Education & Skills Online
Technical Documentation

Published: December 2015

Update 1: October 2016

Update 2: June 2018

Update 3: June 2020

You may cite this document as:

Organisation for Economic Co-operation and Development. (2018). *Education & Skills Online technical documentation*. Paris: Author. http://www.oecd.org/skills/ESonline-assessment/assessmentdesign/technicaldocumentation/

Table of Contents

Chapter 1: Education & Skills Online Overview and Workflow

Education & Skills Online is an assessment tool designed to provide individual level results for literacy, numeracy, problem solving in technology-rich environments (PSTRE), and reading components (basic reading skills) measures that can be used to compare the test taker's results with the those of others both within the test taker's country and internationally. The assessment includes a background questionnaire to collect information on the test taker's age, gender, education level, employment status, and native country and language. This tool also includes noncognitive assessments that measure skill utilization, workforce readiness, career interests, and health indicators.

The cognitive instruments in Education & Skills Online provide information about the skills and knowledge of individuals as assessed in the Organisation for Economic Co-operation and Development's Survey of Adult Skills, which is part of the Programme for the International Assessment of Adult Competencies (PIAAC). PIAAC, a large-scale household assessment, was designed to provide policy-level information about the skills of adults ages 16 to 65. In total, 32 countries participated in the first two rounds of the PIAAC survey from 2008 to 2016. As noted in the PIAAC Technical Report, this computer-based, large-scale assessment of adult skills differed from earlier adult assessments in several important ways. For example, PIAAC was able to address literacy in digital environments by including tasks that required respondents to use electronic texts such as Web pages, emails, and discussion boards. The new domain of PSTRE included computer-based simulation tasks that focused on the cognitive skills required to access and make use of computer-based information to solve problems. The reading components domain, which included measures of vocabulary knowledge, sentence processing, and passage comprehension, provided more information about the skills of individuals with low levels of literacy proficiency than had been available from previous international assessments.

1.1 Cognitive modules

The Education & Skills Online cognitive modules focus on the same domains of literacy, numeracy, PSTRE, and reading components as PIAAC but with an emphasis on providing information about skills as they relate to education and workforce contexts, including postsecondary education or training and workforce readiness. The items in this test reflect the domain definitions and frameworks developed for PIAAC[1] and were designed to provide

[1] These frameworks are available at http://www.oecd.org/site/piaac/publications.htm.

information along PIAAC's described proficiency scales, which capture the progression of task complexity and difficulty for each domain.

Education & Skills Online was developed and validated for a population between ages 16 and 65. It is appropriate for students or out-of-school youth who are interested in transitioning to postsecondary education/training or into the workforce. It is also appropriate for adults of various ages who wish to reenter an educational or training environment or demonstrate their workforce readiness skills. Education & Skills Online can also be used to assess the human capital of enterprises and other entities.

1.1.1 Adaptive testing

Education & Skills Online uses adaptive algorithms for the core literacy and numeracy assessment, optimizing the delivery of test items based on estimated proficiency levels of individuals. The result is more reliable information about skills in a relatively short period of time. Adaptive tests can be roughly categorized as belonging to one of two groups: item-level adaptive tests and multistage adaptive tests. Item-level adaptive tests traditionally have been referred to as computer-adaptive tests (CATs). The multistage adaptive design used in Education & Skills Online is, in a sense, an extension of a CAT in that the CAT algorithm "decides" on the choice of the next item after each response, whereas multistage algorithms allow the choice of the next cluster of items either after one or multiple responses. This provides more information and the opportunity for greater accuracy in the decision of the choice of the difficulty level of the next cluster of items. Using item clusters instead of individual items for adaptive decisions reduces the likely dependence of the stage adaptive selection on item-by-country interactions compared to the effects expected when using item-level adaptive tests.

1.2 Noncognitive modules

Education & Skills Online contains three optional noncognitive modules (a fourth noncognitive module, Behavioral Competencies, was only available from March 2018 to June 2020). They include skill use scales drawn from PIAAC, as well as modules covering noneconomic outcomes and vocational orientation that research has shown to be important for building and maintaining skills among adults and that are of high interest to policy makers and educators.

The noncognitive modules are:

- Skill Use: Utilizes items from PIAAC to assess the specific skills that respondents use in both their work and daily lives as important drivers of skill acquisition as well as critical outcomes affecting their lives. Questions focus on skills associated with reading, writing, use of mathematical information and ideas, and information and communications technology (ICT).

- Career Interest and Intentionality: Measures an individual's preferences for different types of work activities and environments and the level of an individual's intention to seek out new job opportunities and career- and job-related training.

- Subjective Well-Being and Health: Examines the main components of subjective well-being—life evaluation and positive and negative affect—in addition to subjective health and well-being indicators.

- Behavioral Competencies (Available March 2018 to June 2020): Measures selected personality facets based on high relevance and utility for academic and workforce readiness and success.

1.3 Product packages

Education & Skills Online is provided in three different packages to allow organizations purchasing the test to choose the set of modules that best meets their needs. The three package types are described in Table 1.1.

Table 1.1: Package types

Package	Includes
Core assessment package	• Background questionnaire • Literacy and numeracy assessment • PSTRE and reading components assessments
Noncognitive assessment package*	• Background questionnaire • Three noncognitive modules: Skill Use, Career Interest and Intentionality, and Subjective Well-Being and Health
Bundled package (core and noncognitive modules)*	• Background questionnaire • Literacy and numeracy assessment • PSTRE and reading components assessments • Four noncognitive modules: Skill Use, Career Interest and Intentionality, and Subjective Well-Being and Health

*These packages also included the Behavioral Competencies noncognitive module from March 2018 to June 2020.

The test is available in 10 languages (16 country-specific language versions overall):

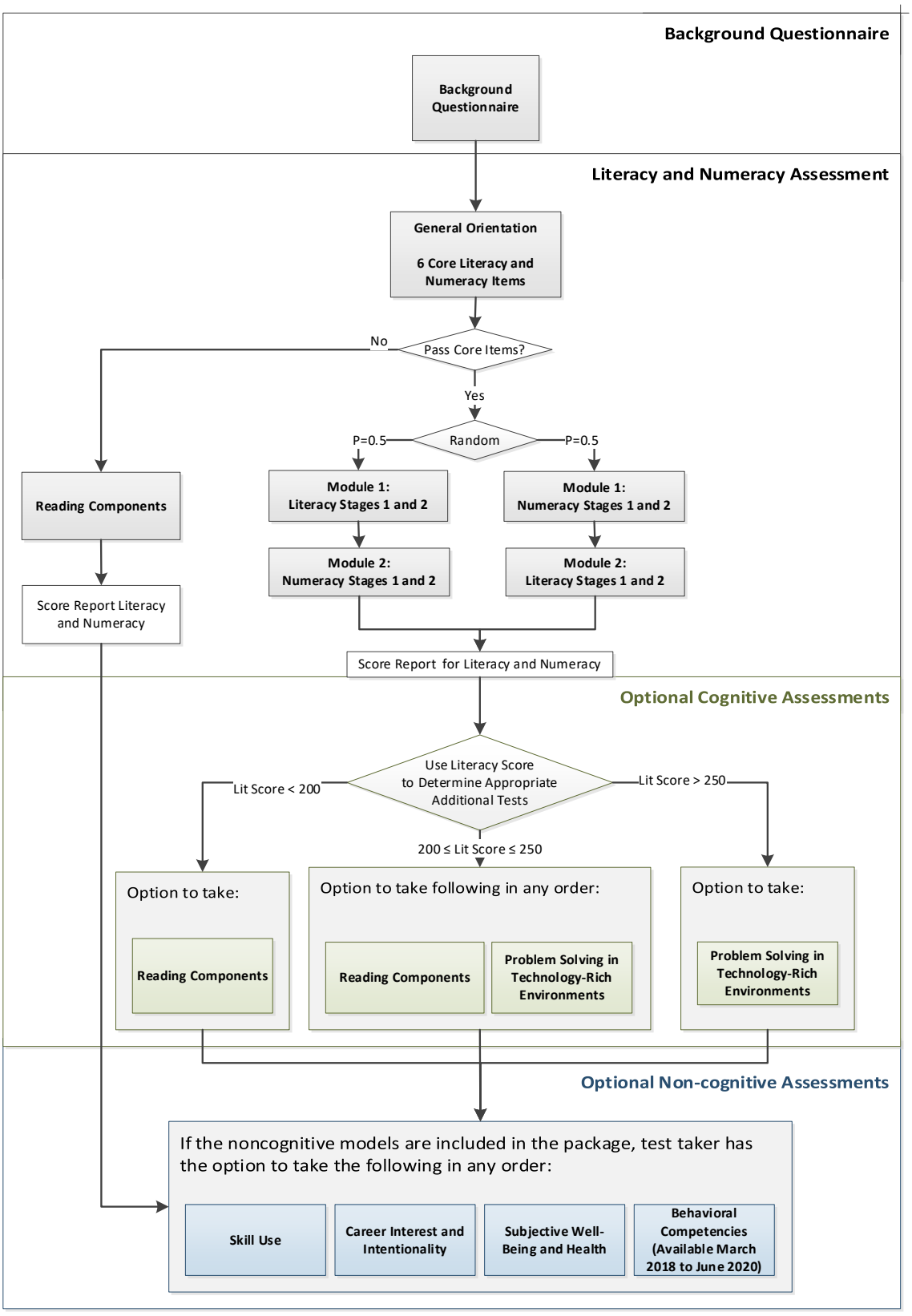
- Czech
- English (Australia, Canada, Ireland, and U.S. versions)
- Estonian
- French (Canada)
- Italian

- Japanese
- Russian (Estonia and Russian Federation versions)
- Slovak
- Slovenian
- Spanish (Chile, Spain, and U.S. versions)

1. 4 Workflow

Education & Skills Online includes three components: a core cognitive assessment of literacy and numeracy, the optional cognitive assessments of PSTRE and reading components, and the optional noncognitive assessments. Optional assessments are ones that are included in the test package; however, because these assessments do not automatically begin the test, the purchaser may direct test takers to skip these assessments. Figure 1.1 illustrates the workflow of the Education & Skills Online assessment, identifying both the core and optional assessments.

Figure 1.1: Education & Skills Online bundled package workflow



1.4.1 Core and optional cognitive modules workflow

1.4.1.1 Core package workflow

All test takers log into the Education & Skills Online test using an authorization code provided by the test purchaser. Education & Skills Online does not collect any personal information such as names, addresses, phone numbers, or email addresses. Any personally-identifying information linking individuals to authorization codes must be maintained by purchasers outside of the Education & Skills Online system. Test takers begin the test by answering background questions.

After completing the background questions, users with a core package or bundled package test code take three literacy and three numeracy items. Test takers with very low literacy scores on these core cognitive items are routed directly to the reading components module. All other test takers continue to the literacy and numeracy assessments.

Choice of first module: The selection of a domain (literacy or numeracy) for the first module is random; test takers have an equal chance of starting with either literacy or numeracy. Each module contains two item clusters, or testlets.

Choice of stage 1 testlet within literacy and numeracy modules: The literacy and numeracy testlets in Stage 1 vary in difficulty. There are three levels of testlets: easiest (Testlet 1), medium (Testlet 2), and most difficult (Testlet 3). Five responses from the background questionnaire determine which testlet is chosen first for a test taker. Item response theory (IRT) scaling performed for PIAAC makes it possible to estimate the relationship between proficiency and key background variables and use that relationship to inform the adaptive algorithm. The background questionnaire variables used are:

- Age
- Education level
- Country of birth (either the country of the test or another country)
- Employment status

Choice of stage 2 testlet for literacy and numeracy: The three literacy and numeracy testlets in Stage 2 also vary in difficulty, ranging from Testlet 1 as the easiest to Testlet 3 as the most difficult. The testlet assignment for Stage 2 depends on the same background variables as in Stage 1 and the test taker's score on the Stage 1 testlet and the core items.

Choice of second module: After completing Module 1 (the two testlets for either literacy or numeracy), the test taker proceeds to Module 2. If the test taker completed literacy as Module 1, then he or she will receive numeracy as Module 2. If the test taker received numeracy as Module 1, then he or she will receive literacy as Module 2. The selection of testlets for Module 2 uses the same variables and process as were used in Module 1.

Score report: The purchaser of the test decides whether a score report, which explains assessment results, is available for the test taker at the end of the test or only to the test administrator in the online test management system. If the option is selected to make it available

to the test taker, after completing the literacy and numeracy modules individuals receive the score report, which explains their literacy and numeracy results. The score report includes the test taker's scale score; an explanation of the skills that were measured based on PIAAC proficiency level standards and how those skills are used in everyday life; and a characterization of the test taker's strengths and weaknesses in the skill areas assessed. In addition, the score report compares the test taker's score to the national and international data from PIAAC by education level, occupation, and age group. Test takers who were routed directly to the reading components test will receive their reading components score report and their literacy and numeracy score report after completing the reading components assessment. Their literacy and numeracy score report will be based on their responses to the six core items they completed.

Choice of core optional module(s): A test taker's literacy score will be used to determine which optional core cognitive modules are appropriate for that individual to take. Individuals with literacy scores below 200 will be given the option to take reading components; individuals with literacy scores between 200 and 250 will be given the option to take both reading components and PSTRE, and individuals with literacy scores above 250 will only be given the option to take PSTRE. After each module, if the test purchaser has selected the option, the test taker will see a score report.

1.4.1.2 Bundled core cognitive and noncognitive package workflow

If the organization purchased the bundle that includes the noncognitive modules, a test taker will have the option to take all three of the noncognitive modules in any order after having completed the core literacy and numeracy modules. If the core optional modules are included as part of the package, the test taker will have the option to take the appropriate core optional module(s) and all three noncognitive modules in any order after having completed the literacy and numeracy modules. At the completion of each module, if the test purchaser has selected the option, the test taker will receive a score report before being returned to the test site, where he or she may start a different module.

1.4.1.3 Noncognitive-only package workflow

If the organization purchased the noncognitive modules only, when a test taker first logs into the test administration site, he or she will be prompted to take the core background questions. The core background questions provide demographic information on the test taker and reduce the need to repeat questions about employment or education status within each module. Following the background questions, the test taker will have the option to take the three noncognitive modules in any order. Following the completion of each module, if the test purchaser has selected the option, the test taker will receive a score report before being returned to the start of the test site, where he or she may start a different module.

1.4.2 Timing

The cognitive portion of the test—including the background questionnaire, literacy and numeracy tests, and reading components or PSTRE modules—takes approximately 95 minutes to complete. The noncognitive portion of the test takes approximately 20 minutes. The bundled package, which includes the core cognitive tests and the noncognitive modules, takes approximately 115 minutes. A breakdown of the average time to complete each portion of the assessment package is included in Table 1.2.

Table 1.2: Average Time to Complete a Test Package

Package	Assessment	Time
Core Cognitive Package	Background Questionnaire	5 minutes
	Core Literacy and Numeracy Assessment	60 minutes
	Reading Components and/or Problem Solving in Technology-Rich Environments	30 minutes
	Total:	**95 minutes**
Noncognitive Package*	Background Questionnaire	5 minutes
	Career Interest and Intentionality	10 minutes
	Skill Use	5 minutes
	Subjective Well-Being and Health	5 minutes
	Total:	**25 minutes**
Core and Noncognitive Bundled Package*	Background Questionnaire	5 minutes
	Core Literacy and Numeracy Assessment	60 minutes
	Reading Components and/or Problem Solving in Technology-Rich Environments	30 minutes
	Career Interest and Intentionality	10 minutes
	Skill Use	5 minutes
	Subjective Well-Being and Health	5 minutes
	Total:	**115 minutes**

*From March 2018 to June 2020 these two packages also included the Behavioral Competencies module which took approximately 15 minutes to complete.

Chapter 2: Development of the Cognitive Instruments

The cognitive items in Education & Skills Online include existing PIAAC (Programme for the International Assessment of Adult Competencies) items as well as newly developed items for the domains of literacy and numeracy. The inclusion of existing PIAAC items was required so the Education & Skills Online and PIAAC results could be linked and comparable scales could be established across the two measures. The newly developed items were field tested in Spring 2013 and 2017 to determine their reliability for the Education & Skills Online language versions. A description of the Field Test process is included in Chapter 5.

2.1 Selection of PIAAC linking items

The instrument for Education & Skills Online includes all of the reading components items and a subset of the problem solving in technology-rich environments (PSTRE) items that were included in the PIAAC Main Study. Because new items needed to be developed for Education & Skills Online, only a subset of the PIAAC Main Study literacy and numeracy items was selected for the final instrument. The selection of these linking literacy and numeracy items was based on:

- the psychometric characteristics of the items,
- an emphasis on including a broader range of difficult items to reflect the workplace focus that is part of Education & Skills Online, and
- a goal of ensuring the intended representation of the domain frameworks as defined in PIAAC.

2.2 Development of new items

To better ensure consistency across the two instruments, the new items for Education & Skills Online were written by assessment developers who had worked on PIAAC. Additionally, items were reviewed by members of the PIAAC Literacy and Numeracy Expert Groups who were familiar with the PIAAC frameworks and the characteristics of items developed for that survey.[1] Assessment developers reviewed all new items to ensure consistency in instructions, response modes, and presentation across domains. In the 2013 Field Test, the new items were centrally translated and then verified by the participating countries. Comments from national teams were

[1] For a complete list of experts in these groups, please see the PIAAC Technical Report, Appendix 6 at http://www.oecd.org/site/piaac/Technical%20Report_Part%206.pdf.

solicited and revisions made as necessary. In the 2017 Field Test, the participating countries translated the new items and the translations were verified by an international contractor. Revisions were made as necessary.

As with PIAAC, a requirement for new items was that they be could be scored automatically by the online test system. This was a necessary feature in order to implement adaptive testing in Education & Skills Online. Developers therefore used response modes that were employed in PIAAC, which included:

- clicking items where respondents were asked to click on graphical elements, cells in a table, links on a Web page, radio buttons, or check boxes;
- numeric entry items where respondents provided answers using the number keys, decimal point (period or comma as appropriate across participating countries), and space bar;
- selection items where respondents indicated an answer using a drop-down menu; and
- highlighting items where respondents could highlight one or more words, phrases, or sentences in a text to answer a question.

Additionally, several new literacy items require respondents to click on a sentence in a text to provide an answer. This is a new response mode included in Education & Skills Online to avoid some of the translation challenges inherent in implementing the highlighting response mode across languages.

Stimulus materials were selected based on specifications provided in the framework for each domain. To the extent possible, stimuli were taken from, or based on, real-world materials such as newspaper and magazine articles, advertisements, books, forms, and Web pages 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 that would be appropriate across cultures and languages.

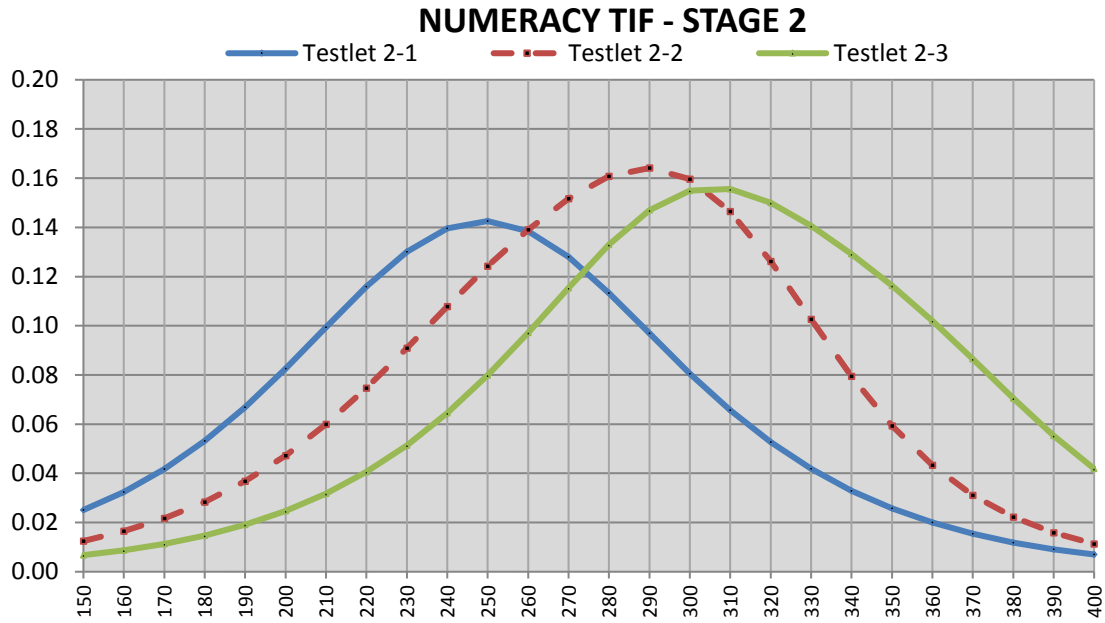
2.2.1 Moving from Field Test to final product instruments

Following analysis of the Field Test data, a number of steps were followed to develop the final product instruments.

- **Item analysis:** Items were evaluated based on their statistical performance in the Field Test, looking at performance within and across countries. The purposes of the Field Test analyses were to ensure that items were reliable, valid and comparable across countries and that the common PIAAC scale could be maintained across countries and assessments.
- **Item selection**: Based on the 2013 Field Test data, developers selected a set of final product items for each domain in September 2014. One challenge for the final product selection process was the need to fit the final set of items within the testlets that made up the adaptive design. A testlet is a cluster of items grouped according to ability levels. As shown in Figure 1, the design for the core cognitive adaptive instrument includes two stages within each domain, divided into a total of six testlets. The goal is for each of the three testlets within a stage to target a score approximately 30-40 points from the next, making the test more

efficient. To accommodate this design, developers needed to look at the difficulty level of items available for the final product and determine the appropriate testlets and blocks for the items. For literacy, the fact that items existed as units (sets of items associated with a single stimulus) posed an additional challenge, particularly in those cases where items within a unit were spread across the defined difficulty levels. Because this same step was necessary for PIAAC, the experience gained from that work helped inform the selection process for Education & Skills Online. As an example, Figure 2.1 shows the test information curve for the Numeracy Stage 2 testlets. This figure illustrates that the three testlets target scale scores of approximately 250, 290, and 310.

Figure 2.1: Test Information Curves for Numeracy Stage 2 Testlets



- **Item corrections**: Countries reviewed the set of selected items, looking for any errors in implementation, during the final national check of the final product instrument. Errors were corrected and the final version reviewed and approved.

The set of items for the final product was balanced in terms of construct representation based on the overall distributions recommendations in the PIAAC framework. A total of 78 items were selected for literacy and numeracy, with the distribution across linking and new items as shown in Table 2.1.

Table 2.1: Distribution of linking and new items by domain

Domain	Linking Items (from PIAAC)	Newly Developed Items
Literacy	20	20
Numeracy	21	17
PSTRE	9	0
Reading components	38 print vocabulary (2 from the PIAAC Field Trial) 22 sentence processing 44 passage comprehension	0

2.3 Distribution of items by domain constructs

As mentioned, the Education & Skills Online items were based on the PIAAC frameworks. Those frameworks included a recommended distribution of items based on the key constructs in each domain. The distribution for PSTRE and reading components items in Education & Skills Online matches that in the PIAAC Main Study. Tables 2.2 through 2.7 show the distribution of the final item pool for literacy and numeracy, including both trend and new items.

