	*	**O**	*	**O**	*	*	*	*	*	*	*
M309320	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M309321	**X**	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*
M309322	*	*	*	*	*	*	**O**	*	**O**	*	*	*	*	*	*	*
N000502	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**W**	*

Table A1 (continued)

	Wave 1						Wave 2						Wave 3			
STEP Item-ID	**BOL – passed-core**	**BOL – failed-core**	**COL – passed-core**	**COL – failed core**	**VNM – passed-core**	**VNM – failed core**	**ARM - passed-core**	**ARM - failed-core**	**GEO - passed-core**	**GEO - failed-core**	**GHA - passed-core**	**GHA - failed-core**	**UKR - passed-core**	**UKR - failed-core**	**KEN- passed-core**	**KEN - failed-core**
M305215	*	*	*	*	*	*	*	*	**O**	*	*	*	*	*	*	*
M305218	*	*	*	*	*	*	*	*	**O**	*	*	*	**W**	*	*	*
M303101	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M303102	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

Note: * denotes common item parameters; **X, O, V, W,** and **Z** denote country-specific item parameters; identical symbols/letters in the same row (or for the same item) for different countries (columns) denote identical item parameters for the specific item in these countries (identical symbols/letters in different rows/items do not).