Literacy

Table 2.2: Distribution of Education & Skills Online literacy items by medium

	Final Item Set		PIAAC Framework Goal
	Number	Percentage	Percentage
Print-based texts	17	43	70-80
Digital texts	23	58	20-35
Total	40	100	100

Note: While the PIAAC Literacy Expert Group expressed a strong interest in including digital texts in that assessment, the distribution provided in the framework document was constrained by the number of trend items that were based on print-based texts and were necessary for linking purposes. The availability of digitally based items that had been newly developed for PIAAC as well as Education & Skills Online allowed for increasing the percentage of such digital texts.

Table 2.3: Distribution of Education & Skills Online literacy items by context

	Final Item Set		PIAAC Framework Goal
	Number	Percentage	Percentage
Work	16	40	15
Personal	11	28	40
Community	11	28	30
Education	2	5	15
Total	**40**	**100**	**100**

Table 2.4: Distribution of Education & Skills Online literacy items by task aspects

	Final Item Set		PIAAC Framework Goal
	Number	Percentage	Percentage
Access and identify	20	50	40
Integrate and interpret	14	35	45
Evaluate and reflect	6	15	15
Total	**40**	**100**	**100**

Note: The increased percentage of "evaluate and reflect" items reflects the emphasis on somewhat more difficult items in Education & Skills Online versus PIAAC.

Numeracy

Table 2.5: Distribution of Education & Skills Online numeracy items by response (process)

	Final Item Set		PIAAC Framework Goal
	Number	Percentage	Percentage
Act upon, use	21	55	50
Identify, locate, or access	6	16	10
Interpret, evaluate	11	29	40
Total	**38**	**100**	**100**

Table 2.6: Distribution of Education & Skills Online numeracy items by context

	Final Item Set		PIAAC Framework Goal
	Number	Percentage	Percentage
Everyday life	12	32	45
Work-related	13	34	23
Society and community	12	33	25
Further learning	1	1	7
Total	**38**	**100**	**100**

Table 2.7: Distribution of Education & Skills Online numeracy items by mathematical content

	Final Item Set		PIAAC Framework Goal
	Number	Percentage	Percentage
Data and chance	13	34	25
Dimension and shape	7	18	25
Pattern, relationships, and change	6	16	20
Quantity and change	12	32	30
Total	**38**	**100**	**100**

2.4 Item translation and adaptation

To support the work of countries in translating and adapting items and implementing computer-based scoring, translation and adaptation guidelines were developed for participating countries. These materials supported the linguistic quality control process. They were designed to help ensure that instruments were comparable across countries and that consistent scoring procedures were implemented.

Countries developed their own national versions of the Field Test assessment materials following the translation, adaptation, and verification processes. Layout checks were conducted by contractors and countries to identify any display issues requiring modification. Revisions were prompted by such issues as text not fitting in a table cell due to longer word lengths in certain languages. These layout issues were fixed on a case-by-case basis and submitted to countries for final review and approval.

During this period, countries were also responsible for defining and adapting the computer-based scoring for their national versions where applicable. That is, all language-dependent scoring rules —such as the highlighting area—were defined by the national centers and verified as part of the quality assurance process. Countries had an opportunity to review items again during the review of the final product prior to its release.

Chapter 3: Development of the Noncognitive Instruments

The noncognitive modules in Education & Skills Online are measures of constructs beyond cognitive ability and knowledge that research has shown to be important for building and maintaining skills among adults. The four modules were developed from existing instruments with known reliability and validity. They are Skill Use, Career Interest and Intentionality, Subjective Well-Being and Health, and Behavioral Competencies. The Behavioral Competencies module was only available from March 2018 to June 2020. An important part of the development work for Education & Skills Online was to analyze and validate the use of the various noncognitive scales across country and language versions; this was done through Field Tests in 2013 and 2017.

3.1 Noncognitive module constructs and relevant utility for Education & Skills Online

The Education & Skills Online noncognitive modules were developed from items available in the public domain as well as licensed items used by permission from third parties.[1] Below are summaries of the constructs measured by each of the noncognitive modules. The accompanying tables below describe the relevance and utility of the modules with respect to various stakeholders.

3.1.1 Skill Use

The Skill Use module utilizes items from the Programme for the International Assessment of Adult Competencies (PIAAC) to assess the specific skills that respondents use in both their work and daily lives as important drivers of skill acquisition as well as critical outcomes affecting their lives. The questions in this module focus on skills associated with reading, writing, use of mathematical information and ideas, and information and communications technology (ICT). These activities are important for building and maintaining skills in literacy, numeracy, and problem solving in technology-rich environments. This module can be linked to the PIAAC skill use scales; Field Test analyses validated the comparability of the scales across the participating countries.

[1] ETS has a fee-based licensing agreement with third-party entities to use the pre-existing intellectual property (IP) within the Career Interest and Intentionality, Subjective Well-Being and Health, and Behavioral Competencies noncognitive modules in Education & Skills Online. ETS does not have the right to grant permission to any other entity for the use of this preexisting, third-party IP.

Table 3.1: Utility of the Skill Use module

Stakeholder	Relevance	Information Included in Score Report
Employers (businesses and trade associations)	• Recruit nontraditional and returning workers	• Skill use indices to identify which skills individuals use regularly in their personal and work lives
Academic systems	• Engage and develop work readiness	• Skill use indices to identify which skills individuals use regularly in their work and personal lives • Skill use indices help identify potential barriers to participation in the workforce
Policymakers	• Research	• Skill use indices to identify which skills individuals use regularly in their work and personal lives • Skill use indices help identify potential barriers to participation in the workforce and society
Individuals	• Job preparation and search	• Skill use indices to identify which skills individuals use regularly in their work and personal lives and identify areas for learning • Career-choice matching based on skill inventory • Skill use indices help identify potential barriers to participation in the workforce

3.1.2 Career Interest and Intentionality

This module, designed exclusively from pre-existing intellectual property (IP), measures an individual's preferences for different types of work activities and environments and the level of an individual's intention to seek out new job opportunities and career- and job-related training. The module includes two sections: the career interest assessment and the career intentionality assessment. Research suggests that career interests not only drive individuals' choices in educational and occupational development but also are key predictors of achievement, including educational attainment, job performance, occupational desirability, and income. When individuals have interests that are congruent or "fit" with their academic or work environments, they tend to be more satisfied, persist longer in their pursuit, perform better, and are more likely to succeed. Career intentionality examines individuals' attitudes and behaviors in their career development, a proximal predictor of eventual career success. This module provides a career interest profile across six dimensions, a career fit index that indicates the level of similarity or dissimilarity with current and desired occupations, and four high/moderate/low designations of intentionality. Additionally, two of these designations can be used to identify gaps between an individual's level of intentionality for finding a job and action taken to do so.

Table 3.2: Utility of the Career Interest and Intentionality module

Stakeholder	Relevance	Information Included in Score Report
Employers (businesses and trade associations)	• Job match • Supply chain planning • Improve tenure and satisfaction of workforce	• Fit indices targeting specific jobs
Academic systems	• Guidance and decision-making	• Career fit indices targeting specific interests
Policymakers	• Talent supply chain planning	• Aggregate interest profiles and fit indices for occupations • Gap charts between measured and expressed interests and job choice
Individuals	• Job search and choice • Decision-making	• Career interest profile • Career fit indices for current and desired job as well as jobs that most fit the individual's interests

3.1.3 Subjective Well-Being and Health

The assessment of Subjective Well-Being and Health is an important information source to policymakers who examine the well-being of the adult population and subpopulations, including workers and those seeking work. Measures of subjective well-being and health offer policymakers a valuable tool in assessing both the impact of policy as well as progress made toward short- and long-term goals. Subjective well-being has become a priority of the Organisation for Economic Co-operation and Development (OECD); interest in the topic in economics literature has increased considerably throughout the past two decades. Research has shown the predictiveness of health for education and work-related outcomes, as healthy individuals are more productive, less likely to be absent from work or school, and better able to help control health care costs. This module examines the main components of subjective well-being: life evaluation and positive and negative affect (using pre-existing IP), in addition to subjective and behavioral health indicators in line with the measures described in the World Health Organization's agenda. The health indicators include measures of subjective health, sleep quality, body mass index (BMI), smoking, diet, and exercise.

Table 3.3: Utility of the Subjective Well-Being and Health module

Stakeholder	Relevance	Information Included in Score Report
Employers (businesses and trade associations)	• Monitor workforce	• Aggregate index of overall well-being
Academic systems	• Monitor student well-being	• Individual and aggregate index of overall well-being
Policymakers	• Planning and research	• Aggregate index of overall well-being
Individuals	• Self-assessment and comparison	• Individual index of overall well-being • Feedback on health indicators and actions to take to improve health indicators

3.1.4 Behavioral Competencies (Available March 2018 to June 2020)

The Behavioral Competencies module was designed as a personality assessment for use in Education & Skills Online. Intended for developmental purposes, this assessment provides scores across 13 traits that are expected to be critical to success in education and the workplace. Findings across a range of studies have demonstrated that personality constructs in particular are important predictors of educational outcomes (Porchea, Allen, Robbins, & Phelps, 2010; Richardson, Abraham, & Bond, 2012; Robbins, Allen, Casillas, Peterson, & Le, 2006) and performance in the workplace (Barrick & Mount, 1991; Campbell, 1990; Campbell & Knapp, 2001). The predominant framework for personality measurement in the extant research continues to be the Big Five or five-factor model. The considerable research on the cross-cultural relevance and portability of the Big Five model of personality traits (Goldberg, 1990)—openness to experience, conscientiousness, extraversion, agreeableness, and emotional stability/neuroticism—in addition to constituent facets such as diligence, collaboration, and creativity, supported inclusion of this module in the international context. The relationship between the five-factor model and the constituent facets is described in detail in Chapter 8. This module, designed exclusively from pre-existing IP, used previously developed items to measure selected personality facets based on their high relevance and utility for academic and workforce readiness and success.

Table 3.4: Utility of the Behavioral Competencies module

Stakeholder	Relevance	Information Included in Score Report
Employers (businesses and trade associations)	• Valid pre-selection • Job match	• Benchmark scores • Developmental profile
Academic systems	• Career readiness and guidance • Accountability/success rates	• Benchmark scores • Developmental profile • Aggregate success rates/benchmark scores
Policymakers	• Gap analysis • Supply chain skill match • Comparative research	• Aggregate success rates/benchmark scores
Individuals	• Employment development • Feedback	• Benchmark scores • Developmental profile

3.2 Item translation and adaptation

The same translation and adaptation guidelines designed for the cognitive items were applied to the items in the noncognitive modules (see Chapter 4). In addition, development of the noncognitive modules followed the established linguistic quality control process designed to help ensure that instruments were comparable across countries. Layout checks were conducted by both contractors and countries to identify any display issues requiring modification.

3.3 Final item selection

After conducting the 2013 Field Test, responses were analyzed to determine which items and scales should be included in the final noncognitive modules. In general, items were evaluated to make sure they were appropriate for all countries included in the Field Test and that they did not duplicate information provided by other items within each module. In 2017, a second Field Test

was conducted with additional countries. The modules administered to the 2017 Field Test recipients only included those items selected for the final product in 2015. Responses from the 2017 Field Test examinees were analyzed to ensure that the results were similar to those obtained from the 2013 Field Test countries.

3.3.1 Skill Use final item selection

The Education & Skills Online 2013 Field Test included 1,963 examinees who completed the Skill Use module. Items associated with 11 of the 13 scales retained for reporting in PIAAC were administered as part of Education & Skills Online. Therefore, the scales from PIAAC available for potential inclusion were the eight skill use scales (reading, writing, numeracy, and ICT, both at home and at work) as well as scales related to informal training (learning at work) and nonliteracy skills used at work (influencing, planning at work).

Three primary criteria were used to identify which items/scales to include in the final module:

- Criterion 1: Scale reliability—When reporting subscale results, it is important that the scores have sufficient reliability to allow for defensible inferences to be made based on those scores. For cognitive measures, reliabilities of 0.80 or higher are generally preferred. But if this criterion were used, nearly two-thirds of the potential scales would have been flagged for possible exclusion. Therefore, a slightly relaxed criterion was used because the self-reporting scales include only small numbers of questions and were not intended for high stakes application; rather they were designed to describe statistical associations among self-reports of work-related behavior and skills assessed in the direct cognitive skills measures. In order to be considered for inclusion, the mean reliability across countries had to be greater than 0.6, as characterized by Cronbach's alpha.

- Criterion 2: Scale correlations—In addition to being sufficiently reliable, subscores should provide unique information about the measured background characteristics. Scales that provide redundant information may be of little utility; hence, the correlation between scales was considered. Potential scales with mean correlations across countries greater than or equal to 0.7 were flagged for possible exclusion.

- Criterion 3: Between-country differences—When item parameters are estimated for measures administered across countries, there is potential for item-by-country interactions that may lead to item misfit within countries if item parameters are not country-specific. Stated differently, the empirical response curves across countries may differ appreciably from the expected curves based on international item parameters. These differences may occur for individual items or for all or most items in a subscale.

Analyses based on the criteria above led to the exclusion of the scales related to informal training (learning at work) and nonliteracy skills used at work (influencing, planning at work), leaving eight scales altogether. The final Skill Use module for Education & Skills Online includes 57 items. For each scale included in the final module, test takers were asked to identify how frequently they use skills identified with that scale, using a five-point range of never (1) to every day (5). The final items are divided into the following scales:

Table 3.5: Items in final Skill Use module

Scale	Number of items (57)*
Reading at work	8
Reading at home	8
Writing at work	4
Writing at home	4
Numeracy at work	6
Numeracy at home	6
ICT at work	6
ICT at home	7

*The module also included 8 routing questions

The Education & Skills Online 2017 Field Test included 3,158 examinees who took the Skill Use module. To evaluate the fit of the data in these six countries to the previously estimated skill use item parameters from the PIAAC main study, we conducted an IRT analysis where we fixed the item parameters for each of the scales then examined the root mean squared deviations (RMSD) for each item across all scales and countries. Using a RMSD misfit criterion of 0.2, we found that the item parameters fit the data well; that is, the items had RMSDs below the 0.2 criterion.

3.3.2 Career Interest and Intentionality final item selection

Of the Education & Skills Online 2013 Field Test examinees, 2,636 completed the Career Interest and Intentionality module. The module includes two sections: the career interest assessment and career intentionality assessment. The career interest assessment consists of 60 items from the O*NET Interest Profiler Short Form (Rounds, Su, Lewis, & Rivkin, 2010). This set of items is composed of 10 items from each of the six RIASEC scales (realistic, investigative, artistic, social, enterprising, and conventional). All items have a five-point Likert response scale from strongly dislike (0) to strongly like (4). From the Field Test responses, scale scores were calculated for each RIASEC dimension by adding the 10 item scores within each dimension. The mean interest profiles are similar across most countries/languages, with a few exceptions (e.g., Japan, Czech Republic). Each of the six RIASEC scales had high internal consistency, with Cronbach's alpha reliability values ranging from .89 to .93. The scales were also highly reliable across countries.

The career intentionality assessment consists of 26 items. This set is composed of 6 items that measure job-seeking intentionality, 6 that measure training intentionality, 4 that measure job-seeking and training self-efficacy, and 10 that measure taking active steps. The job-seeking intentionality, training intentionality, and job-seeking and training self-efficacy scales had a six-point response scale from strongly disagree (1) to strongly agree (6). Scale means were calculated for each scale by averaging item scores. Each item in the "taking active steps" scale had a binary response of yes (1) or no (0). The total number of yes responses was used as the scale score. Field Test responses indicated that overall individuals had moderate intention to get a new job or seek additional job training and relatively high self-efficacy to do so but had taken very few active steps to find work. The level of career intentionality varied across countries/languages. Individuals from Japan had lower scores on all of the career intentionality scales. All career intentionality scales

had high internal consistency, with Cronbach's alpha reliability values ranging from .84 to .97. The scales were also highly reliable across countries.

Due to the high reliability across countries for both items in the career interest and career intentionality assessments, all items and scales included in the Field Test were retained in the final product.

Of the Education & Skills Online 2017 Field Test examinees, 3,136 took the Career Interest and Intentionality module. Within-country reliabilities for all six interest scales (*Realistic*, *Investigative*, *Artistic*, *Social*, *Enterprising*, and *Conventional*) ranged from .86 to .93. Pooled (over country) scale intercorrelations ranged from .31 (Artistic, Conventional) to .62 (Artistic, Social), with good comparability between countries. Within-country reliabilities for the four Career Intentionality Scales (*Job-Seeking Intentionality*, *Training Intentionality*, *Jobs and Training Self-Efficacy*, and *Taking Active Steps*) ranged from .81 to .97. Pooled (over country) scale intercorrelations ranged from .17 (Self-Efficacy, Taking Active Steps) to .55 (Job Seeking Intentionality, Taking Active Steps), with good comparability between countries. The results suggest no significant problems with comparability for the additional countries.

3.3.3 Subjective Well-Being and Health final item selection

Of the Education & Skills Online 2013 Field Test examinees, 1,105 completed the Subjective Well-Being and Health module. These examinees came from five countries: Canada, the Czech Republic, Japan, the United States, and Spain. The module contains two sections: subjective well-being as well as subjective and behavioral health.

Subjective well-being focuses on personal feelings or attitudes toward one's life. The extant literature suggests that subjective well-being is characterized by both cognitive and emotional life assessments. This module employs the Satisfaction with Life Scale (SWLS) (Diener, Emmons, Larsen, & Griffin, 1985) as a cognitive measure, while the emotional component of subjective well-being is measured using an adapted scale based on the Positive and Negative Affect Schedule (PANAS) (Watson, Clark, & Tellegen, 1988) and I-PANAS-SF (Thompson, 2007), an internationally validated short form of the instrument.

The SWLS (Diener et al., 1985) is a well-validated, multi-item, cognitive measure of global life satisfaction. It has been translated into over 30 languages and is included in numerous international surveys and research initiatives including the World Values Survey, the German Socio-Economic Panel, and the British Household Panel Survey. The original SWLS is a five-item instrument with a seven-point scale that elicits respondents' global judgments of their lives. The psychometric properties of the SWLS include reported internal consistencies of 0.80 or greater (Alfonso et.al, 1996; Diener et al., 1985; Pavot et al., 1991) and test-retest reliability from 0.84 in a two-week interval to 0.54 over a four-year interval, demonstrating stability and a sensitivity to change over time (Alfonso et. al, 1996; Pavot & Diener, 1993; Magnus et al., 1992). The Education & Skills Online adaptation of the SWLS includes four of the original items. Reliabilities of 0.70 or higher are considered sufficient and are particularly good for scales with as few items as the scales in the Subjective Well-Being and Health module. Cronbach's alpha was used as the indicator of scale reliability and was evaluated at the country level because drastic differences in scale reliability between countries would not be desirable.

The Education & Skills Online implementation of SWLS includes four of the original items and a modified six-point response scale including strongly disagree, disagree, slightly disagree, slightly agree, agree, and strongly agree. Within-country reliability estimates for the Education & Skills Online implementation of SWLS were consistent with those found in the literature, ranging from 0.82 to 0.88 for the four-item version. The decrease in reliability was very small considering how few items make up the SWLS, and the decline was consistent across countries. Reliability analysis suggested the Cronbach's alpha estimates would increase by 0.01 to 0.03 with the fifth SWLS item.

The second element of subjective well-being is an emotional evaluation, which is more descriptive of the degree to which people emotionally experience their lives. While life satisfaction is assessed on a single dimension, the emotional evaluation is composed of both positive affect and negative affect, which are two distinct dimensions. The Education & Skills Online measure for emotional evaluation is an adapted version of the PANAS (Watson, Clark, & Tellegen, 1988) and I-PANAS-SF (Thompson, 2007). While the PANAS is composed of 10 positive affect items and 10 negative affect items, the I-PANAS-SF contains five positive affect and five negative affect items. The Education & Skills Online adapted scale is composed of two original items, four items from the original PANAS scale (Watson et al., 1988), and three items from the PANAS/I-PANAS-SF (Thompson, 2007) all measured with a five-point scale ranging from "very slightly or not at all" to extremely. Respondents are asked to rate their experience of each emotion during the previous week.

While the internal consistencies of the PANAS positive affect and negative affect scales using Cronbach's alpha are estimated from 0.85 to 0.89, cultural differences in emotional experience and expression as well as linguistic issues of comparable naming often make global comparisons problematic. The reliability estimates for the Education & Skills Online version of PANAS varied between countries for both positive and negative affect. For positive affect, reliabilities using Cronbach's alpha range from 0.70 to 0.84. Though these estimates are still at or above the minimum acceptable, the range here suggests quite a bit of inconsistency between countries. Reliabilities for negative affect are not as variable between countries as those for positive affect, ranging from 0.72 to 0.78.

Health is a complex, multidimensional construct whose definition has evolved from a purely biological focus to include psychosocial factors critical to well-being. Gathering health data is an integral component of the ongoing effort to monitor economic and social progress across countries and promote policies aimed at improving overall quality of life (Organisation for Economic Co-operation and Development, 2012). The Education & Skills Online measures of subjective and behavioral health include 14 survey items on the feelings and behaviors most relevant to health as described in the OECD agenda. These include items on subjective health, BMI, nutrition, exercise, sleep, and smoking status.

Subjective health is a single item measure of an individual's perceived health, consistent with other subjective health measures. Behavioral health indicators outlined by the World Health Organization include BMI, nutrition, physical activity, smoking, and sleep. Responses from two questions on height and weight are used to calculate BMI, an internationally accepted health measure. As an indicator of nutrition, based on the international nutritional recommendations, four questions eliciting daily and weekly consumption of fruits and vegetables are included. Two items ask about duration and quality of sleep, as they are core features of commonly accepted sleep

recommendations. Four items examining physical activity in terms of frequency, duration, and intensity are presented in addition to a single item on smoking status.

Based on analysis of the data collected in the 2013 Field Test, several changes were made from the original battery of measures: The *Domain Satisfaction* and *Eudaimonic Well-Being* scales were eliminated due to poor psychometric properties (*Domain Satisfaction*) or redundancy with other scales (*Eudaimonic Well-Being* was redundant with *Satisfaction with Life*). Several items were removed from several of the scales due to poor psychometric properties. Anchoring vignettes for Healthy Behaviors were eliminated due to the amount of time it took to complete them. This resulted in a final module of 4 items in *Satisfaction with Life Scale*, 4 items measuring *Positive Affect* and 5 items measuring *Negative Affect* in the *Positive Affect Negative Affect Schedule* (PANAS), 13 Health Behaviors items, and 2 Body Mass Index (BMI) items.

Of the Education & Skills Online 2017 Field Test examinees, 3,147 took the Subjective Well-Being and Health Module. Final items in the module resulted in good scale reliabilities and good cross-country comparability. For the SWLS and the PANAS, means and standard deviations for items in the scales were reviewed for each language version. Reliability was also reviewed for each language version by examining item-total correlations, alpha reliability coefficients for the scale, and alpha-if-item-deleted values for the scale. For the SWLS, the within-country reliability ranged from 0.83 (Estonia-Russian) to 0.92 (Australia and Slovakia), which is consistent with the results from the 2013 Field Test. For the PANAS, the within-country reliabilities for positive affect for four countries were consistent with the 2013 results, ranging from 0.70 (Chile) to 0.85 (Estonia-Russian and Slovakia). One country, Slovenia, showed a lower within-country positive affect reliability of 0.55. Reliabilities for negative affect ranged from 0.79 (Slovenia) to 0.85 (Estonia-Russian and Slovakia), which is consistent with the Round 1 results.

For the health indicators and behaviors questions, we reviewed distribution of responses across categories for both the pooled country-level data and by country. The distribution of responses across categories for each country was fairly consistent.

3.3.4 Behavioral Competencies final item selection

Personality traits have been used to predict workplace behaviors with varying reliabilities depending on the measures. The predominant framework for personality measurement is the Big Five, as described in section 3.1.4. The Behavioral Competencies module of Education & Skills Online is designed as a personality assessment intended for developmental purposes that assesses 13 personality traits that are components of the Big Five that are expected to be critical to educational and workplace success. The Behavioral Competencies assessment consists of 208 statements that represent 13 traits indicative of important workplace behaviors, employing a forced-choice methodology that combines those items into 104 pairs of statements, where respondents are required to choose the statement in the pair that most reflects their personality. Forced-choice methodology is resistant to test faking as each of the items in a pair are equally desirable.

Of the Education & Skills Online 2013 Field Test examinees, 2,517 completed the Behavioral Competencies module. Published coefficients of stability for personality scales generally show a wide range of values, from around 0.50 to 0.90. A meta-analysis (Viswesvaran & Ones, 2000) of these reliabilities for Big Five constructs (Goldberg, 1990), showed average mean reliabilities of

0.72, 0.69, 0.71, 0.76, and 0.75 for conscientiousness, agreeableness, openness to experience, extraversion, and emotional stability, respectively. The Behavioral Competencies module produced overall reliabilities of scales by Big Five domain of 0.87 for agreeableness, 0.88 for conscientiousness, 0.86 for extraversion, 0.88 for emotional stability, and 0.90 for openness; comparable to other personality measures. An examination of Behavioral Competencies scale descriptive statistics and distributions, using scale reliabilities, means, and standard deviations for each country, produced reliabilities of the 13 reported Behavioral Competencies scales ranging from 0.79 to 0.86, which are equally comparable. Reliabilities for each scale, grouped by Big Five domain, are detailed in Table 3.6.

Table 3.6: Behavioral Competencies scale reliabilities

Big 5 Domain	BPC Scale	α
Agreeableness	Collaboration	0.86
	Generosity	0.84
Conscientiousness	Diligence	0.85
	Organization	0.86
	Dependability	0.82
	Self-Discipline	0.79
Extraversion	Assertiveness	0.85
	Friendliness	0.80
Emotional Stability	Stability	0.82
	Optimism	0.84
Openness to Experience	Inquisitiveness	0.83
	Creativity	0.83
	Intellectual Orientation	0.81

Of the Education & Skills Online 2017 Field Test examinees, 2,861 from 6 countries (7 language versions) took the Behavioral Competencies module. Reliability estimates were comparable to those found for the nine language versions analyzed in 2013, with values within the averages observed in the published literature. For data collected in both the 2013 and 2017 Field Tests, scale intercorrelations were low, all less than approximately 0.50. These correlations indicated that each scale provides unique information about the examinees. Finally, these correlations were relatively consistent across each of the Field Test countries, indicating that operation and use of the scales across the locales is reasonable for use with the current application as a developmental module.

References

Alfonso, V. C., Allison, D. B., Rader, D. E., & Gorman, B. S. (1996). The extended satisfaction with life scale: Development and psychometric properties. *Social Indicators Research*, *38*(3), 275-301.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1-26.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol. 1, pp. 687-731). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P., & Knapp, D. J. (2001). Exploring the Limits in Personnel Selection and Classification, New Jersey: Lawrence Erlbaum Associates.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49, 71-75.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216.

Magnus, K. B., & Diener, E. (1991). *Factors of happiness: A longitudinal analysis of personality, life events, and subjective well-being.* Paper presented at the 63rd Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Organisation for Economic Co-operation and Development (2011). How's life?: Measuring well-being. Paris, France: Author. doi:10.1787/9789264121164-en

Pavot, W., & Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment, 5*(2), 164-172.

Pavot, W., Diener, E., Colvin C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment, 57*(1), 149-161.

Porchea, S. F., Allen, J., Robbins, S., & Phelps, R. P. (2010). Predictors of long-term enrollment and degree outcomes for community college students: Integrating academic, psychosocial, socio-demographic, and situational factors. *The Journal of Higher Education*, 81, 750-778.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138,* 353-387.

Robbins, S. Allen, J., Casillas, A., Peterson, C., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology*, 98, 598-616.

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). *O*NET® interest profiler short form psychometric characteristics: Summary*. Raleigh, NC: National Center for O*NET Development.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology, 38*, 227-242. doi:10.1177/0022022106297301

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*(2), 224-235.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Chapter 4: Translation, Adaptation, and Validation of Education & Skills Online Field Test Assessment Instruments

4.1 Overview

The Education & Skills Online final product instruments use the translated versions of the assessment developed for the Field Test. The 2013 Field Test instruments, comprising both cognitive test units and noncognitive modules, were prepared for administration to participating adults in 9 countries in 11 language versions. As two countries did not complete the 2013 Field Test, final versions are available in 9 language versions for 7 countries. In 2017, the Field Test instruments were prepared for 6 countries in 7 language versions (1 language version, Australian English, borrowed translations from the Irish English version of Education & Skills Online rather than going through the translation, adaptation, and validation process described in this chapter). Localization (translation, adaptation for local use, and independent validation) of the instruments was a key aspect of the development process for the Field Test.

Education & Skills Online builds on the PIAAC Round 1 survey, with a legacy of previously localized instrumentation accompanied by materials newly developed for the Education & Skills Online Field Test. The localization process was a complex operation involving staff from various organizations and components that followed different processes.

In 2013, 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.
- BranTra Premium Translation Services was responsible for translating newly created materials, while cApStAn retrieved localized PIAAC materials that could be utilized for preparing PIAAC linking items included in Education & Skills Online.
- cApStAn independently verified materials translated by BranTra and adapted them into additional versions of the same language.
- BranTra processed the verification feedback from cApStAn and harmonized newly created and linking materials.
- Representatives of participating countries reviewed the harmonized materials to propose changes before finalization.

- cApStAn reviewed and selectively implemented the changes proposed by countries. ETS further implemented any necessary layout adjustments. As a last step (once wording and layout of the materials were final), cApStAn verified language-dependent automated scoring rules.

In 2017, the process was somewhat different:

- cApStAn Linguistic Quality Control, Inc. in close co-operation with ETS, developed the localization design and was responsible for preparing instructions and training for countries for the translation process.
- Countries were responsible for translating all new materials that could not be retrieved from PIAAC Rounds 1 or 2.
- cApStAn independently verified materials translated by countries.
- cApStAn was responsible for transferring PIAAC Round 1 cognitive materials to Education & Skills Online xliff templates. This process was not necessary for countries that participated in PIAAC R2 (Chile, Slovenia), as the R2 xliffs are compatible with Education & Skills Online.
- Countries reviewed the transferred materials and had the opportunity to request changes to correct outright errors or outdated adaptations.
- cApStAn verified these change requests, in collaboration with the ETS test developers, and implemented centrally those changes that were accepted.

ETS provided overall guidance and technical support for the export and import of XLIFF files used with the assessment delivery software (XLIFF is an XML-based format for exchanging data during the text translation process), setup and maintenance of the Education & Skills Online Item Management Portal, and layout adaptation, and reviewed scoring rules. The seven countries in the 2013 Field Test for which Education & Skills Online-localized instruments are available are Canada, the Czech Republic, Ireland, Italy, Japan, Spain, and the United States. The six countries in the 2017 Field Test for which the Education & Skills Online materials were fully finalized are Australia, Chile, Estonia, Slovak Republic, Slovenia and Russia.

4.2 Localization design, including LQA and LQC processes

The Education & Skills Online localization design in both the 2013 and 2017 Field Tests was based on the design used for the Programme for the International Assessment of Adult Competencies (PIAAC), which was in turn based on the design used for the Programme for International Student Assessment (PISA). Newly created cognitive materials followed a number of steps to ensure linguistic quality, including: preparation of the source materials for localization, double translation by two independent translators; creation of a merged version by a reconciler; independent verification of the materials by professional and appropriately trained and monitored staff; and documentation of all steps leading to the final localized national versions.

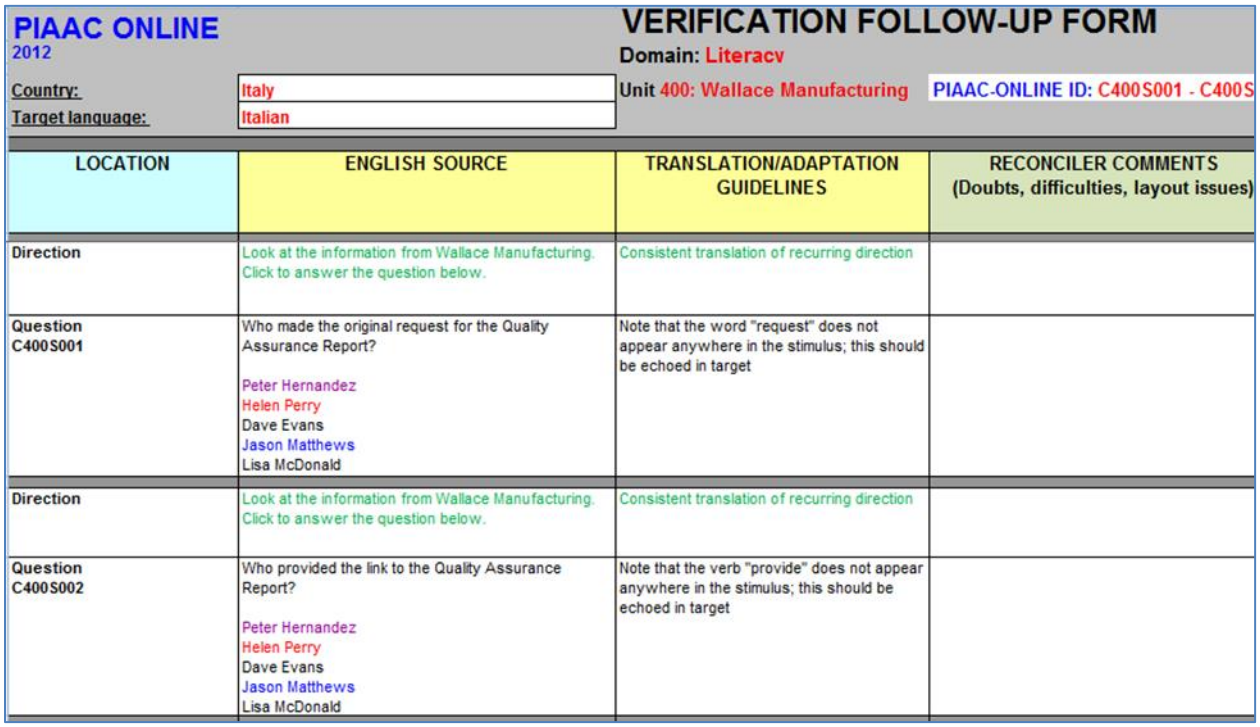
In 2013, an important difference versus PIAAC was the limited role of participating countries in the Education & Skills Online localization process. Whereas participating countries were responsible for their translated instruments in PIAAC (under the guidance of the PIAAC

Consortium), in Education & Skills Online they were invited to review "centrally produced" localized instruments only at the end of the process (see section 4.8).

For Education & Skills Online, the LQA processes implemented by cApStAn in cooperation with ETS and BranTra included:

- Early resolution of potential localization issues via preliminary scrutiny of source assessment materials in order to anticipate adaptation issues, ambiguities, cultural issues, or item translatability problems, with suggestions for either rewording or adding item-specific translation/adaptation guidelines. This was an upstream LQA process that aimed to reduce the difficulties and workload encountered later downstream.
- Reuse of the PIAAC Translation/Adaptation Guidelines, a key document setting out requirements and roles, offering pointers on 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. 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). A sample of a VFF is shown in Figure 4.1.

Figure 4.1: Sample VFF showing item-specific guidelines



PIAAC ONLINE 2012

VERIFICATION FOLLOW-UP FORM

Domain: Literacy

Country: Italy

Target language: Italian

Unit 400: Wallace Manufacturing

PIAAC-ONLINE ID: C400S001 - C400S

LOCATION	ENGLISH SOURCE	TRANSLATION/ADAPTATION GUIDELINES	RECONCILER COMMENTS (Doubts, difficulties, layout issues)
Direction	Look at the information from Wallace Manufacturing. Click to answer the question below.	Consistent translation of recurring direction	
Question C400S001	Who made the original request for the Quality Assurance Report? Peter Hernandez Helen Perry Dave Evans Jason Matthews Lisa McDonald	Note that the word "request" does not appear anywhere in the stimulus; this should be echoed in target	
Direction	Look at the information from Wallace Manufacturing. Click to answer the question below.	Consistent translation of recurring direction	
Question C400S002	Who provided the link to the Quality Assurance Report? Peter Hernandez Helen Perry Dave Evans Jason Matthews Lisa McDonald	Note that the verb "provide" does not appear anywhere in the stimulus; this should be echoed in target	

- Participation in preparation and delivery of training sessions for BranTra's translation teams, mostly administered through webinars.
- Continued assistance from cApStAn to BranTra throughout the localization process (help desk), liaising with ETS as needed.
- Throughout the localization process, cApStAn took care of an errata management process, whereby errors in the source identified by translators or reconciles from BranTra or cApStAn verifiers were tracked and listed for correction in all translated versions.

The implemented LQC processes included:

- Verification by cApStAn verifiers of translated versions submitted by BranTra (for newly translated materials) and quality checking of translated versions obtained through reuse of PIAAC materials. Verification involved sentence-by-sentence comparison versus the source versions with reporting of residual errors and expert advice where corrective action was suggested.
- 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.
- Verification by cApStAn staff, assisted by verifiers as needed, of language-dependent automated scoring rules (for the items with "highlight in stimulus" response mode).

In 2017, the localization design followed the PIAAC model in which countries are responsible for translating any new content, and cApStAn's role is to verify the translated materials, as well

as to implement requested (and approved) changes in linking units centrally. In 2017, the LQA processes implemented by cApStAn in cooperation with ETS included:

- Reuse of the Translation/Adaptation Guidelines from PIAAC as well as from the Education & Skills Online 2013 Field Test.
- Updating the Verification Follow-up Forms (VFF) used in the 2013 Field Test to accommodate the changes in the localization process from 2013 to 2017. This tool provides a detailed history of all changes and comments made throughout the translation and verification process.
- Preparation of a searchable online translation memory of all PIAAC Round 1 translations to be used as references by the countries' translation teams.
- Training the national teams in the double-translation and reconciliation process, and in using the computer-aided translation tool OLT (Open Language Tool), a separate widget for producing a file that merges translation 1 and 2 into one single file as well as the searchable online translation memory to ensure consistency with PIAAC materials.
- Preparation of role-specific instructions for the translator and reconciler for the translation process and providing support to countries (helpdesk) throughout the translation process.
- Semi-automated transfer of PIAAC linking cognitive units to the Education & Skills Online environment (for PIAAC Round 1 countries only).

4.3 Translation/adaptation procedures for newly developed cognitive materials

In 2013, the cognitive materials newly developed for Education & Skills Online were translated and adapted from the international English source version into 9 national versions for 7 countries comprising 6 languages, as shown in Table 4.1. The materials used in English-speaking countries (Canada, Ireland, and the US) underwent an adaptation procedure to local usage. The same was done for the Spanish version used in the US.

Table 4.1: Localization process for new cognitive materials by country and language version

Country	Language	Localization process for new cognitive materials
Canada	English	Adaptation of international English source
Canada	French	Adaptation of France-French version
Czech Rep	Czech	Double translation and reconciliation
Ireland	English	Adaptation of UK-English version
Italy	Italian	Double translation and reconciliation (including PIAAC problem solving)
Japan	Japanese	Double translation and reconciliation (except PIAAC reduced set reading components: single translation and reconciliation)
Spain	Spanish	Double translation and reconciliation
US	English	Adaptation of international English source
US	Spanish	Adaptation of Spain-Spanish version

The translation approach used for this part of the project was double translation by two independent translators, followed by reconciliation. Double translation helps identify misinterpretations or ambiguities, idiosyncratic wording, or translator oversights; moreover, it offers stylistic variants among which to choose to achieve a more fluent translation. The procedure is a state-of-the art approach for assessment instruments.

In the double translation-plus-reconciliation process, the main task of the reconciler is to "merge" the two independent translations in such a way that:

- the resulting national version is as equivalent as possible to the source version,
- all possible translation errors have been corrected, and
- the wording is as fluent as possible.

This means that the reconciler's role is not limited to just selecting the "best" translation out of the two and briefly proofreading it. First-hand translations always need accurate, in-depth reworking. The reconciler has to:

- read both translations, sentence by sentence,
- check each sentence against the source version and consult any item-specific guidelines,
- carefully rework the translated text in order to make it as accurate and fluent as possible, and
- preview the final text (against the source) to spot any layout issues.

The aim is to strike the right balance: The translation must not be literal to the point that it sounds awkward, but neither should it deviate too far from the source version, which could affect the functioning of the assessment items in unexpected ways.

Translators are skilled practitioners at translating into their mother tongue and experienced or trained in translation of survey instruments. Reconcilers have strong language skills in both the source and target languages and are knowledgeable about the subtleties of reconciliation.

The translators and reconcilers received general guidelines prepared jointly by BranTra and cApStAn, based on the PIAAC Translation/Adaptation Guidelines. This document stressed the need for high-quality translation in order to collect internationally comparable data—with the additional challenge, in the case of cognitive materials, to "retain the cognitive equivalence of tasks as much as possible." It laid down requirements for translators and reconcilers, addressed security/confidentiality aspects, translation traps, and the general principles for cultural adaptations, and explained the LQC processes.

Translators and reconcilers also were provided with item-specific translation guidelines (also referred to as "translation and adaptation notes" or "item-by-item notes") that were echoed in the VFF.

Remote training was organized for the translators and the reconcilers. The major focus of these training sessions was to familiarize translators with the guidelines for translating and adapting tasks. That is, in addition to stressing the importance of accurate translations, the workshops were used to emphasize the key role that the assessment construct played in helping to develop the adaptation guidelines. In order to accomplish these goals, these workshops were used to

provide a brief overview of the construct, demonstrate sets of specific items, and share and discuss specific guidelines for the proposed items. In addition, translators and reconcilers received precise instructions concerning the technical environment for the localization process: the OLT (Open Language Tool) translation software used on XLIFF files exchanged via the Education & Skills Online Item Management Portal.

Throughout the translation process (from the initial double translations to reconciliation), translators and reconcilers could consult with BranTra staff in cases of queries or concerns. If needed, queries were relayed to cApStAn, which in turn could also consult with ETS. In addition, reconcilers were encouraged to write comments for the attention of the verifier in the VFF to call attention to difficulties, purposeful deviations, and decisions made.

Table 4.2 shows the different components of all Education & Skills Online instruments with their word counts and a summary description of the process followed for production of the localized versions.

Table 4.2: Localization process components

Component	Source Words	Localization Process
PIAAC literacy and numeracy units	9,545	Copied and pasted, quality checked, and reviewed for harmonization with newly created cognitive materials
PIAAC problem-solving units	11,088	Copied and pasted, quality checked, and reviewed for harmonization with newly created cognitive materials, except for Italy (double translated, reconciled and verified)
Help and orientation modules	2,724	Copied and pasted, quality checked, and reviewed for harmonization with newly created cognitive materials, except for Italy's problem-solving sections (single translated, reviewed by reconciler and verified)
Reading components (full set)	2,252	Copied and pasted, quality checked, and reviewed for harmonization with newly created cognitive materials, except for Japan (see below)
Reading components (reduced set)	2,007	For Japan, a reduced set (without word meaning section) was single translated, reviewed by reconciler, and verified
Newly developed literacy and numeracy units	11,916	Double translated, reconciled, verified, reviewed post-verification, including for harmonization with PIAAC cognitive materials; adapted (in the case of different-country, same-language versions)
Core background questionnaire, noncognitive modules, navigation & transition screens	7,577	Single translated, reviewed by reconciler and verified, reviewed post-verification; adapted (in the case of different-country, same-language versions)
ISCO-ISCED* drop-down menu items	2,844	Researched from official sources and compiled in Excel files

* International Standard Classification of Occupations-International Standard Classification of Education

In 2017, the cognitive materials newly developed for Education & Skills Online were translated and adapted from the international English source version into 7 national versions for 6 countries comprising 6 languages, as shown in Table 4.3.

Table 4.3: Localization process for new cognitive materials by country and language version

Country	Language	Localization process for new cognitive materials
Australia	English	Used Education & Skills Online 2013 materials from Ireland with updates to the currency and background questionnaire only
Chile	Spanish	Adaptation of Spain-Spanish version (from Education & Skills Online 2013)
Estonia	Estonian	Double translation and reconciliation
Estonia	Russian	Adaptation of Russia-Russian version (from Education & Skills Online 2017)
Russia	Russian	Double translation and reconciliation
Slovenia	Slovene	Double translation and reconciliation
Slovak Rep.	Slovak	Double translation and reconciliation

The process applied to translating new materials was the same as in 2013; the main difference was that the translations were produced by the participating countries, not centrally by BranTra.

Training was organized for the countries in the form of webinars. There were two main focus areas in the training webinars: to explain the concept and benefits of the double-translation and reconciliation process, and to explain the tools necessary to perform the required processes. The following tools and functionalities were explained using practical examples:

- VFFs (Verification Follow-up Forms): the forms in Excel format that include all item-specific guidelines and need to be consulted throughout the translation process.
- OLT (Open Language Tool): the computer-aided translation tool needed for translating the materials in xliff format. How to take advantage of the translation memory functionality was also explained.
- Searchable online translation memory for looking up translations used in PIAAC Round 1 and 2, and hence to ensure consistency between the instruments (this tool was introduced in 2017, and was not used in the Education & Skills Online 2013 Field Test).
- Merging widget: a simple online tool designed to merge the two independent translations (T1 and T2) into one xliff file, which makes it easier for the reconciler to toggle between the two versions (this tool was introduced in 2017 and was not used in the Education & Skills Online 2013 Field Test).

Table 4.4 shows the different components of Education & Skills Online instruments that were produced for the 2017 Field Test, with their word counts and a summary description of the process followed for production of the localized versions.

Table 4.4: Localization process components

Component	Source Words	Localization Process
PIAAC literacy and numeracy units	9,545	Transferred from monolingual xliff files using a semi-automated method whereby the PIAAC files were used to create translation memories (TM) and these TMs were then used to automatically populate the Education & Skills Online 2017 xliffs. These 'pre-translated' xliffs where then checked by a linguist to detect residual errors stemming from the transfer process. Countries were asked to review the transferred units and request changes if errors were identified. These change requests were verified, and accepted changes were implemented centrally by cApStAn. Highlight scoring blocks in highlighting items (Literacy items only) were defined by countries and verified by cApStAn.
PIAAC problem-solving units	11,088	Same as above
Help and orientation modules	2,724	Same as above
Reading components (full set)	2,252	Transfer from Word files used in PIAAC using the same semi-automated method. Otherwise same as above.
Newly developed literacy and numeracy units	11,916	Double translated and reconciled by countries, verified by cApStAn. In the case of Chile and Estonian Russian: an adaptation process from the materials translated for Spain in Education & Skills Online 2013 Field Test or by Russia in the Education & Skills Online 2017 Field Test, respectively. Adaptations made by countries were verified by cApStAn. Highlight scoring blocks in highlighting items (Literacy items only) were defined by countries and verified by cApStAn.
Transition screens	152	Single-translated by countries, verified by cApStAn
Core background questionnaire, noncognitive modules	7,577	Double-translated and reconciled by countries, verified by cApStAn
ISCO * drop-down menu items	2,844	Researched from official sources and compiled in Excel files, reviewed by countries

* International Standard Classification of Occupations

4.4 Translation/adaptation procedures for newly developed noncognitive materials

In 2013, the noncognitive materials, comprising the Core BQ (background questionnaire), four noncognitive modules, and various navigation and transition screens, were single translated or adapted into the same national versions (see Table 4.1). These materials were subsequently reviewed and proofread by the reconciler. This approach, which is more cost effective than double translation and reconciliation (but likewise followed by independent verification), is deemed sufficiently robust for materials that are less sensitive in terms of data collection.

The guidelines document distributed to translators and reconcilers also covered issues to be considered in the translation of noncognitive materials, with a call to produce a translation that maintains the measurement properties and the meaning of the source questions while at the same time being as understandable as possible. Furthermore, item-specific translation and adaptation guidelines were also provided in the VFFs for noncognitive materials, offering further clarifications (e.g., on the meaning of terms or phrases or on characteristics of response categories) for a certain number of items.

As for cognitive materials, throughout the translation process, the translators and reconcilers could consult with BranTra staff, liaising with cApStAn and ETS as needed.

In 2017, the countries were also asked to apply the double-translation and reconciliation model in the translation of the Background Questionnaire and other noncognitive modules. In the Skill Use module existing translations from PIAAC were transferred to the translatable Excel files by cApStAn before the files were made available to countries. The entire double-translation and reconciliation process for each noncognitive module took place in an Excel file that also included detailed instructions on how to proceed, with notes about which items come from PIAAC and which are repeated within the module, so that consistency between identical items could be kept.

One point of detail: for the 2013 and 2017 Field Tests, long lists of International Standard Classification of Occupations descriptors (occupation titles), which were used as response options for one question included in the Core BQ and one question included in the Career Interest and Intentionality module, were not translated; instead, official translations were researched, checked, and collated. In the 2013 Field Test, official translations of the International Standard Classification of Education (academic study fields), which is used in the Core BQ, were also researched, checked, and collated. In both the 2013 and 2017 Field Tests, countries had the opportunity to review these fields and request changes.

4.5 Translation/adaptation procedures for PIAAC cognitive and noncognitive materials

Many cognitive and noncognitive materials used in Education & Skills Online could be retrieved from PIAAC and required little or no changes or updates: literacy items, numeracy items, problem-solving items (except for Italy), help and orientation modules, and reading components (except for Japan).

The reusable localized PIAAC materials had already been extensively validated (full verification and final check at Field Test followed by focused verification and final check at the Main Study

phase—see PIAAC Technical Report[1]) and thus did not need to go through verification again. However, the assessment delivery platform changed for Education & Skills Online versus PIAAC, so the materials had to be transferred into new XLIFFs and then "quality checked" by verifiers.

The transfer was performed by a mix of cApStAn staff and cApStAn verifiers, all of them proficient in the language(s) on which they worked. In 2013, they were instructed to copy from the XML printout of the PIAAC XLIFF to the Education & Skills Online XLIFF, segment by segment, and then to launch the preview of the XLIFF on the Education & Skills Online portal to check that all content was correctly transferred and that there were no layout or technical issues.

After the transfer, the files went through a quality check procedure. This step was also performed by cApStAn and by some of its verifiers, but with a rule that all materials would be checked by a different person than the one who performed the copying and pasting. This check consisted of:

- reading all sentences/elements to make sure all texts were translated and that there were no residual errors (e.g., double periods, untranslated button labels, layout issues, incorrect date format, decimal or thousand separators, defective graph captions, etc.)
- simulating respondent action to see all screens and check that items functioned as expected
- making all necessary changes in the XLIFF
- reporting any residual (not correctable) formatting or layout issue in the VFF (for the attention of ETS technical staff).

In the Education & Skills Online 2017 Field Test, a more automated process was implemented for the transfer: the PIAAC Round 1 xliffs were only available as monolingual files (i.e., both source and target segments had the same translatedtext). Because of this, the xliffs could not be used directly for creating a translation memory (TM), instead they had to be first 'aligned' to the corresponding English text using a specialized software. The outcome of this alignment process was a TM that could be used to populate the Education & Skills Online xliffs in a semi-automated way. This process is less error-prone than a manual transfer, however it does include some manual work and therefore the resulting xliffs need to be carefully quality-checked after the transfer. The entire transfer and review process was carried out by cApStAn in-house staff in co-operation with the relevant linguists. The finalized PIAAC xliffs were then uploaded by ETS on the previewing portal for the countries to review.

4.6 Verification and adaptation procedures

4.6.1 Verification

Verification is the LQC process put in place to check to what extent the translation procedure was successful, correcting course as needed. Thus, the verifiers' task was to:

- ensure linguistic correctness and cross-country equivalence of the different language versions of the Education & Skills Online instruments,

[1] The PIAAC Technical Report can be found at http://www.oecd.org/site/piaac/PIAAC%20Tech%20Report_Section%205_update%201SEP14.pdf.

- achieve the best possible balance between faithfulness to source and fluency in target language, and
- document interventions.

The verifiers were selected from cApStAn's experienced team. They are native speakers of each of the target languages and are highly proficient in English as a source language and as a working language to document their findings. They are trained to assess whether translation and adaptation guidelines are followed and to document possible deviations, insert corrections as needed, and provide expert linguistic advice. They are knowledgeable about equivalence issues, translation traps, and meaning shifts that are likely to affect response patterns in assessments. They also have experience in assessing the relevance of cultural adaptations in data collection instruments. They are all familiar with the use of "verifier intervention categories" and verifier comments in a standardized form.

Verifiers received detailed procedures and technical instructions to achieve their task. They were instructed about the Education & Skills Online Item Management Portal, OLT software, and the particularities of the different instruments to be verified.

The verifier was asked to:

- Perform a focused quality control task to deliver error-free files, as in a regular proofreading task. In particular, they were asked to examine whether the target version was linguistically correct and if it struck the right balance between faithfulness to the source and fluency in the target language.
- Avoid making preferential choices, replacing words with their synonyms, or making stylistic improvements if the translated version seemed correct and acceptable. However, the verifier could suggest a change by inserting a comment in the monitoring instrument (VFF) without making the change in the XLIFF file.
- When necessary, propose corrective actions in the target file and document these interventions in English in the VFF. Documentation involves selecting an intervention category to identify the type of issue and writing an explanatory comment. The verifier was also asked to report any residual (but not corrected) formatting and layout issues for the ETS technical team.
- Double check that the translation/adaptation guidelines listed in the VFF were followed and that any comment appearing in the "Reconciler Comments" column of the VFF was considered.
- Check national versions against the latest Education & Skills Online errata list.

In the 2017 Field Test, the verification process was the same as in the 2013 Field Test.

4.6.2 Adaptation

For the English versions in the 2013 Field Test (Canada, Ireland, US), the French version for Canada, and the Spanish version for the US, verifiers were asked to play the role of adapters and be in charge of making the changes needed to fit the national context.

In particular, adapters were asked to focus on precise elements that are frequently adapted such as:

- spelling conventions
- lexical choices
- date and time formats
- punctuation conventions
- currencies/units of measure
- fictitious names and addresses

The adapter was responsible for the consistent implementation of all necessary adaptations, ensuring that they did not contradict the translation and adaptation guidelines, and implementing the latest errata discovered during verification stage.

Adaptation for the above-mentioned national versions was carried out for all newly created cognitive and noncognitive materials, but also for some linking materials retrieved from PIAAC. In particular, there were no US-Spanish cognitive materials, so they were adapted from Chile-Spanish materials. By way of exception, Spain-Spanish cognitive materials were adapted from Chile-Spanish materials.

Verifiers and adapters were monitored and assisted by cApStAn staff throughout the process, liaising as needed with ETS on content and/or technical issues, before materials were delivered to BranTra for post-verification review and harmonization.

Verifiers' and adapters' suggested corrections were mostly implemented in the materials except in some cases where verifiers reported layout issues that they could not correct, or made suggestions that were better not implemented but left to the reconciler's initiative. Such exceptions were always explicitly stated in the VFF (by default, verifiers' entries in the VFFs described problems that they had corrected).

In the 2017 Field Test, the adaptation task—just like the double-translation and reconciliation task—was given to the countries. In 2017 there were two national versions that were adapted from an existing verified version: the Chile-Spanish version that was produced by adapting the Spain-Spanish version translated in 2013; and the Estonian-Russian version that was adapted from the Russian version from Russia, produced for the Education & Skills Online 2017 Field Test. The country adapter was asked to document all adaptations (i.e., changes vs. the base version) in the VFF. In addition, automated DIF reports were produced to show all differences between the base version and the adapted version. The verifier was then asked to verify that the proposed changes are

- linguistically correct;
- correctly and consistently implemented;
- in compliance with the general and item-specific guidelines.

This 'focused' verification process did not include a full sentence-by-sentence comparison against the source, as the base materials had already gone through the entire quality control process in 2013 (Spanish version for Spain) and in 2017 (Russian version for Russia).

4.6.3 Verification of change requests to PIAAC linking units

In the 2017 Field Test, the PIAAC trend items underwent a country review step where the countries had the opportunity to review the units and request changes, if they identified errors, outdated adaptations or layout issues. Such change requests were documented in a change request form (Excel file) and submitted for verification to cApStAn. The verifiers' task was to evaluate if the change request is acceptable, and if yes, implement it in the xliff file. Generally, the following kind of change requests were accepted:

- Obvious errors (typos, text left in English, mistranslations)
- Outdated adaptations
- Harmonization of recurring instructions with the new units
- Translations that did not follow the item-specific guidelines

While the following type of change requests were generally rejected:

- Changes that would make the item easier/more difficult than the international master, for example, by introducing an additional explanation when no such explanation was in the source; or by breaking a literal match between the stimulus and question stem
- Linguistically incorrect changes
- Unnecessary rephrasing of an already correct translation
- Requests that are not in compliance with the general and/or item-specific guidelines

4.7 Post-verification review and overall harmonization

After verification and adaptation, the files went through a post-verification review included in an overall harmonization procedure to ensure consistency between linking and newly developed materials. In the 2013 Field Test reconcilers, under BranTra guidance, performed this step. In the 2017 Field Test the task was assigned to the country's reconciler.

The reconciler's role in this procedure was to:

- Address verifier's comments in the VFF by accepting or discarding suggested edits in the XLIFFs: mark the verifier's comments in the VFF with an "OK" or an explanation why a suggestion was rejected and/or a change was undone.
- Ensure consistency between PIAAC and Education & Skills Online materials to harmonize recurrent elements such as form of address, recurring directions and instructions, position of currency sign, abbreviations for weights and measures, date and time formats, punctuation and other typographic conventions, and so on.
- Check that the latest errata were correctly addressed.
- Perform a "final optical check" to make sure that all residual layout issues were correctly described in the VFF and none were missed. This required using the preview function on the portal and reuploading XLIFFs whenever changes were made. To properly carry out a

complete final optical check, the reconciler was asked to act like the respondent and perform each task to ensure that all parts were translated (such as pop-up messages, reference documents, help features, etc.) and fully functioning. The outcomes of this final optical check were entered in the VFF and layout issues were highlighted for easy identification by the ETS technical team responsible for technical corrections.

After this post-verification and overall harmonization procedure, the XLIFF files went to the ETS technical team to solve any residual layout and technical bugs reported at verification and post-verification stage.

4.8 Evaluation and selective implementation of changes suggested by representatives of participating countries (2013 Field Test only)

After the steps described above, representatives of participating countries were given the opportunity to review materials before finalization and to propose changes. The focus of the task was limited to identifying errors with the translation and adaptation. ETS provided online access for each country to a secured portal with the translated versions of the items along with written instructions for navigating across the units on the portal, reviewing the translated content, and documenting comments and suggested changes. Country feedback was provided in the form of comments and suggested edits in the VFFs, sometimes complemented by annotated screenshots.

In some cases, the extent of country feedback was surprisingly high compared to other countries, which led to a discussion over the policy to be adopted by cApStAn and ETS. It was agreed that cApStAn would evaluate each request for change, categorizing it and acting on it as follows:

- Needed change (error that should be corrected): Implemented by cApStAn verifiers in both linking items and newly created materials.
- "Cosmetic" or preferential change (a change that has very little or no impact on the construct being measured): Implemented by cApStAn verifiers only in newly developed materials in order to preserve trend.
- Incorrect or risky change: Not implemented, with an explanation in the VFF.

In the 2017 Field Test, the countries were involved in the translation process early on, so this step was not part of the process.

4.9 Verification of language-dependent automated scoring rules

To verify the correct scoring of literacy items with the highlight response mode, a procedure was put in place and carried out by cApStAn using an "evaluation" widget specially created on the Education & Skills Online portal for this procedure (see the PIAAC Technical Report for more detailed information about highlighting items).

In 2013, after countries revised their units, including checking and correcting the text blocks and testing the highlight items, cApStAn carried out checks on a sample of the more difficult/sensitive items by performing a certain number of testing steps and checking that the same expected results were obtained. A sample of 19 items was selected, for which the scoring was tested in all national versions.

The scoring check included:

- checking that all text blocks were correctly and precisely defined (text block definition was done at the word level, not at the character level, as in PIAAC);
- testing all minimum correct responses (consisting of one text block);
- testing all maximum correct responses;
- testing correct responses for which more than one text block needs to be highlighted, in separate locations (on two different screens for example);
- systematically double checking some known recurring bugs discovered during the procedure in all versions.

Whenever needed, verifiers could correct imprecise text blocks, add those that were missing, and report any issue that they were unable to correct, as well as a list of interventions made, in an ad hoc Excel file sent to ETS.

Also, if any additional layout problems or residual issues at the content level were spotted during the scoring verification stage, they could be addressed before test delivery packages were finalized and delivered to the participating countries.

In 2017, countries were responsible for defining the scoring blocks. A separate webinar was organized to explain the logic of the scoring definitions and the process for defining the blocks in the target language. Once countries finished defining the scoring blocks on the portal and documenting any issues they may have had in a specially prepared Excel monitoring sheet, cApStAn verified that these scoring definitions reflected the guidelines. If any issues were detected, these were documented in the Excel monitoring form that was then delivered back to the country. If the country spotted functionality issues, these were corrected by ETS after verification and prior to delivering the verification feedback to the countries.

4.10 Translation of the score reports

Education & Skills Online also includes score reports for each of the modules presented to the test taker. These score reports were developed by ETS in English and approved by the Organisation for Economic Co-operation and Development. The score reports were sent to representatives from each of the countries that completed the Field Test for translation. Countries translated and adapted text into the country-specific language versions. The translated and adapted score reports were not reviewed or validated by ETS or cApStAn. Countries reviewed the text of the score reports when reviewing the final assessment package.

Chapter 5: Field Test Procedures and Administration

Field tests are an integral part of all large-scale assessments and surveys. Their primary purposes are to test the survey operations procedures, identify and correct poorly performing items, identify scoring issues, and examine item characteristics. Field tests also provide an opportunity to examine issues that might be associated with translation and adaptation or other survey procedures. In Education & Skills Online, the Field Tests also served to evaluate the equivalence of item parameters in relation to the International Adult Literacy Survey, the Adult Literacy and Lifeskills survey, and the Programme for the International Assessment of Adult Competencies (PIAAC). The Field Tests for Education & Skills Online were conducted in two rounds. The first Field Test was conducted in 2013 and included the following countries and language versions:

- Canada (English and French)
- Czech Republic (Czech)
- Ireland (English)
- Italy (Italian)
- Japan (Japanese)
- Spain (Spanish)
- United States (English and Spanish)

The second Field Test was conducted in 2017 and included the following countries and language versions:

- Australia (English)
- Chile (Spanish)
- Estonia (Estonian and Russian)
- Russian Federation (Russian)
- Slovak Republic (Slovak)
- Slovenia (Slovenian)

The Education & Skills Online Field Test included a fully computerized measure of cognitive and noncognitive skills. For the 2013 Field Test, ETS and Internet Testing Systems (ITS) began preparations for the delivery system in January 2013 with the development of the online delivery portal, hosted by ITS, and the import of translations for the assessment content. In the months leading up to the Field Test start date in April 2013, the delivery system was assembled and tested internally by ETS and ITS staff. Prior to administration, a second review phase took place within each of the participating countries to test administration procedures and system functionality.

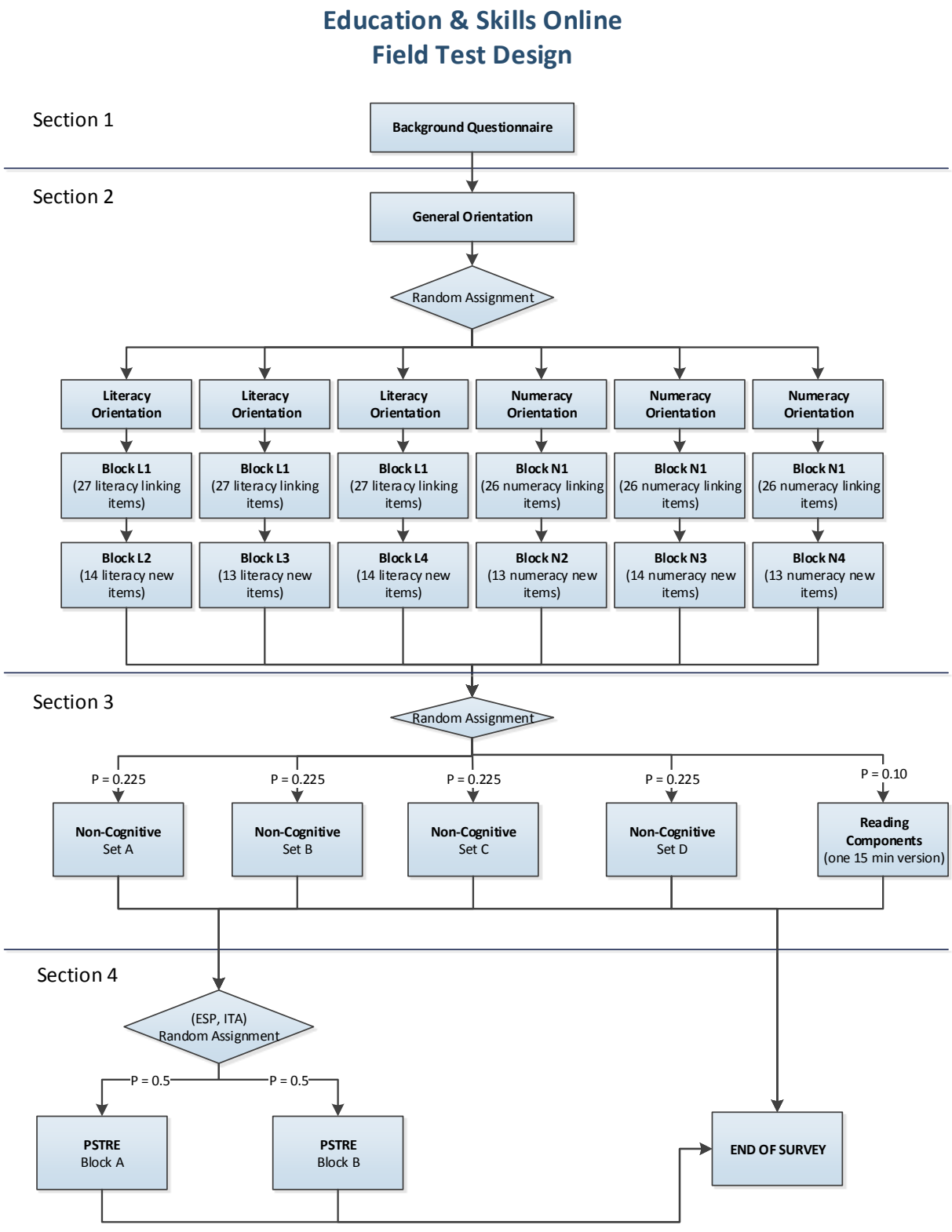
For the 2017 Field Test, ETS and ITS began preparations for the delivery system in September 2016. In the months leading up to the Field Test start date in February 2017 and May 2017, the delivery system was assembled and tested internally by ETS and ITS staff. Prior to administration, a second review phase took place within each of the participating countries to test administration procedures and system functionality.

Countries were responsible for several aspects of the Field Test. In 2013, countries were responsible for reviewing and verifying the translation of the instruments, and in 2017 countries were responsible for the translation and adaptation of the instruments (see Chapter 4). Countries also conducted tests of the delivery system, and country representatives recruited organizations and/or individuals to participate in the Field Test and monitored the progress of data collection. Country representatives worked closely with ETS staff to track completion of tests and check the distribution of completed tests within the targeted population.

5.1 Workflow of the Field Test administration – 2013 Field Test

The Education & Skills Online Field Test was designed to take approximately 90 minutes, with the exception of Italy, and Spain, where it was designed to take 120 minutes due to inclusion of the problem solving in technology-rich environments instrument. Figure 5.1 shows the workflow diagram for the Field Test. The assessment was administered in four sections. The sections were presented sequentially, beginning with the Core Background Questionnaire. For countries not administering the problem-solving blocks, there were 30 unique paths possible through the assessment. For countries administering the problem-solving blocks, there were 60 unique paths. All respondents were assigned one module within each section.

Figure 5.1: Workflow of the 2013 Field Test Administration



The second section of the assessment introduced the literacy and numeracy cognitive modules; each began with a general orientation to the functionality of the module. Respondents were then randomly assigned to one of six cognitive modules, each comprising either two literacy or two numeracy blocks. Each block contained three or more units of questions. Within each of the cognitive modules, respondents were permitted to go back to a previous question if they were still within the unit. Respondents received a confirmation message upon completion of all questions within a unit. There were approximately 40 questions in each cognitive module.

All literacy cognitive modules began with the literacy orientation, followed by two literacy blocks. Each literacy cognitive module, comprising approximately 40 questions, began with block L1. Block L1 was paired once with each of the remaining literacy blocks (L2–L4) to create the three literacy cognitive modules. All numeracy cognitive modules began with the numeracy orientation, followed by block N1. Block N1 was paired once with each of the remaining numeracy blocks (N2–N4) to create the three numeracy cognitive modules.

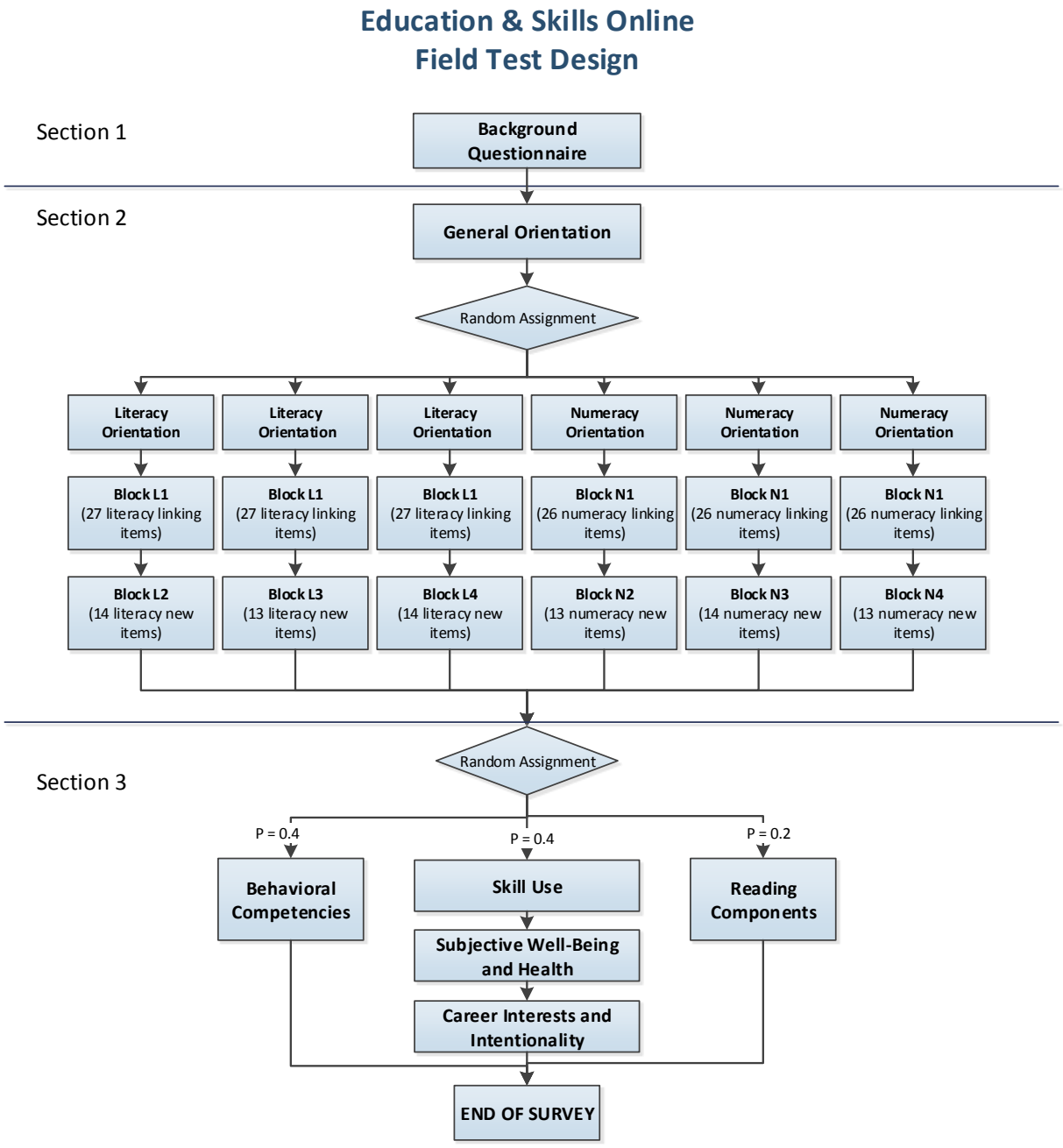
In Section 3 of the assessment, respondents were randomly assigned to one of four noncognitive modules or the reading components module. The assignment was done according to the probabilities specified in the workflow diagram. There were four noncognitive modules: Skill Use in Set A, Career Interest and Intentionality in Set B, Behavioral Competencies in Set C, and Subjective Well-Being and Health in Set D. These noncognitive modules varied in length and complexity of question layout and routing. Reading components was the fifth module included in this section.

After completing Section 3, respondents in Italy and Spain were randomly assigned to one of two problem-solving blocks (PS-A or PS-B) as shown in Section 4. For countries not administering the problem-solving blocks, after finishing the last question within the module assigned in Section 3 module, the system indicated completion of the assessment and instructed respondents to close the browser.

5.2 Workflow of the Field Test administration – 2017 Field Test

The Education & Skills Online 2017 Field Test was designed to take approximately 90 minutes. Figure 5.2 shows the workflow diagram for the Field Test. The assessment was administered in three sections. The sections were presented sequentially, beginning with the Core Background Questionnaire. There were 18 unique paths possible through the assessment.

Figure 5.2: Workflow of the 2017 Field Test Administration



Section 2 of the workflow remained the same in both 2013 and 2017 Field Tests.

The other sections of the workflow and content of the 2017 Field Test differed from the 2013 Field Test in a few key ways:

- The content of Sections 1 and 3 (the Background Questionnaire, Skill Use module, and Subjective Well-Being and Health module) were updated to only include those items that

were selected for the final product in 2015. This was done to reduce respondent burden and reduce translation burden for the participating countries. The content of the Reading Components module remained the same.

- The routing of Section 3 was updated so that respondents were randomly assigned to one of three paths rather than one of five paths. These three paths were: Behavioral Competencies module only, Reading Components module only, or the Skill Use module followed by the Subjective Well-Being and Health module and the Career Interests and Intentionality module. The probabilities of being assigned to each path are included in Figure 5.2. This update was made to obtain larger sample sizes for the non-cognitive modules. Even though respondents were administered additional modules in one of the paths, this updated routing did not impact the time it took respondents to answer all questions because the number of questions in some modules had been reduced from the 2013 Field Test.

- Section 4 was removed because all the countries participating in the 2017 Field Test had included problem solving in their PIAAC administrations.

5.3 Country recruitment and participation

Countries were responsible for managing Field Test recruitment and monitoring progress of the administration. The sample sizes were determined by ETS and varied by country and language according to the availability of item parameter information from PIAAC based on existing language versions. Countries were required to locate organizations, institutions, and/or individuals to gain cooperation for Field Test participation. The examples shown in Tables 5.1, 5.2, and 5.3 represent the desired distribution for a range of sample sizes associated with various country-language versions. Countries were asked to make an effort to achieve a balanced gender distribution in each category (represented by a cell in the table).

Table 5.1: US (English) – 2,000 cases

Labor Force Status	**Employed**			**Unemployed/Inactive**			
Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**	**Total**
High school or below	148	148	148	74	74	74	**666**
Some post-secondary education	148	148	148	74	74	74	**666**
Full university degree or above	148	148	148	74	74	74	**666**
Total	**444**	**444**	**444**	**222**	**222**	**222**	**2,000**

Table 5.2: US (Spanish), Canada (French), Italy, Czech Republic, Japan, Estonia, Russian Federation, Slovenia, Slovak Republic – 1,200 cases

Labor Force Status	Employed			Unemployed/Inactive			
Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**	**Total**
High school or below	89	89	89	44	44	44	**400**
Some post-secondary education	89	89	89	44	44	44	**400**
Full university degree or above	89	89	89	44	44	44	**400**
Total	**267**	**267**	**267**	**132**	**132**	**132**	**1,200**

Table 5.3: Canada (English), Ireland (English), Estonia (Russian), Australia (English), Chile (Spanish) – 900 cases

Labor Force Status	Employed			Unemployed/Inactive			
Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**	**Total**
High school or below	67	67	67	33	33	33	**300**
Some post-secondary education	67	67	67	33	33	33	**300**
Full university degree or above	67	67	67	33	33	33	**300**
Total	**200**	**200**	**200**	**100**	**100**	**100**	**900**

5.4 Field Test data collection

The Field Test instrument was accessible online. ETS worked with a representative within each country to identify the number of authorization codes required for each language version (where applicable). Each country received an adequate number of authorization codes, based on its sample size, which were valid for the duration of the Field Test. Countries were responsible for assigning these codes to participants and for keeping track of these individuals in case technical problems arose. Each code could be used only once. The national data file produced by the system contained only country and access codes, as well as the responses provided to the Field Test instruments. All other information about the participating organization or individuals, such

as names, addresses, or phone numbers, remained with the countries and was not accessible by the contractors.

Participants accessed the Education & Skills Online administration portal through a URL. The login screen required an initial system check to confirm computer readiness and an authorization number to start the testing session. To facilitate the administration process, ETS prepared documentation for countries to translate and disseminate to survey administrators and test takers. It contained frequently asked questions about the Education & Skills Online assessment and step-by-step instructions for accessing the online administration login site.

Data collection for the 2013 Field Test ran from April 2013 through April 2014. Data collection for the 2017 Field Test began on a staggered schedule. Three language versions: Estonian, Russian (Russia), and Spanish (Chile) began data collection in February 2017; three language versions: Slovak, Slovenian, and Russian (Estonia) began data collection in May 2017; and Australia began data collection in June 2017. Data collection ran through January 2018. Throughout the data collection period, ETS and ITS provided technical support and managed the process of generating codes and monitoring completion of codes on a weekly basis. ETS provided country representatives with weekly updates on the progress of data collection within their country during the Field Test period. This information was provided in a format that allowed managers to monitor the progress of the Field Test according to the sample design.

Tables 5.4 to 5.16 summarize the final sample sizes for each country/language.

Table 5.4: Australia-English Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**		
Level of Education	**High school or below**	43	40	27	86	24	41	261	**300**
	Some post-secondary	39	50	52	23	16	50	230	**300**
	Full university degree or above	28	47	70	8	16	45	214	**300**
	Total (actual)	**110**	**137**	**149**	**117**	**56**	**136**	**705**	
	Total (target)	**200**	**200**	**200**	**100**	**100**	**100**		**900**

Female	Male
438	267

Table 5.5: Canada-English Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**		
Level of Education	**High school or below**	39	19	14	89	64	39	264	**300**
	Some post-secondary	25	36	67	13	12	15	168	**300**

	Full university degree or above	89	145	172	13	13	9	441	**300**		
	Total (actual)	**153**	**200**	**253**	**115**	**89**	**63**	**873**		**Female**	**Male**
	Total (target)	**200**	**200**	**200**	**100**	**100**	**100**		**900**	619	254

Table 5.6: Canada-French Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	24	21	28	38	40	27	178	400		
	Some post-secondary	12	13	12	8	4	1	50	400		
	Full university degree or above	79	181	132	18	14	26	450	400		
	Total (actual)	**115**	**215**	**172**	**64**	**58**	**54**	**678**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	451	227

Table 5.7: Chile-Spanish Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	86	78	61	118	74	43	460	**300**		
	Some post-secondary	56	52	29	30	34	16	217	**300**		
	Full university degree or above	145	132	59	34	39	16	425	**300**		
	Total (actual)	**287**	**262**	**149**	**182**	**147**	**75**	**1102**		**Female**	**Male**
	Total (target)	**200**	**200**	**200**	**100**	**100**	**100**		**900**	599	503

Table 5.8: Czech Republic-Czech Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**		
Level of Education	**High school or below**	112	217	190	179	64	88	**850**	**400**
	Some post-secondary	20	23	14	8	8	2	**75**	**400**

	Full university degree or above	107	131	143	54	27	26	**488**	**400**		
	Total (actual)	**239**	**371**	**347**	**241**	**99**	**116**	**1413**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	651	762

Table 5.9: Estonia-Estonian Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	81	65	39	290	133	114	722	**400**		
	Some post-secondary	22	38	41	63	74	87	325	**400**		
	Full university degree or above	115	181	116	96	185	163	856	**400**		
	Total (actual)	**218**	**284**	**196**	**449**	**392**	**364**	**1903**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	1275	628

Table 5.10: Estonia-Russian Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	82	54	34	74	41	33	318	**300**		
	Some post-secondary	64	60	53	32	30	48	287	**300**		
	Full university degree or above	98	77	77	37	38	59	386	**300**		
	Total (actual)	**244**	**191**	**164**	**143**	**109**	**140**	**991**		**Female**	**Male**
	Total (target)	**200**	**200**	**200**	**100**	**100**	**100**		**900**	605	386

Table 5.11: Ireland-English Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**		
Level of Education	**High school or below**	52	48	68	115	30	36	**349**	**300**
	Some post-secondary	58	67	60	48	34	37	**304**	**300**
	Full university degree or above	49	93	31	23	29	17	**242**	**300**

	Total (actual)	**159**	**208**	**159**	**186**	**93**	**90**	**895**		**Female**	**Male**
	Total (target)	**200**	**200**	**200**	**100**	**100**	**100**		**900**	467	428

Table 5.12: Italy-Italian Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	55	212	238	167	78	120	**870**	**400**		
	Some post-secondary	0	18	8	1	2	4	**33**	**400**		
	Full university degree or above	35	122	58	41	30	13	**299**	**400**		
	Total (actual)	**90**	**352**	**304**	**209**	**110**	**137**	**1202**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	691	511

Table 5.13: Japan-Japanese Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	9	73	63	30	23	23	**221**	**400**		
	Some post-secondary	4	58	62	2	29	20	**175**	**400**		
	Full university degree or above	76	314	282	41	64	60	**837**	**400**		
	Total (actual)	**89**	**445**	**407**	**73**	**116**	**103**	**1233**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	411	822

Table 5.14: Russian Federation-Russian Sample

	Labor Force Status	**Employed**			**Unemployed/Inactive**			**Total Actual**	**Total Target**		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	87	76	40	195	46	46	490	**400**		
	Some post-secondary	94	91	93	67	46	56	447	**400**		
	Full university degree or above	127	143	87	124	129	89	699	**400**		
	Total (actual)	**308**	**310**	**220**	**386**	**221**	**191**	**1636**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	1008	628

Table 5.15: Slovak Republic-Slovak Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target		
	Age	16-30	31-45	46-65	16-30	31-45	46-65				
Level of Education	High school or below	309	127	106	256	71	83	952	**400**		
	Some post-secondary	13	1	5	1		7	27	**400**		
	Full university degree or above	76	164	157	24	26	30	477	**400**		
	Total (actual)	**398**	**292**	**268**	**281**	**97**	**120**	**1456**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	810	646

Table 5.16: Slovenia-Slovenian Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target		
	Age	16-30	31-45	46-65	16-30	31-45	46-65				
Level of Education	High school or below	85	104	110	68	53	53	473	**400**		
	Some post-secondary	17	33	53	33	27	30	193	**400**		
	Full university degree or above	105	187	113	66	52	42	565	**400**		
	Total (actual)	**207**	**324**	**276**	**167**	**132**	**125**	**1231**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	846	385

Table 5.10: Spain-Spanish Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target		
	Age	16-30	31-45	46-65	16-30	31-45	46-65				
Level of Education	High school or below	47	69	61	150	86	79	**492**	**300**		
	Some post-secondary	47	36	45	21	26	27	**202**	**300**		
	Full university degree or above	64	167	138	35	47	29	**480**	**300**		
	Total (actual)	**158**	**272**	**244**	**206**	**159**	**135**	**1174**		**Female**	**Male**
	Total (target)	**200**	**200**	**200**	**100**	**100**	**100**		**900**	661	513

Table 5.11: US-English Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	125	112	96	160	90	59	**642**	**666**		
	Some post-secondary	100	90	145	26	22	18	**401**	**666**		
	Full university degree or above	88	177	204	17	14	9	**509**	**666**		
	Total (actual)	**313**	**379**	**445**	**203**	**126**	**86**	**1552**		**Female**	**Male**
	Total (target)	**444**	**444**	**444**	**222**	**222**	**222**		**2000**	991	561

Table 5.12: US-Spanish Sample

	Labor Force Status	Employed			Unemployed/Inactive			Total Actual	Total Target		
	Age	**16-30**	**31-45**	**46-65**	**16-30**	**31-45**	**46-65**				
Level of Education	**High school or below**	85	50	36	168	112	21	**472**	**400**		
	Some post-secondary	13	19	10	46	17	8	**113**	**400**		
	Full university degree or above	19	31	14	9	11	2	**86**	**400**		
	Total (actual)	**117**	**100**	**60**	**223**	**140**	**31**	**671**		**Female**	**Male**
	Total (target)	**267**	**267**	**267**	**132**	**132**	**132**		**1200**	489	182

Chapter 6: Data Analysis, Scaling, and Calculation of Proficiency Values

6.1 Introduction

This chapter focuses on data analysis and scaling for the cognitive items in Education and Skills Online. Please note that a summary of the noncognitive data analysis is available Chapter 3. The cognitive modules in Education & Skills Online are based on the literacy (including reading components), numeracy, and problem solving in technology-rich environments domains from the Programme for the International Assessment of Adult Competencies (PIAAC). Several steps were taken to assure comparability of the cognitive domains in Education & Skills Online to those in PIAAC in terms of instrumentation, target populations, and survey operations.

A number of items from PIAAC were used to create a link between it and Education & Skills Online so that the two tests could be measured on a common scale. These "linking" items were selected to represent the PIAAC literacy, numeracy, and problem-solving frameworks. New items were developed for literacy and numeracy as well, based on the PIAAC frameworks.

The target population for the Education & Skills Online Field Tests was a subset of the total adult population (ages 16-65). All items were administered on computer via an Internet-based platform. The systems for test administration, scoring, and evaluating scoring accuracy were comparable to those for the computer-based PIAAC assessment. The analysis methods and procedures for Education & Skills Online were based on identical psychometric principles as used for PIAAC.

The Education & Skills Online Field Test 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 item 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 Field Test 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 from the Field Test 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) and reviewing item analyses (percent of correct responses per item) within and across countries to detect potential errors in translation or scoring. 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 Education & Skills Online 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 Education & Skills Online Field Test data for every linking item by country. Using essentially the same IRT item parameters for the linking items assured that the scale linkage of Education & Skills Online to PIAAC could be established, meaning the inferences that can be made from the PIAAC scale scores (i.e., level descriptors, representative tasks along the scale) are true for Education & Skills Online as well. To achieve this, the majority of item parameters for linking items in Education & Skills Online were the same as in PIAAC (common item parameters); only a few items needed unique item parameters in certain countries. These newly estimated item parameters were necessary in cases where the items showed no fit to the common item parameters obtained in PIAAC.

In the following sections, the data evaluation process and the scaling model used for Education & Skills Online are described.

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

The assurance of data quality was an important step prior to 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.

6.2.1 Scoring and handling of missing data

The Education & Skills Online Field Test followed the same scoring guidelines and procedures as those applied in PIAAC for the computer-based administration. The literacy and numeracy items were dichotomously scored: correct responses were scored as 1, and incorrect responses as 0. The problem-solving items received polytomous scores: partly correct or fully correct responses were scored as 1 or 2, or as 1, 2, or 3; incorrect responses were still scored as 0. For data analysis purposes, missing data were handled with a procedure similar to that used in PIAAC in order to maintain comparability between the two studies. The structure of missing responses is derived from the matrix sampling design:

1. Missing by design (scored as 9): Items that were not presented to each respondent due to the matrix sampling design used in the Education & Skills Online Field Test. Accordingly, these structural missing data, unrelated to respondents' literacy, numeracy, and problem-solving skills, were ignored when calculating respondent proficiencies.

2. Omitted responses (scored as 8): 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 (scored as 9): Missing responses at the end of the test 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.

Some respondents who answered a sufficient number of background questionnaire (BQ) questions may not have been able to respond to the cognitive items or were unwilling to respond to them. The treatment of these cases is described in the next section.

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

Some respondents completed the entire background questionnaire (BQ) but responded to only a few cognitive items. For most countries, the proportion of respondents with fewer than 5 cognitive item responses (in the literacy or numeracy domains) was less than 4 percent; within the Estonia-Estonian (ETI) and Estonia-Russian (RUE) samples, that proportion was approximately 14%.

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 skill-related reasons), and those unwilling to respond (i.e., not for skill-related reasons). The handling of the missing data in these cases followed the same procedure as described above (section 6.2.1).

In general, the data were prepared as described above, including respondents with fewer than 5 cognitive responses for analysis purpose (scores are not produced for respondents with fewer than 5 cognitive responses). The evaluations 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.

6.2.3 Item analyses

Once the data were prepared, item analyses were conducted separately for each country and each cognitive domain (literacy, numeracy, problem solving, and reading components). The purpose of the item analyses was to identify outliers or unexpected patterns that might signify issues with translations of items or scoring rules. ETS reviewed an item analysis report including the following statistics for each item and item block:

Item block:

- Statistics for the computation of the alpha reliability coefficient and standard error of measurement for the test.

- Summary statistics for the literacy and numeracy block scores (L1, L2, L3, L4, N1, N2, N3, N4). The block score is the sum of correct responses for each country.

Item response categories within block:

NOT RCH	Subjects who did not respond to or omitted the question and did not respond to any subsequent question, or for which a response is missing.
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.
0	Subjects who responded incorrectly.
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 within block:

R BIS	The R-biserial indicates the correlation between test takers' performance on an individual question and their performance on the criterion score. It is a measure of a question's power to discriminate among test takers of different abilities. A relatively high R-biserial indicates that test takers who scored higher on the criterion score were more likely than test takers 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 test takers 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. A smaller delta index indicates an easier item and a larger number indicates a difficult item. The index can vary often between 1 and 25.

The Education & Skills Online Field Test item analyses completed for all countries are summarized in Tables 6.1.a and b[1]. Overall the items performed appropriately and as expected. On average: not reached and omitted rates were 2% and 6% in literacy and 3% and 9% in numeracy; P+ values were 0.54 in literacy and 0.56 in numeracy; and R BIS value were 0.68 in literacy and 0.68 in numeracy. Some differences were observed across countries and languages. US Spanish had the highest omitted rates and the lowest average performance while Japan had the lowest not reached and omitted rates and highest average performance (highlighted in Tables 6.1.a and b).

The broad range of item difficulties permitted the selection of easy and difficult items in the building of the final version of the assessment to be able to measure the proficiency of lower as well as higher performing respondents. Through item analyses and the IRT scaling, issues could be identified and fixed for some items. Items that were found to function poorly or not to have the desired characteristics were not selected for the final instrument.

[1] Statistics were computed only for sample sizes greater or equal to 75.

Table 6.1.a: Average (standard deviation) of literacy item statistics

Country / Language	Literacy							
	% Not Rch		% Omit		P+		R Bis	
All	3	(7)	9	(8)	0.54	(0.25)	0.68	(0.19)
CSY	0	(0)	4	(5)	0.60	(0.27)	0.60	(0.22)
ENA	4	(3)	11	(5)	0.60	(0.20)	0.80	(0.16)
ENC	2	(2)	7	(5)	0.59	(0.23)	0.74	(0.21)
ENI	2	(2)	6	(5)	0.57	(0.23)	0.64	(0.15)
ENU	1	(0)	7	(5)	0.53	(0.23)	0.74	(0.16)
ESL	3	(3)	10	(9)	0.44	(0.23)	0.64	(0.12)
ESN	2	(1)	8	(6)	0.56	(0.24)	0.69	(0.19)
EST	1	(1)	28	(12)	0.25	(0.18)	0.78	(0.23)
ETI	14	(14)	5	(4)	0.65	(0.23)	0.67	(0.18)
FRC	2	(2)	8	(6)	0.57	(0.23)	0.68	(0.16)
ITA	1	(1)	14	(8)	0.44	(0.26)	0.64	(0.19)
JPN	0	(0)	4	(6)	0.61	(0.33)	0.59	(0.25)
RUE	11	(12)	7	(6)	0.58	(0.23)	0.68	(0.18)
RUS	1	(1)	8	(8)	0.50	(0.24)	0.69	(0.15)
SKY	3	(2)	7	(6)	0.50	(0.27)	0.61	(0.16)
SLV	0	(0)	7	(7)	0.56	(0.23)	0.65	(0.15)

Table 6.1.b: Average and (standard deviation) of numeracy item statistics

Country / Language	Numeracy							
	% Not Rch		% Omit		P+		R Bis	
All	2	(3)	6	(6)	0.56	(0.27)	0.68	(0.20)
CSY	0	(0)	1	(2)	0.68	(0.24)	0.65	(0.17)
ENA	2	(2)	10	(4)	0.58	(0.22)	0.78	(0.14)
ENC	2	(2)	5	(3)	0.60	(0.25)	0.74	(0.20)
ENI	2	(1)	3	(3)	0.56	(0.26)	0.66	(0.23)
ENU	0	(0)	6	(4)	0.45	(0.25)	0.75	(0.26)
ESL	3	(2)	9	(8)	0.49	(0.25)	0.67	(0.15)
ESN	1	(1)	5	(4)	0.57	(0.26)	0.66	(0.14)
EST	1	(1)	12	(9)	0.27	(0.27)	0.56	(0.39)
ETI	7	(7)	4	(6)	0.68	(0.22)	0.68	(0.15)
FRC	2	(1)	5	(4)	0.59	(0.26)	0.72	(0.16)
ITA	0	(1)	10	(8)	0.47	(0.28)	0.60	(0.20)
JPN	0	(0)	1	(3)	0.75	(0.22)	0.69	(0.20)
RUE	5	(4)	4	(5)	0.66	(0.24)	0.69	(0.17)
RUS	1	(1)	7	(4)	0.55	(0.22)	0.74	(0.15)
SKY	1	(1)	5	(5)	0.57	(0.27)	0.63	(0.16)
SLV	0	(0)	5	(5)	0.62	(0.24)	0.69	(0.13)

6.3 IRT scaling: Estimation of item parameters

The IRT scaling provided the estimations of item parameters and the proficiency distribution of the population and was carried out separately for the domains of literacy, numeracy, and problem solving (no IRT analyses were computed for reading components). Similar to PIAAC, Education & Skills Online used the two-parameter logistic model (2PL; Birnbaum, 1968) for dichotomously scored responses (literacy, numeracy) and the general partial credit model (GPCM; Muraki, 1992) for polytomous data (problem solving). As noted above, incorrect responses were coded as 0, correct responses in the 2PL model as 1 or 2, or 1, 2, or 3 in the GPCM; omitted responses were treated as incorrect responses, and missing at random as missing values.

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 (θj) of the respondent (j) and two item parameters characterizing the properties of the item (item difficulty βi and item discrimination αi).

The *GPCM*, like the 2PL model, is a mathematical model for the probability that an individual will respond in a certain response category on a particular item. While the 2PL model is suitable for dichotomous responses only, the GPCM can be used with polytomous and dichotomous responses considering m_i+1 ordered response categories for an item *i*. The GPCM reduces to the 2PL model when applied to dichotomous responses.

For more details about the models and IRT scaling process, see the PIAAC Technical Report (OECD, 2013, chapter 17, pp. 2–6).[2]

Because Education & Skills Online used items from PIAAC to link the two surveys, those linking items were analyzed to see if they did, in fact, work similarly in Education & Skills Online. For this, the item parameters in the IRT scaling of the Education & Skills Online 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 linking items in the assessment. Because the sample size for some countries was rather small, the data from countries were compiled into language groups for the domains of literacy and numeracy in order to provide more reliable parameter estimates, while the data from two countries were available for problem solving (see Tables 6.2 and 6.3).

Item parameters for new Education & Skills Online items based on the PIAAC frameworks were estimated independently from the PIAAC item parameters by setting equality constraints so that the item parameters for every item were estimated equally in each language group (common Education & Skills Online item parameters for new items). Concurrent calibration was used to determine if items worked comparably in each language group or if there were differences due to scoring, translation or other issues.

[2] The specific technical report chapter can be found at http://www.oecd.org/site/piaac/PIAAC%20Tech%20Report_Section%205_update%201SEP14.pdf

Table 6.2: Sample sizes of participating countries and language groups in the Education & Skills Online Field Test for the domains of literacy and numeracy

	Language Group used in IRT Analyses	Country (Language Code Appearing in Data)	n per Country	n per Language Group
1	Czech	Czech Republic (CSY)	1,413	1,413
2	English	Canada (ENC)	873	4,022
		USA (ENU)	1,552	
		Ireland (ENI)	895	
		Australia (ENA)	702	
3	Spanish	Spain (ESN)	1,174	2,947
		USA (EST)	671	
		Chile (ESL)	1,102	
4	French	Canada (FRC)	678	678
5	Italian	Italy (ITA)	1,202	1,202
6	Japanese	Japan (JPN)	1,233	1,233
7	Estonian	Estonia (ETI)	1,903	1,903
8	Russian	Russia (RUS)	1,636	2,627
		Estonia (RUE)	991	
9	Slovak	Slovak Republic (SKY)	1,456	1,456
10	Slovenian	Slovenia (SLV)	1,231	1,231

Table 6.3: Sample sizes of participating countries and language groups in the Education & Skills Online Field Test for the domains of problem solving

	Language Group used in IRT Analyses	Country	n per Country/Language Group
1	Spanish	Spain (ESN)	1,174
2	Italian	Italy (ITA)	1,202

Linking items in the Education & Skills Online scaling that showed deviations from the common PIAAC item parameters were assumed to work differently in Education & Skills Online, meaning they would harm the link to PIAAC. Similarly, new items in the Education & Skills Online scaling that showed deviations from the common Education & Skills Online item parameters (parameters were estimated equally in all language groups) were assumed to work differently in certain language groups. To examine deviations in the IRT scaling, so-called item fit statistics were used to test the fit of the model for each item. Like PIAAC, Education & Skills Online used the root mean squared deviation (RMSD) and the mean deviation (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 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 the 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 ICCs in Education & Skills Online (for linking items and new items) were revealed using an RMSD > 0.15, and 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 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 group-specific unique item parameters were estimated in a second step (unique meaning that item parameters are unique for one country or a small group of countries).

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

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 and the GPCM were estimated as multiple-group IRT models allowing for different sample distributions (one for each language group). Each distribution was modeled using three moments (mean, standard deviation and skewness) updated at each iteration during IRT calibration.

In the Education & Skills Online analyses in most cases, the linking item responses across language groups were accurately described by the common PIAAC item parameters, and most new items received common item parameters across the language groups. The deviation pattern of linking items from the common PIAAC item parameters as well as the deviation pattern of new items from the common Education & Skills Online item parameters was not consistent for any one particular language group. Table 6.4 provides the number of unique item parameters per language group for items in the final test (see appendix for detailed information).

[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 6.4: Number of unique item parameters for each language group and proficiency scale

Language-Group	Number of Group-Specific Item Parameters		
	Literacy (40 items)	Numeracy (38 items)	Problem Solving (9 items)
Czech	5	2	0
English	2	2	0
English (Australia)	3	0	0
Estonian	8	3	0
French	6	1	0
Italian	3	1	0
Japanese	8	9	0
Russian	10	7	0
Russian (Estonia)	3	1	0
Slovak	5	5	1
Slovenian	7	1	0
Spanish	1	1	0
Spanish (Chile)	2	4	0

6.4 Estimation of proficiencies used for routing

Education & Skills Online uses a respondent's demographic information to calculate a starting estimate of literacy and numeracy proficiency which determines the difficulty of the first-stage set of items the respondent answers in the assessment. The answers to the first-stage set of items and the estimate of literacy or numeracy proficiency are used to determine the difficulty of the second-stage set of items the respondent answers. The routing algorithm uses beta weights (regression coefficients) for specific demographic categories to calculate the estimated proficiency. The demographic categories considered are age, education, employment status, and whether or not the respondent was born in the country of the test. The beta weights were calculated for each country using data from the PIAAC assessment by regressing dummies for the demographic variables on the PIAAC proficiency values. The estimated beta weights were then confirmed using data from the E&S Online field test by regressing the same demographic variables on the WLE proficiency estimate from the field test administration. When confirming the beta weight estimates using the E&S Online field test data, only those respondents who answered at least five cognitive items were included in the calculation.

6.5 Calculation of proficiency values for individual test takers (weighted likelihood estimation)

Based on the item parameters estimated in the IRT scaling (see the above section), individual test scores for the ability of test takers can be estimated for each test taker using the information from his or her responses on the test. For each cognitive domain (Literacy, Numeracy, PSTRE), a

separate test score is estimated. The fewer the number of items in a domain, the higher the potential for bias in estimates of ability and the lower the measurement reliability. The most common estimators of latent ability (θ) are maximum likelihood (ML) and expected a posteriori (EAP). The former does not incorporate any bias correction, whereas the latter is a Bayesian approach that shrinks estimates toward the mean as a function of score reliability. As an alternative to EAPs, Warm (1989) proposed a weighted likelihood estimator for dichotomously scored responses that essentially serves as a bias-corrected ML estimator. David Firth (1992, 1993) proposed a more general approach using the Jeffrey's prior, and Samejima (1998) expanded Warm's approach to general discrete responses. Park and Muraki (2003), von Davier (1997, 2001), and Penfield and Bergeron (2005) extended Warm's methodology to polytomous models such as the partial credit model (Masters, 1982) and generalized partial credit model.

The maximum likelihood estimate of θ for a given individual is equal to the value of θ that maximizes the log likelihood, $L = L(\theta)$, of the associated response pattern given a fixed set of item parameters. This estimate is obtained iteratively through the use of an iterative maximization. In the Newton-Raphson algorithm, for example, the estimate at iteration *t* is equal to

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{L'}{L''} \tag{1}$$

In Equation (1), L' and L'' are given by

$$L' = \sum_{i=1}^{N} \sum_{j=0}^{J} u_{ij} D a_i (j - \lambda_1), \tag{2}$$

$$L'' = -\sum_{i=1}^{N} D^2 a_i^2 (\lambda_2 - \lambda_1^2), \tag{3}$$

where $\lambda_k = \sum_{j=0}^{J} j^k P_{ij}$ and P_{ij} is the expected response probability from the IRT model. Under this formulation $\lambda_1 = \sum_{j=0}^{J} j P_{ij}$ and $\lambda_2 = \sum_{j=0}^{J} j^2 P_{ij}$. The standard error for $\hat{\theta}$ is equal to

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I}} \tag{4}$$

where $I = I(\theta)$ is the information of the test at θ, and is computed as

$$I = \sum_{i=1}^{N} D^2 a_i^2 (\lambda_2 - \lambda_1^2) \tag{5}$$

Extending this approach, the weighted likelihood estimator of θ is based on maximizing a penalized likelihood function. In the case of exponential families, this penalty function takes the form of the square root of the information function (Warm, 1989; Firth, 1992) and the weighted likelihood to be maximized is

$$L^*(\theta) = L(\theta)\sqrt{I(\theta)} \tag{6}$$

so that the weighted log likelihood equals

$$\log[L^*(\theta)] = \log[L(\theta)] + \frac{1}{2}\log[I(\theta)]. \tag{7}$$

Therefore, the weighted likelihood estimation $\hat{\theta}_t$ at iteration t is equal to

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{W'}{W''} = \hat{\theta}_{t-1} - \frac{L' + B'}{L'' + B''} \tag{8}$$

where W is the weighted log likelihood (i.e., the bias-corrected log-likelihood, Penfield and Bergeron, 2005) and B' and B'' are given by

$$B' = \frac{\sum_{i=1}^{N} D a_i^3(\lambda_3 - 3\lambda_1\lambda_2 + 2\lambda_1^3)}{2\sum_{i=1}^{N} a_i^2(\lambda_2 - \lambda_1^2)} \tag{9}$$

$$B'' = \frac{AB - 2C^2}{B^2} \tag{10}$$

where

$$A = \sum_{i=1}^{N} D^2 a_i^4(\lambda_4 - 4\lambda_1\lambda_3 - 3\lambda_2^2 + 12\lambda_1^2\lambda_2 - 6\lambda_1^4) \tag{11}$$

$$B = 2\sum_{i=1}^{N} a_i^2(\lambda_2 - \lambda_1^2) \tag{12}$$

$$C = \sum_{i=1}^{N} D a_i^3(\lambda_3 - 3\lambda_1\lambda_2 + 2\lambda_1^3) \tag{13}$$

Because B is proportional to the likelihood, it cannot be estimated directly (Warm, 1989). As such, an iterative maximization method is required. The standard error is the same as that obtained for the ML estimate.

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.

Firth, D. (1992) Bias reduction, the Jeffrey's prior and GLIM. In L. Fahrmeir, B. J. Francis, R. Gilchrist and G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling* (pp 91-100). New York, NY: Springer.

Firth, D. (1993). Bias reduction of maximum likelihood estimates [Correction: 95V82 p667]. *Biometrika, 80*, 27-38.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177.

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

Park, C., & Muraki, E. (2003). Bias of ability estimates using Warm's weighted likelihood estimator (WLE) in the generalized partial credit model (GPCM). In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New Developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society IMPS2001*. Osaka, Japan, July 15–19, 2001 (pp. 199-206). Tokyo: Springer Japan.

Penfield, R. D., & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement, 29*, 218-233.

Samejima, F. (1998). Expansion of Warm's weighted likelihood estimator of ability for three-parameter logistic model to general discrete responses. Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego).

von Davier, M. (1997, updated 2001). *Winmira user manual*. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften [Institute of Educational Sciences].

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427-450.

Appendix

Table 6A.1: Items per domain and language group that received group-specific item parameters in the IRT scaling

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

LITERACY														
Item-ID	**Link to PIAAC**	**Czech**	**English**	**English (Australia)**	**Estonian**	**French**	**Italian**	**Japanese**	**Russian**	**Russian (Estonia)**	**Slovak**	**Slovenian**	**Spanish**	**Spanish (Chile)**
C300AC02	link	*	*	*	*	*	*	*	*	*	*	*	X	X
C301AC05	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C302BC02	link	X	*	*	*	X	*	O	*	*	*	*	*	*
C305A215	link	*	*	*	*	*	X	*	*	*	*	*	*	*
C305A218	link	*	*	*	*	X	*	*	O	*	*	*	*	*
C305S001	new	*	*	*	X	*	*	*	X	*	*	X	*	*
C305S003	new	*	*	*	X	*	*	O	Y	*	*	Y	*	*
C311B701	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C320P001	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C320P003	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C320P004	link	X	*	*	*	*	*	*	*	*	*	*	*	*
C321P001	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C321P002	link	X	*	*	*	*	*	*	*	*	X	X	*	*
C322P001	link	*	*	*	*	*	*	*	*	*	X	*	*	*
C322P002	link	*	*	*	*	*	*	X	*	*	*	*	*	*
C322P003	link	X	X	*	O	Y	*	*	Z	*	*	A	*	*
C322P004	link	*	*	*	X	O	*	Y	*	*	*	Z	*	*
C322P005	link	*	*	*	*	*	*	*	X	X	*	*	*	*
C323P002	link	*	*	*	*	X	*	X	*	*	*	*	*	*
C323P004	link	*	*	*	*	*	X	*	O	*	*	*	*	*
C329P002	link	*	X	*	O	*	*	Y	*	*	*	*	*	*
C329P003	link	*	*	*	X	*	*	*	O	O	*	Y	*	*
C400S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C400S002	new	*	*	*	*	*	*	X	*	*	O	*	*	*
C400S003	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C401S002	new	*	*	*	*	*	*	*	*	*	X	*	*	*
C401S003	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C403S001	new	*	*	*	*	*	X	*	O	*	*	*	*	*
C403S004	new	*	*	*	*	*	*	*	*	*	*	*	*	*

LITERACY														
Item-ID	**Link to PIAAC**	**Czech**	**English**	**English (Australia)**	**Estonian**	**French**	**Italian**	**Japanese**	**Russian**	**Russian (Estonia)**	**Slovak**	**Slovenian**	**Spanish**	**Spanish (Chile)**
C403S005	new	*	*	*	*	*	*	*	X	*	*	*	*	*
C404S002	new	*	*	*	X	O	*	*	Y	*	*	*	*	*
C405S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C405S002	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C405S003	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C407S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C407S003	new	*	*	*	X	*	*	O	*	X	X	X	*	X
C407S004	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C407S005	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C409S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C409S002	new	X	*	*	*	*	*	*	*	*	*	*	*	*

NUMERACY														
Item-ID	**Link to PIAAC**	**Czech**	**English**	**English (Australia)**	**Estonian**	**French**	**Italian**	**Japanese**	**Russian**	**Russian (Estonia)**	**Slovak**	**Slovenian**	**Spanish**	**Spanish (Chile)**
C600AC04	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C601AC06	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C602A502	link	X	*	*	*	*	*	*	O	*	*	*	*	*
C602A503	link	*	*	*	*	*	*	X	O	*	*	*	*	*
C611A516	link	*	*	*	*	*	*	X	O	*	*	*	*	Y
C611A517	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C612A518	link	*	X	*	*	*	*	*	O	*	*	*	*	*
C614A601	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C618A607	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C618A608	link	*	*	*	*	*	*	X	*	*	*	*	*	*
C620A610	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C620A612	link	*	*	*	*	*	*	X	*	*	O	*	*	*
C622A615	link	*	*	*	*	*	*	X	*	*	*	*	*	*
C623A616	link	*	*	*	*	*	*	*	*	*	*	*	X	X
C623A617	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C632P001	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C635P001	link	*	*	*	*	*	*	*	*	*	*	*	*	*
C636P001	link	*	*	*	*	*	*	X	O	O	*	*	*	*
E645001S	link	*	*	*	*	*	*	*	*	*	*	*	*	*

NUMERACY														
Item-ID	**Link to PIAAC**	**Czech**	**English**	**English (Australia)**	**Estonian**	**French**	**Italian**	**Japanese**	**Russian**	**Russian (Estonia)**	**Slovak**	**Slovenian**	**Spanish**	**Spanish (Chile)**
C651P002	link	*	*	*	*	*	*	X	*	*	O	*	*	*
C655P001	link	*	X	*	*	*	*	*	*	*	*	*	*	O
C700S001	new	*	*	*	*	*	X	*	*	*	*	*	*	*
C700S002	new	X	*	*	*	*	*	*	*	*	*	*	*	O
C701S001	new	*	*	*	X	*	*	*	*	*	*	*	*	*
C701S002	new	*	*	*	*	*	*	*	*	*	X	*	*	*
C701S003	new	*	*	*	X	*	*	*	*	*	*	*	*	*
C702S002	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C702S003	new	*	*	*	X	O	*	*	*	*	*	*	*	*
C704S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C704S002	new	*	*	*	*	*	*	X	*	*	O	*	*	*
C707S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C710S001	new	*	*	*	*	*	*	X	*	*	*	*	*	*
C711S002	new	*	*	*	*	*	*	*	X	*	*	*	*	*
C713S001	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C713S002	new	*	*	*	*	*	*	*	*	*	*	*	*	*
C713S003	new	*	*	*	*	*	*	*	X	*	*	*	*	*
C714S002	new	*	*	*	*	*	*	*	*	*	*	X	*	*
C714S003	new	*	*	*	*	*	*	*	*	*	X	*	*	*

PROBLEM SOLVING														
Item-ID	**Link to PIAAC**	**Czech**	**English**	**English (Australia)**	**Estonian**	**French**	**Italian**	**Japanese**	**Russian**	**Russian (Estonia)**	**Slovak**	**Slovenian**	**Spanish**	**Spanish (Chile)**
U01A000P	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U01B000S	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U02X000P	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U03A000S	link	*	*	*	*	*	*	*	*	*	X	*	*	*
U07X000S	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U11B000P	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U19A000S	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U19B000P	link	*	*	*	*	*	*	*	*	*	*	*	*	*
U21X000S	link	*	*	*	*	*	*	*	*	*	*	*	*	*

Chapter 7: Cognitive Modules Score Reporting

7.1 Introduction

Education & Skills Online provides test takers and test administrators with individual-level results for the cognitive modules that are tied to the PIAAC proficiency scales. This chapter describes the proficiency scales for the core cognitive modules. A discussion of how individual scores are estimated is included in Chapter 6.

Proficiency scales for literacy, numeracy, and problem solving in technology-rich environments range from 0 to 500 and are designed so the scores represent degrees of proficiency in a particular aspect of the domain. Scores are reported in 10-point increments and are grouped into five levels. There are easier and harder tasks for each proficiency scale.[1] Each scale is divided into proficiency levels based on the knowledge and skills required to complete the tasks within those levels. Below Level 1 represents the lowest level of proficiency, while each succeeding level represents higher proficiency. For reading components, scores are reported as either low, medium, or high in each of the skill areas assessed in the module.

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.

The Programme for the International Assessment of Adult Competencies proficiency scales used in Education & Skills Online were defined by the PIAAC Expert Groups in December 2012 and January 2013. For a complete list of experts in these groups, please see the PIAAC Technical Report (Appendix 6).[2]

Copies of the score reports are included in Appendix B. In addition, test administrators may download result information from the Administration Portal. A list of the fields included in the Administration Portal's data download is in Appendix C. For specific information on how the weighted likelihood estimate (WLE) score is computed for literacy, numeracy, and PSTRE, please see Chapter 6, section 6.4.

[1] See Appendix A for the complete list of Education & Skills Online items in each domain organized by difficulty.

[2] The PIAAC Technical Report appendices are located at: http://www.oecd.org/site/piaac/Technical%20Report_Part%206.pdf.

7.2 Literacy

The Education & Skills Online literacy items were developed and selected to represent three major aspects of processing continuous and noncontinuous texts and documents: accessing and identifying, integrating and interpreting, and reflecting on and evaluating information.

- *Access and identify* tasks require the reader to locate information in a text or document. While some tasks can be relatively straightforward because the information requested in the question matches clearly with information that is easily located in the text, not all tasks in this category are necessarily easy. Inferences may need to be made and rhetorical understanding may be required.

- *Integrate and interpret* tasks require the reader to relate different parts of the text to each other. Requiring respondents to compare and contrast, understand problems and solutions, and identify cause/effect relationships are examples of this task type. These relationships may be explicitly signaled (e.g., the text states "the cause of X is Y") or may require the reader to make inferences. The text components to be related may be contiguous and therefore easier to locate and integrate, or may be found in different paragraphs in the same text or in separate documents.

- *Evaluate and reflect* tasks require the reader to draw on knowledge, ideas, or values external to the text. The reader must assess the relevance, credibility, argumentation, and truthfulness of the information presented in the text within a context of information that is not present in the text. The reader may also evaluate the purposefulness, register, structure or reader awareness of the text, or the success with which the author uses evidence and language to argue or persuade. Tasks of this type were judged to be particularly important to include in the context of Education & Skills Online's digital texts, where it is readers must be alert to a text's accuracy, reliability, and timeliness.

The PIAAC literacy framework, which was used to develop Education & Skills Online, defined features of stimulus texts and tasks that were anticipated to impact the difficulty of tasks included in the assessment.[3] These included the following:

- Transparency of information in the text as it relates to the presented task or question

- Degree of complexity necessary to make required inferences

- Semantic and syntactic complexity of the text and/or question

- Amount of text that must be processed

- Prominence of needed information in the text

[3] For the full text of the PIAAC Literacy Framework, see Chapter 3 of Organisation for Economic Co-operation and Development (2012) http://www.oecd.org/site/piaac/PIAAC%20Framework%202012--%20Revised%2028oct2013_ebook.pdf.

- Competing information in the text
- Text features that facilitate or hinder understanding relationships among parts of the text

The literacy proficiency scale is defined in terms of five levels (this differs from PIAAC's six levels due to the merging of Levels 4 and 5 for Education & Skills Online). Levels may have some shared properties. Tasks were placed along each scale so that someone at that point on the scale would have a 67 percent chance of answering that item correctly, referred to as an RP67. Stated another way, the average person within each level would be expected to get 67 percent of the items within that level correct. In all, the literacy scale includes 40 tasks ranging in difficulty from an RP67 of 75 to 500. Those tasks are distributed by level as follows:

- Below Level 1 (1 – 175): 2 tasks
- Level 1 (176 – 225): 2 tasks
- Level 2 (226 – 275): 7 tasks
- Level 3 (276 – 325): 17 tasks
- Level 4/5 (326 – 500): 12 tasks

In the score report, test takers receive a numerical score, which is rounded to the nearest 10 points. Because the scores at the extreme ends of the scale are less precise, no test taker will receive a score below 150 or above 400. The score report also includes a description of the proficiency level of the test taker. Each of the five proficiency levels included in the score reports is defined below. Test administrators will be able to download in an Excel spreadsheet the test taker's score as well as the start and end date and time for the module and which testlets were administered to the test taker.

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 texts were noncontinuous. Tasks below Level 1 do not make use of any features specific to digital texts.

Adults in this level are able to locate specific information from a text with a few sentences or paragraphs about familiar topics. For example, they are likely able to:

- Locate a phone number or address of a store from a newspaper advertisement
- Locate the date and time of a community art show from a flyer

- Identify the winner of an employee contest from a company announcement
- Identify key ingredients from a food package label

They might sometimes have trouble using literacy skills to understand longer unfamiliar texts or to complete a form. For example, they might find it challenging to:

- Complete a short form to order a magazine subscription
- Submit a vote for or against a new workplace dress code on an employer's Web page
- Locate the link on a theater's website that would be used to find information about the theater
- Use a table in a newspaper article to identify the top three companies with the most employees
- Name two reasons stated in a newspaper article for an increase in local food prices
- Use a music store's Web page to compare and contrast several reviews to determine which song to download based on the price and the type of music one likes

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 that is identical to or synonymous with the information given in the question or directive. Some tasks, in the case of some noncontinuous text, may require the respondent to enter personal information into a document. 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 paragraph text is expected.

Adults in this level are typically able to understand longer texts about familiar topics. For example, they are likely able to:

- Identify key ingredients from a food package label
- Complete a short form to order a magazine subscription
- Submit a vote for or against a new workplace dress code on an employer's Web page
- Locate the link on a theater's website that would be used to find information about the theater
- Use a table in a newspaper article to identify the three companies with the most employees

They might sometimes have trouble understanding longer and more complicated texts. For example, they might find it challenging to:

- Determine what forms are needed to return a damaged telephone according to instructions in the warranty brochure
- Identify information in a camera store's single Web page that explains how this year's photo contest rules differ from those in previous years
- Name two reasons stated in an employee newsletter for an increase in company sales
- Use a music store's Web page to compare and contrast several reviews to determine which song to download based on the price and the type of music one likes

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
- Navigate within digital texts to access and identify information from various parts of a document

Adults in this level are typically able understand longer and more complicated texts about unfamiliar topics. For example, they are likely able to:

- Submit a vote for or against a new workplace dress code on an employer's Web page
- Determine what forms are needed to return a damaged telephone according to instructions in the warranty brochure
- Identify information in a camera store's single Web page that explains how this year's photo contest rules differ from those in previous years
- Name two reasons stated in an employee newsletter for an increase in company sales

They might sometimes experience frustration understanding longer and more complicated digital and printed texts with a variety of text features. For example, they might find it challenging to:

- Find out whether a utility company accepts the same type of payment if paid by mail or online using information from a monthly billing statement

- Use a music store's Web page to compare and contrast several reviews to determine which song to download based on the price and the type of music one likes

- Search several Web pages of a national health organization for evidence supporting the claim that exercise can lead to greater work productivity

- Determine which parents in a childcare discussion forum share a similar viewpoint by comparing their comments

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 inference. Many tasks require the respondent to 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.

Adults in this level are typically able to understand longer and more complicated digital and print texts with a variety of text features. For example, they are likely able to:

- Name two reasons stated in an employee newsletter for an increase in company sales

- Find out whether a utility company accepts the same type of payment if paid by mail or online using information from a monthly billing statement

- Use a music store's Web page to compare and contrast several reviews to determine which song to download based on the price and the type of music one likes

- Search several Web pages of a national health organization for evidence supporting the claim that exercise can lead to greater work productivity

They might find it challenging to:

- Use online search results for books on alternative energy to identify a book that includes arguments both for and against solar energy

- Evaluate posts in a discussion forum on health remedies by comparing the information they provide against that in a website from a well-known medical center

- Use several links in a city's transportation Web page to locate information about special fares or services on holidays

- From a list of workplace safety suggestions, determine which a company will be likely to adopt based on a complex chart showing the company's existing policies and procedures

Literacy Level 4/5 **326 to 500**

Tasks at the lower end of 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.

At the higher end of 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.

Adults at this level are able to use literacy skills to understand longer and more complicated texts from a number of different sources. For example, they are likely able to:

- Find out whether a utility company accepts the same type of payment if paid by mail or online using information from a monthly billing statement

- Use a music store's Web page to compare and contrast several reviews to determine which song to download based on the price and the type of music one likes

- Search several Web pages of a national health organization for evidence supporting the claim that exercise can lead to greater work productivity

- Evaluate posts in a discussion forum on health remedies by comparing the information they provide against that in a website from a well-known medical center

- Use several links in a city's transportation Web page to locate information about special fares or services on holidays

- Determine which claims in a newspaper article about the benefits of sleep are supported by information and graphs in two long research article

7.3 Numeracy

The PIAAC numeracy framework, which was used for Education & Skills Online, includes a definition of the domain as well as a description of numerate behavior.[4] Numeracy tasks were developed to cover a range of difficulty as a result of combining variables that include:

- The kind and degree of interpretation and reflection required by the problem
- The kind of representation skills required
- The kind and level of mathematical skill required (e.g., single-step vs. multistep problems, or more advanced mathematical knowledge, complex decision making, and problem-solving and modeling skills)
- The kind and degree of mathematical argumentation required
- The degree of familiarity with the context
- The extent to which tasks require reproduction of known procedures and steps or present novel situations requiring nonroutine and perhaps more creative responses

The numeracy proficiency scale is defined in terms of five levels (this differs from PIAAC's six levels due to the merging of Levels 4 and 5 for Education & Skills Online) and includes 38 tasks with difficulty values ranging from 129 to 384. Based on RP67, these tasks are distributed by level as follows:

- Below Level 1 (1 – 175): 2 tasks
- Level 1 (176 – 225): 4 tasks
- Level 2 (226 – 275): 11 tasks
- Level 3 (276 – 325): 13 tasks
- Level 4/5 (326 – 500): 8 tasks

In the score report, test takers receive a numerical score, which is rounded to the nearest 10 points. Because the scores at the extreme ends of the scale are less precise, no test taker will receive a score below 150 or above 400. The score report also includes a description of the proficiency level of the test taker. Each of the six proficiency levels is defined below. Test administrators will be able to download in an Excel spreadsheet the test taker's score as well as the start and end date and time for the module and which testlets were administered to the test taker.

[4] For the full text of the PIAAC Numeracy Framework, see Chapter 4 of OECD (2012) http://www.oecd.org/site/piaac/PIAAC%20Framework%202012--%20Revised%2028oct2013_ebook.pdf.

Numeracy Below Level 1 **0 to 175**

Tasks at this level are set in concrete, familiar contexts where the mathematical content is explicit with little or no text or distractors and require only simple processes such as counting, sorting, performing basic arithmetic operations with whole numbers or money, or recognizing common spatial representations.

Adults at this level are typically able to do simple arithmetic in familiar situations. For example, they are likely able to:

- Figure out how much money it will cost to buy a few common items in the grocery store
- Identify the amount that corresponds to an unlabeled mark on a measuring cup
- Find the range in daily temperatures by subtracting the lowest from the highest temperature

They may have trouble using numeracy skills that require computing with percents and decimal numbers, or understanding mathematical information in a table. For example, they might find it challenging to:

- Figure out the price of a shirt that will be discounted by 25 percent
- Determine the price of a single bottle of water when given the cost of an entire case of bottles
- Look at a weekly timesheet to find out which employee worked the most hours in a single day

Numeracy Level 1 **176 to 225**

Tasks in this level require the respondent to carry out basic mathematical processes in common, concrete contexts where the mathematical content is explicit with little text and minimal distractors. Tasks usually require simple one- or two-step processes involving, for example, performing basic arithmetic operations; understanding simple percents such as 50%; or locating, identifying, and using elements of simple or common graphical or spatial representations.

Adults at this level are typically able to compute with percents and decimal numbers, or understand mathematical information in a table. For example, they are likely able to:

- Identify the amount that corresponds to an unlabeled mark on a measuring cup
- Find the range in daily temperatures by subtracting the lowest from the highest temperature

- Figure out the price of a shirt that will be discounted by 25 percent
- Determine the price of a single bottle of water when given the cost of an entire case of bottles

They may have trouble using numeracy skills that require performing an intermediate computation before being able to answer a question, having to interpret a graph, or using ratios. For example, some adults with scores similar to yours might find it challenging to:

- Determine how many months in a year had sales above the mean sales for the year from a table of monthly sales
- Identify which predicted monthly gasoline price was most accurate based on line graphs of predicted and actual gasoline prices for a year
- Determine the amount of concentrated lemonade flavoring and water needed to make a large container of lemonade that is in the same ratio of flavoring to water as a smaller amount of lemonade

Numeracy Level 2 **226 to 275**

Tasks in this level require the respondent to identify and act upon mathematical information and ideas embedded in a range of common contexts where the mathematical content is fairly explicit or visual with relatively few distractors. Tasks tend to require the application of two or more steps or processes involving, for example, calculation with whole numbers and common decimals, percents, and fractions; simple measurement and spatial representation; estimation; and interpretation of relatively simple data and statistics in texts, tables, and graphs.

Adults at this level are typically able to perform an intermediate computation before being able to answer a question, understand mathematical information in a table, or interpret a simple graph. For example, they are likely able to:

- Figure out the price of a shirt that will be discounted by 25 percent
- Determine the price of a single bottle of water when given the cost of an entire case of bottles
- Determine how many months in a year had sales above the mean sales for the year from a table of monthly sales
- Identify which predicted monthly gasoline price was most accurate based on line graphs of predicted and actual gasoline prices for a year

They may have trouble using numeracy skills that require using ratios, reading a complex graph, or comparing changes in percentages. For example, they might find it challenging to:

- Determine the amount of concentrated lemonade flavoring and water needed to make a large container of lemonade that is in the same ratio of flavoring to water as a smaller amount of lemonade

- Read a complex graph, comparing the amount of salt, sugar, and fat in a typical diet for men versus a typical diet for women, to determine the amount of sugar consumed by men

- Convert the number of students enrolled in a university each year into percentages, and then compute the change in the percentage of students enrolled each year

Numeracy Level 3 **276 to 325**

Tasks in this level require the respondent to understand mathematical information that may be less explicit, embedded in contexts that are not always familiar, and represented in ways that are more complex. Tasks require several steps and may involve the choice of problem-solving strategies and relevant processes. Tasks tend to require the application of, for example, number sense and spatial sense; recognizing and working with mathematical relationships, patterns, and proportions expressed in verbal or numerical form; and interpretation and basic analysis of data and statistics in texts, tables, and graphs.

Adults at this level are typically able to use ratios, understand mathematical information in a table, or read a complex graph. For example, they are likely able to:

- Determine the price of a single bottle of water when given the cost of an entire case of bottles

- Determine how many months in a year had sales above the mean sales for the year from a table of monthly sales

- Identify which predicted monthly gasoline price was most accurate based on line graphs of predicted and actual gasoline prices for a year

- Determine the amount of concentrated lemonade flavoring and water needed to make a large container of lemonade that is in the same ratio of flavoring to water as a smaller amount of lemonade

- Read a complex graph, comparing the amount of salt, sugar, and fat in a typical diet for men versus a typical diet for women, to determine the amount of sugar consumed by men

They may have trouble using numeracy skills that require using percentages, using rates, or understanding how quantities are related. For example, they might find it challenging to:

- Convert the number of students enrolled in a university each year into percentages, and then compute the change in the percentage of students enrolled each year

- Determine how much medicine to give to a child when the dosage is based on the child's body weight

- Calculate profit from a table containing lists of income and expense sources

Numeracy Level 4/5 **326 to 500**

Tasks at the lower end of this level require the respondent to understand a broad range of mathematical information that may be complex, abstract, or embedded in unfamiliar contexts. These tasks involve undertaking multiple steps and choosing relevant problem-solving strategies and processes. Tasks tend to require analysis and more complex reasoning about, for example, quantities and data; statistics and chance; spatial relationships; change; proportions; and formulas. Tasks in this level may also require comprehending arguments or communicating well-reasoned explanations for answers or choices.

Tasks at the higher end of this level require the respondent to understand complex representations and abstract and formal mathematical and statistical ideas, possibly embedded in complex texts. Respondents may have to integrate multiple types of mathematical information where considerable translation or interpretation is required; draw inferences; develop or work with mathematical arguments or models; and justify, evaluate, and critically reflect upon solutions or choices.

Adults at this level are typically able to use percentages and rates, interpret information presented in various ways, or understand how quantities are related. For example, they are likely able to:

- Identify which predicted monthly gasoline price was most accurate based on line graphs of predicted and actual gasoline prices for a year

- Determine the amount of concentrated lemonade flavoring and water needed to make a large container of lemonade that is in the same ratio of flavoring to water as a smaller amount of lemonade

- Convert the number of students enrolled in a university each year into percentages, and then compute the change in the percentage of students enrolled each year

- Read a complex graph, comparing the amount of salt, sugar, and fat in a typical diet for men versus a typical diet for women, to determine the amount of sugar consumed by men

- Determine how much medicine to give to a child when the dosage is based on the child's body weight

- Calculate profit from a table containing lists of income and expense sources

7.4 Reading components

The reading components assessment is designed to better understand the difficulties faced by test takers who demonstrate poor reading skills. Reading components represent the basic set of skills necessary to gain meaning from written text—knowledge of vocabulary, ability to process meaning at the level of the sentence, and fluency in reading passages.

- Vocabulary – Items testing print vocabulary consist of a picture of an object and four printed words, one of which refers to the pictured object. Respondents are asked to click on the word that matches the picture.

- Sentence comprehension – The sentence comprehension items require the respondent to assess whether a sentence makes sense in terms of the properties of the real world or the internal logic of the sentence. The respondent reads the sentence and clicks on yes if the sentence makes sense or no if the sentence does not make sense.

- Passage comprehension – In items assessing passage comprehension, respondents are asked to read a passage in which they are required at certain points to select the word that makes sense from the two alternatives provided.

Scores for each of these reading components skills are measured in terms of accuracy and rate. The rate is calculated as the average time to complete each item in each of the three sections (vocabulary, sentence completion, and passage comprehension). The accuracy is the percentage of correct responses in each section. Test administrators will be able to see the rate and accuracy for each of the three skills in the data download.

For each skill, results are reported in the score report and the data download in terms of low, medium, and high skills. For each literacy score we have calculated the high and low cut points for vocabulary, sentence comprehension, and passage comprehension for both rate and accuracy. The high and low cut points are set at the 25^{th} and 75^{th} percentiles for rate and accuracy for that particular literacy score (based on 10-point increments) for that particular language. These cut points are included in Appendix D.

Test takers receive the following descriptions of their skills:

- High accuracy and fast rate: Basic reading skills are good. The focus can be on building comprehension skills

- High accuracy and low or medium rate: Basic reading skills are good. The focus can be on building comprehension skills and increasing rate

- Low or medium accuracy and fast rate: The individual might be trying to go too fast. He or she needs to build basic skills

- Low or medium accuracy and low or medium rate: Work is needed on basic skills and getting faster

The Japanese language version of the reading components module has been modified in two ways due to the particular characteristics of the Japanese language and because Japan did not include reading components as part of its PIAAC Main Study administration. First, the Japanese reading components module does not include the vocabulary section. Second, the Japanese reading components score report does not provide test takers with any scores for the two remaining skill areas (sentence comprehension and passage comprehension) because there is not sufficient data to create reliable and accurate cut points for the Japanese language version. Test administrators will be able to see, however, the test taker's rate and accuracy for sentence comprehension and passage comprehension in the data download.

7.5 Problem solving in technology-rich environments (PSTRE)

The PSTRE questions measure how well one uses different types of technology to solve everyday problems and complete tasks to successfully meet goals. They also measure how well one understands and uses information in different environments, such as email, Web pages, or spreadsheets. In this test, a problem is any situation where one does not already have a good idea about how to achieve a goal. This may be because the strategy use does not appear obvious or because one has never tried such a task in the past.

Most adults use problem-solving skills in technology environments to find information or answer questions, use online tools and functions that can make tasks easier, and communicate with others. For example, one uses these skills to:

- Read and answer emails from friends or co-workers
- Search for a website with information about treatment for a medical issue
- Use a spreadsheet to set up a budget and keep track of spending
- Help a friend figure out how to install a new software program
- Set up folders on your computer to organize your emails or files
- Evaluate whether information on a Web page comes from a reliable source

The PSTRE domain is organized around three core dimensions: the cognitive strategies and processes a person uses to solve a problem, the tasks or problem statements that trigger and condition problem solving, and the technologies through which the problem solving is conducted. Variations within and across all of those dimensions were expected to contribute to the overall difficulty of the problems presented in the assessment. For example, a problem is likely to be more complex if it is ill defined as opposed to explicitly stated, requires complex problem-solving strategies such as defining goals and resolving impasses, and/or requires the use of multiple technology environments (e.g., respondents must utilize both emails and spreadsheets).

In order to explain how proficiency can be affected by the three dimensions of PSTRE, the problem-solving proficiency scale was divided into three levels as shown in Table 7.1 below. In this section, we describe the essential features of tasks at each of these three levels.

Table 7.1: Technology, task, and cognitive characteristics of problems at each of 3 main levels of proficiency

Level	Technology features	Task features	Cognitive processes
Level 1	• Generic applications • Little or no navigation required • Relevant information is directly available • Use of facilitating tools not required	• Few steps • Single operators	• Reach a given goal • Apply explicit criteria • Minimal monitoring demands • Simple relevance match • Categorical reasoning • No integrate or transformation
Level 2	• Both generic and novel applications (e.g., Web-based services) • Some navigation required to acquire information or perform actions • Use of tools facilitates operations	• Multiple steps • Multiple operators	• Goal may need to be defined • Apply explicit criteria • Generally higher monitoring demands • Generally involves resolving impasses • Some evaluation of relevance • Some integrate or transformation • Inferential reasoning
Level 3	• Generic and novel applications • Some navigation required to acquire information or perform actions • Use of tools required to efficiently solve the problem	• Multiple steps • Multiple operators	• Goal may need to be defined • Establish and apply criteria • Generally high monitoring • High inferential reasoning and integration • Evaluate relevance and reliability • Generally involves resolving impasses

In the score report, test takers receive a numerical score, which is rounded to the nearest 10 points. Because the scores at the extreme ends of the scale are less precise, no test taker will receive a score below 150 or above 400. The score report also includes a description of the proficiency level of the test taker. Test administrators will be able to download in an Excel spreadsheet the test taker's score as well as the start and end date and time for the module.

The proficiency levels of PSTRE are defined as follows:

PSTRE Below Level 1 **0 to 240**

Tasks are based on well-defined problems involving the use of only one function within a generic interface to meet one explicit criterion without any categorical, inferential reasoning, or transforming of information. Few steps are required and no subgoal has to be generated.

Adults at this level are typically able to complete tasks that are quite routine for them using familiar technology programs. For example, they are likely able to:

- Use a familiar email program to open and read emails
- Write a short summary of a club meeting using a word processing program they know well
- Enter the name of a local store into a search engine they have used in the past to find the store's phone number

They might sometimes have trouble using technology to solve problems that are more complex. For example, they might find it challenging to:

- Open and read email using an unfamiliar email program similar to one they regularly use
- Select a website from the results of a search and locate specific information on the homepage of that website
- Organize a small set of emails into one or two folders

PSTRE Level 1 **241 to 290**

At this level, tasks typically require the use of widely available and familiar technology applications, such as email software or a Web browser. There is little or no navigation required to access the information or commands required to solve the problem. The problem may be solved regardless of one's awareness and use of specific tools and functions (e.g., a sort function). The task involves few steps and a minimal number of operators. At a cognitive level, the person can readily infer the goal from the task statement; problem resolution requires one to apply explicit criteria; and there are few monitoring demands (e.g., the person does not have to check whether he or she has used the adequate procedure or made progress toward the solution). Identifying contents and operators can be done through simple match: only simple forms of reasoning, for example, assigning items to categories are required. There is no need to contrast or integrate information.

Adults at this level are typically able to use unfamiliar software programs that work like ones they have used in the past to solve problems where the goal is clear and a limited number of steps are required. For example, they are likely able to:

- Open, read, and respond to email using an unfamiliar email program

- Locate specific information on the homepage of a website that a friend has recommended
- Set up a system of folders that allow files or emails to be organized and easily retrieved

They might sometimes have trouble using technology to solve problems that are more complex. For example, they might find it challenging to:

- Figure out how to send an email message to a number of contacts using an unfamiliar bulk email function
- Use a sorting tool to make it easier to locate sales numbers for a specific product in a company spreadsheet
- Conduct a Web search to find out how to solve a problem with other software, such as how to view a column that won't display properly in a spreadsheet
- Find an email message or file that has been "lost" somewhere on a computer hard drive

PSTRE Level 2 **291 to 340**

At this level, tasks typically require the use of both generic and more specific technology applications. For instance, the person may have to make use of a novel online form. Some navigation across pages and applications is required to solve the problem. The use of tools (e.g., a sort function) can facilitate the resolution of the problem. The task may involve multiple steps and operators. In terms of cognitive processing, the problem goal may have to be defined by the person, though the criteria to be met are explicit. There are higher monitoring demands. Some unexpected outcomes or impasses may appear. The task may require evaluating the relevance of a set of items to discard distractors. Some integration and inferential reasoning may be needed.

Adults at this level are typically able to use software they have never seen before to solve problems that are more complex, even when unexpected impasses/outcomes occur. For example, they are likely able to:

- Figure out how to send an email message to a number of contacts using an unfamiliar bulk email function
- Use a sorting tool to make it easier to locate sales numbers for a specific product in a company spreadsheet
- Conduct a Web search to find out how to solve a problem with other software, such as how to view a column that won't display properly in a spreadsheet
- Find an email message or file that has been "lost" somewhere on a computer hard drive

They might sometimes have trouble using technology to solve problems that are more complex. For example, they might find it challenging to:

- Establish criteria for narrowing a Web search, documenting results using a spreadsheet, and communicating the results to others through email

- Evaluate a number of Web search results to determine which has the most relevant and reliable information. Part of this process includes evaluating and refining a search to determine if additional or different types of websites should be considered

- Use a software program that they have never seen before with limited or unclear directions based on general experience with technology or by consulting other online resources including websites or user blogs

- Select from among a number of choices the best software to use for a particular task.

PSTRE Level 3 **341 to 500**

At this level, tasks typically require the use of both generic and more specific technology applications. Some navigation across pages and applications is required to solve the problem. The use of tools (e.g., a sort function) is required to make progress toward the solution. The task may involve multiple steps and operators. In terms of cognitive processing, the problem goal may have to be defined by the person, and the criteria to be met may or may not be explicit. There are typically high monitoring demands. Unexpected outcomes and impasses are likely to occur. The task may require evaluating the relevance and the reliability of information in order to discard distractors. Integration and inferential reasoning may be needed to a large extent.

Adults at this level are typically able to use one or more complex software programs to solve ill-defined problems with multiple goals. For example, they are likely able to:

- Conduct a Web search to find out how to solve a problem with other software, such as how to view a column that won't display properly in a spreadsheet

- Figure out how to send an email message to a number of contacts using an unfamiliar bulk email function

- Evaluate a number of Web search results to determine which has the most relevant and reliable information. Part of this process includes evaluating and refining a search to determine if additional or different types of websites should be considered

- Use a software program that they have never seen before with limited or unclear direction. Success may be based on a user's general experience with technology or information may be gathered by consulting other online resources including websites or user blogs

- Select from among a number of choices the best software to use for a particular task

7.6 Comparison charts in literacy, numeracy, and PSTRE score reports

The literacy, numeracy, and PSTRE score reports include three charts that show the test taker how he or she compares to a series of averages from both the test taker's country and the overall OECD average.

The first chart shows the test taker how his or her score compares to the average score for people with different education levels. The three comparison education levels are below upper secondary, upper secondary, and tertiary. These comparison groups were created from the OECD International Data Explorer variable EDCAT6. Table 7.2 provides information on how the EDCAT6 categories were collapsed.

Table 7.2: Education categories used in Education & Skills Online cognitive score reports

Group in Education & Skills Online	EDCAT6 Category
Below upper secondary	Lower secondary or less (ISCED 1,2, 3C short or less)
Upper secondary	Upper secondary (ISCED 3A-B, C long)
Tertiary	Post-secondary, non-tertiary (ISCED 4A-B-C)
	Post-secondary professional degree (ISCED 5B)
	Post-secondary bachelor degree (ISCED 5A)
	Post-secondary master/research degree (ISCED 5A/6)
	Tertiary - bachelor/master/research degree (ISCED 5A/6)

The second chart shows the test taker how his or her score compares to the average score for people with different occupations. The four comparison occupation levels are from the OECD International Data Explorer variable ISCOSKIL4 (occupational classification of respondent's job [4 skill-based categories], last or current [derived]). The four occupation skill levels are: skilled, semi-skilled white collar, semi-skilled blue collar, and elementary.

The third chart shows the test taker how his or her score compares to the average score for people of different ages. The five age comparison groups are from the OECD International Data Explorer variable AGEG10LFS (age in 10-year bands [derived]). The five age bands are 24 or less, 25 to 34, 35 to 44, 45 to 54, and 55 and over.

Spain and Italy did not administer the PSTRE module during PIAAC 2012; therefore, no data are available to make a country comparison in the score reports for the Italian and Spanish (Spain) tests. Score reports for those languages will have international comparison charts only.

References

Organisation of Economic Co-operation and Development (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD Survey of Adult Skills*. Paris, France: Author. doi:org/10.1787/9789264128859-en

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

Chapter 8: Noncognitive Modules Score Reporting

8.1 Introduction

Education & Skills Online provides test takers and test administrators with individual-level results for the noncognitive modules. Test administrators determine whether the score reports appear automatically for test takers after each module. Regardless of whether score reports are presented to the test taker, a copy of the score report is saved in the Administration Portal and can be downloaded by the test administrator at any time. Examples of the score reports for each module are included in Appendix B.

8.2 Skill Use

The Skill Use module for Education & Skills Online includes 57 items across 8 scales. The scales focus on the frequency with which test takers use the skills associated with reading, writing, numeracy, and information and communications technology (ICT) at home and at work.

The reading scale measures how often test takers use the skills required to read documents such as directions, instructions, letters, memos, emails, articles, books, manuals, bills, invoices, diagrams, and maps. The writing scale measures how often test takers use the skills required to write documents such as letters, memos, emails, articles, reports, and fill-in forms. The numeracy scale measures how often test takers use the skills required to calculate prices, costs, or budgets; use fractions, decimals, or percentages; use calculators; prepare graphs or tables; use algebra or formulas; and use advanced math or statistics. The ICT scale measures how often test takers use the skills required to use email, the Internet, spreadsheets, word processors, and programming languages; conduct transactions online; and participate in online discussions (conferences, chats).

For each scale, respondents are asked four to eight questions about how often they use these skills in their home or work lives. The response options are 1) never, 2) less than once a month, 3) less than once a week but at least once a month, 4) at least once a week but not every day, and 5) every day. For ICT skill use, respondents were first asked whether they had ever used a computer; questions assessing the domain are not presented to those without any previous contact with computers. In contrast, reading, writing, and numeracy skills used at home are assessed for all respondents. The corresponding scales for skills used at work are assessed only for those respondents who are part of the labor force or have been in the labor force at some time, as determined by their answer to the background question on their current employment status and a question at the beginning of the Skill Use module asking if they have ever been employed.

Test takers receive scores of not applicable, low, moderate, or high for each of the skill use scales, defined as follows:

- Not applicable: The test taker reported that he or she never engaged in any of the activities involving this skill.
- Low: The test taker reported that he or she rarely engaged in most of the activities involving this skill.
- Moderate: The test taker reported that his or her engagement in activities varied in terms of how many activities were done and how often they were done.
- High: The test taker reported that he or she engaged in most activities on most days or every day.

Test takers receive a score of not applicable when they indicate that they have never engaged in any of the activities mentioned in the module for that scale. To determine whether a test taker should receive a score of low, moderate, or high for a particular skill, the test taker's responses are compared to responses from participants in the 24 countries in the Programme for the International Assessment of Adult Competencies (PIAAC). Test takers receive a score of low if their responses indicate they are in the bottom quintile (one-fifth of the distribution) of individuals internationally who use that skill. Test takers receive a score of moderate if their responses indicate they are in the middle three quintiles of individuals internationally who use that skill. Test takers receive a score of high if their responses indicate they are in the top quintile of individuals internationally who use that skill.

8.3 Career Interest and Intentionality

The Career Interest and Intentionality portion of the assessment consists of 60 items from the O*NET Interest Profiler Short Form (Rounds, Su, Lewis, & Rivkin, 2010). This set of items is composed of 10 items from each of the six RIASEC scales (realistic, investigative, artistic, social, enterprising, and conventional). All items had a five-point Likert response scale from 1) strongly dislike to 5) strongly like. Scores for each RIASEC dimension are calculated by averaging the 10 item values within each dimension. The test taker will receive a score of 0 to 40 in each interest area. Higher scores indicate the test taker's interests are more aligned with that interest area.

Table 8.1: Interest area description

Interest Area	Description	Examples of Work
Realistic	People with realistic interests like work that includes practical, hands-on problems and answers. Often people with realistic interests do not like careers that involve paperwork or working closely with others.	• Working with plants and animals • Real-world materials like wood, tools, and machinery • Outside work
Investigative	People with investigative interests like work that has to do with ideas and thinking rather than physical activity or leading people.	• Searching for facts • Figuring out problems
Artistic	People with artistic interests like work that deals with the artistic side of things, such as acting, music, art, and design.	• Creativity in their work • Work that can be done without following a set of rules
Social	People with social interests like working with others to help them learn and grow. They like working with people more than working with objects, machines, or information.	• Teaching • Giving advice • Helping and being of service to people
Enterprising	People with enterprising interests like work that has to do with starting up and carrying out business projects. They like taking action rather than thinking about things.	• Persuading and leading people • Making decisions • Taking risks for profits
Conventional	People with conventional interests like work that follows set procedures and routines. They prefer working with information and paying attention to details rather than working with ideas.	• Working with clear rules • Following a strong leader

The career interest assessment uses the interest profile for 436 occupations from the O*NET database to determine how well the test taker's interests match the interest profile of the test taker's current and desired occupations. Using the test taker's RIASEC profile, a job fit score from minus-100 to 100 is calculated for the current and desired occupations. The occupation is considered a low fit if the job fit score is less than 10 points, a moderate fit if the score is between 10 and 50 points, and a high fit if the score is 50 points or above. If the test taker indicates in the background questionnaire that he or she is "unemployed, not looking for work" then no score will be provided for the fit of current and desired jobs. A job fit score also is calculated for each of the 436 occupations, and a list of the highest scoring 20 occupations and the lowest scoring 10 occupations are provided to the test taker in the score report. The highest scoring occupations are considered the best fit for the test taker's interests and the lowest scoring are considered the worst fit.

Occupational interest profiles in O*NET were developed using subject matter expert ratings. Two groups of three trained raters considered those occupations included in the O*NET database. The appropriateness of each RIASEC category for each occupation based on O*NET

data for the occupation was evaluated. The mean rating for the three reviewers was calculated for each of the six interest dimensions across occupations. Inter-rater agreement and validity evidence were also assessed. A high degree of rater reliability was found, as was alignment to Holland's theoretical RIASEC model.

The career intentionality portion of the assessment consists of 26 items. This set of items is composed of six items that measure job-seeking intentionality, six that measure training intentionality, four that measure job-seeking and training self-efficacy, and 10 that measure taking active steps. Job-seeking intentionality, training intentionality, and job-seeking and training self-efficacy scales had a six-point response scale from 1) strongly disagree to 6) strongly agree. Scores are calculated for each scale by averaging item responses. The "taking active steps" scale had a binary response of yes or no. The total number of yes responses is used as the score. If the test taker indicates in the background questionnaire that he or she is "unemployed, not looking for work" then no score will be provided for the career intentionality assessment.

The job-seeking intentionality, training intentionality, and job-seeking and training self-efficacy scales use stanine (nine-point standard scale) scores to determine low, moderate, and high scoring groups. Scores were placed on the scale, using lower and upper bounds to establish a range for each group (note: stanines 8 and 9 nine were combined due to small N-count). Cutoffs for low, moderate, and high are included in Table 8.2. Though gaps existed in the values of the upper and lower bounds, none of the calculated intentionality scores will equal those values. For the taking active steps scale, low (~50% of norm group; score = 0), moderate (~25%; score between 1 and 3), and high groups (~25%; score between 4 and 8) were established using test takers' raw scores from the Field Test in 2014.

Table 8.2: Career Intentionality score cutoffs

Scale	Score		
	Low	Moderate	High
Job-seeking intentionality	1.00 - 1.67	1.83 - 4.84	5.00 - 6.00
Training intentionality	1.00 - 2.67	2.83 - 5.00	5.16 - 6.00
Job-seeking and training self-efficacy	1.00 - 3.25	3.50 - 5.25	5.50 - 6.00
Taking active steps	0	1.00 - 3.00	4.00 – 8.00

8.4 Subjective Well-Being and Health

The subjective well-being portion of the module focuses on a test taker's attitudes and feelings toward his or her life, using cognitive and emotional measures of life satisfaction. The cognitive measure is an adapted version of the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). The adapted SWLS includes four items on a six-point Likert type

response scale including 1) strongly disagree, 2) disagree, 3) slightly disagree, 4) slightly agree, 5) agree, and 6) strongly agree. Scores are calculated by averaging the four item responses, resulting in a total score from 1 to 6, and then comparing them to the scores collected during the Field Test. A quartile (one-fourth of the distribution) approach, appropriate for cross-cultural comparisons, was used to define scoring cutoffs for reporting. Scores in the first quartile are reported as low, in the second and third quartiles as moderate, and in the fourth quartile as high.

Table 8.3: Life Satisfaction score report elements

Score	Result
High	Your score shows that you are very satisfied with your life and feel good about how it is going. Generally, people who score in this range take on life's challenges without feeling overwhelmed.
Moderate	Your score shows that you are somewhat satisfied with your life. You may feel as though you are doing well in some areas while feeling other areas need improvement. People who report having a moderate level of life satisfaction for long periods of time may want to think about why this is. After reflection, it is important for them to try to make positive changes in their lives.
Low	Your score shows that you are not very satisfied with your life. When possible, changes in circumstances (e.g., schedule, activities), attitudes, and behaviors are recommended for people with a low score. These changes may result in positive ways of dealing with difficult situations and improvement in life satisfaction.

The second element of subjective well-being is an emotional evaluation describing the test taker's emotional experience of his or her life. While life satisfaction is assessed on a single dimension, the emotional evaluation is composed of two distinct dimensions: positive affect (PA) and negative affect (NA). The Education & Skills Online measure for emotional evaluation is an adapted version of the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) and I-PANAS-SF (Thompson, 2007), an internationally validated short form of the instrument. The Education & Skills Online scale is composed of nine items, including four PA items and five NA items. Respondents are asked to rate their experience of each emotion during the previous week, measured using a five-point scale including 1) very slightly or not at all, 2) a little, 3) moderately, 4) quite a bit, and 5) extremely. Scores for PA are calculated by averaging the four positive item responses, while NA is calculated by averaging the five negative item responses, resulting in one total score for each dimension. Dimension scores are then compared to the 1,890 scores collected during the Field Test. A quartile approach, appropriate for cross-cultural comparisons, is used to define scoring cutoffs, reporting PA and NA scores in the first quartile as low, in the second and third quartiles as moderate, and in the fourth quartile as high.[1]

[1] Analysis of the Field Test data for Slovenia indicate that the positive affect score for this country version should be interpreted with caution as the reliability of the positive affect scale was not as high as that of the other countries in the Field Test.

Table 8.4: Positive and negative affect score report elements

Score	Result provided to test taker
High Positive Affect	Your score shows that you had positive moods and emotions in the past week. People who usually score high in this category feel happiness and are often quick to smile, energetic, and enjoy their work.
Moderate Positive Affect	Your score indicates that you experienced moderate positive moods and emotions in the past week. People who score in this range can appear emotionally controlled while being hard to read due to a lack of obvious enthusiasm.
Low Positive Affect	Your score shows that you had low levels of positive moods and emotions in the past week. People who score in this range have had fewer positive experiences and felt sadness in the past week, which sometimes results in feeling tired and little activity.
Low Negative Affect	Your score shows that you experienced low levels of negativity in the past week. People who score in this range appear calm and composed.
Moderate Negative Affect	Your score shows that you experienced moderately negative moods and emotions in the past week. People who score in this range appear somewhat angry, annoyed, and tense.
High Negative Affect	Your score shows that you experienced negative moods and emotions multiple times in the past week. People who score high in negative affect experience negative feelings more often than others. They are often frustrated and depressed.

Health is a complex multidimensional construct whose definition has evolved from a purely biological measure to include psychosocial factors considered critical to the assessment of overall well-being. Gathering health data is an integral component of the ongoing effort to monitor economic and social progress across countries and promote policies aimed at improving overall life quality (Organisation for Economic Co-operation and Development, 2012). The Education & Skills Online measures of subjective and behavioral health include 14 survey items on the feelings and behaviors most relevant to health as described in the OECD agenda. These include items on subjective health, body mass index (BMI), nutrition, exercise, sleep, and smoking status. Reporting of subjective and behavioral health is presented in an informational format as the respondent's self-reported health perceptions and behaviors are compared to the accepted international health recommendations (World Health Organization; WHO, 2015) for each category.

Subjective health is a single item measure on a test taker's self-perception of his or her health, measured with a six-point response scale, including 1) very poor, 2) poor, 3) fair, 4) good, 5) very good, and 6) excellent. Based on the response, a health outlook is reported as poor for scores 1 or 2, fair for 3 or 4, and positive for 5 or 6.

BMI, an internationally accepted health measure, is calculated using the self-reported responses for height and weight. The corresponding report element offers a definition of BMI as well as a classification of underweight, normal weight, or overweight based on the international classification of the WHO (2015).

As an indicator of nutrition and based on the international nutritional recommendations (WHO, 2015), four questions elicit the number of servings per day and days per week that the respondent consumes fruits and vegetables. Self-reported servings and frequencies of fruit and vegetable consumption are used to provide an evaluation of the test taker's diet. The corresponding report element on diet and nutrition details the current recommendation and indicates whether the respondent consumes no fruits and vegetables, insufficient amounts of fruits and vegetables, or greater than or equally sufficient amounts of fruits and vegetables per the current recommendations.

One item elicits the smoking status of the respondent by asking if he or she currently smokes any tobacco products including cigarettes, cigars or pipes, offering response options of no, yes, sometimes, or "yes, daily." Because smoking is a recognized health risk and not recommended at any level, the corresponding report element describes the health risk of smoking and indicates whether the respondent has reported smoking behavior.

Based on the international recommendations (WHO, 2015) for physical activity, which suggest 75 minutes of intense activity or 150 minutes of moderate activity per week, four items eliciting information on the frequency and intensity of physical activity are included. These four items request the amount of time and number of days a respondent engages in both moderate and vigorous exercise. Total times are calculated for both moderate and intense activity and equated to a common scale (weekly minutes of intense exercise times two), which is compared to the 150 minutes recommendation. The corresponding report element on exercise details the current recommendation and indicates whether the respondent does not exercise, exercises at a level insufficient to the stated standard, or meets or exceeds the recommendation.

Two items elicit duration and quality of sleep, as these are core features of commonly accepted sleep recommendations. The item on sleep quality offers responses of very bad or fairly bad for indicators of insufficient sleep quality, and fairly good or very good as indicators of adequate sleep quality. The sleep duration item requests the average amount of sleep in hours for the past month. The health behavior report details the current recommendation of seven to nine hours of quality sleep and related health benefits while reporting whether the test taker meets the recommendation.

Table 8.5: Subjective health and behavioral report elements

Score Category	Recommendation included in the score report
Health Outlook	Being optimistic about your health given your particular situation is important to managing illness. Focusing more on what you can do and less on what you cannot do in terms of your health, can positively impact your ability to cope with and recover from health challenges
Body Mass Index (BMI)	Body Mass Index (BMI) is a simple index of weight for height and is commonly classified as Underweight, Normal Weight, or Overweight. Maintaining a healthy weight is important for your overall health. It can lower your risk for many illnesses and conditions, while increasing your energy level.
Diet/Nutrition	A healthy diet includes eating fruits and vegetables every day. This can reduce your risk for illnesses such as heart disease, cancer, and diabetes. Current recommendations suggest that we eat 400 grams (approximately 3 cups or 5 servings) of fruits and vegetables per day to maintain good health.
Smoking	Smoking is a large risk factor for serious illnesses such as heart attack, stroke, and cancer. Avoid smoking and second hand smoke to positively impact your health.
Exercise	Regular physical exercise is important to reducing stress, managing your weight, and maintaining good health. The current recommendation is 150 minutes of moderate or 75 minutes of intense exercise each week. Walking regularly, taking the stairs, and starting a new sport are ways to increase physical activity.
Sleep	Good sleep habits lead to better mood and functioning and reduce the risk of illness. Too little sleep can lead to illness, irritability, and difficulty concentrating. It is best to get 7 to 9 hours of quality sleep per night. A regular sleep schedule, including habits that encourage uninterrupted sleep, such as limiting caffeine and alcohol, are helpful.

8.5 Behavioral Competencies (Only Available March 2018 to June 2020)

The Behavioral Competencies module was designed as a personality assessment for use in Education & Skills Online. Intended for developmental purposes, this assessment provides scores across 13 traits that are expected to be critical to success in education and the workplace. Research has found that noncognitive traits, such as those assessed by the Behavioral Competencies module, are malleable and that their improvement may aid in improving outcomes such as grades in school, anxiety reduction, and enhanced work-relevant skills (Heckman & Kautz, 2013; Hembree, 1988; Roberts, Walton, & Viechtbauer, 2006). Consequently, the personality scores provided by the module may provide valuable feedback to test takers in identifying areas most in need of improvement.

Although a variety of theoretical frameworks for describing personality have been proposed, the five-factor model of personality (Goldberg, 1990) has emerged as the predominant to approach framing personality measurement. The 13 personality facets assessed by the Behavioral

Competencies module can be described based on their relation to broad categories of personality, which will be described here. The five-factor model is described by five traits. Openness to experience reflects people's willingness to make adjustments to existing attitudes and behaviors once they have been exposed to new ideas or situations (Flynn, 2005). Conscientiousness reflects the degree to which a person is hard working, dependable, and detail oriented (Berry, Ones, & Sackett, 2007). Extraversion reflects the tendency to be sociable, assertive, active, and experience positive affects such as energy and zeal (Judge et al., 2002). Agreeableness reflects the degree to which a person is likable, easy to get along with, and friendly (Berry et al., 2007). Emotional stability/neuroticism reflects the degree to which a person is secure, calm, has low anxiety, and has low emotionality (Berry et al., 2007). These domains consist of lower order personality traits, including diligence, optimism, and creativity as elements of conscientiousness, agreeableness and openness, respectively, for example. The structure of the five-factor model consistently has been identified across a range of international contexts and data sources (e.g., Hendriks et al., 2003; McCrae & Costa, 1999; McCrae & Terracciano, 2005). The measurement of the Big Five and constituent lower order traits has been found useful in predicting job performance (Neal, Yeo, Koy, & Xiao, 2012; Tett, Jackson, & Rothstein, 1991).

The 13 facets of personality measured by the Behavioral Competencies module were drawn from a comprehensive taxonomy of 21 lower-order personality facets derived by Drasgow et al. (2012). These narrow trait domains were determined through the factor analysis of data from a sample of individuals responding to seven major personality inventories over a period of five years and an analysis of the lexical structure of those inventories (see Drasgow et al., 2012, for a complete description of this process). For each higher-order personality dimension, an exploratory factor analysis was performed, the results of which were used to identify groups of subscales describing logically similar aspects of the dimension.

The Behavioral Competencies assessment consists of 208 statements that represent 13 traits indicative of important workplace behaviors. The assessment employs a forced choice methodology that combines those items into 104 statement pairs, where respondents are required to choose the statement in the pair that is most like them. Forced choice methodology is resistant to test faking as the statements in an item pair have equal social desirability. Across the 104 pairs, each personality facet was represented by 16 statements, so that the number of opportunities for responding to each was held constant. The statements and statement parameters (α, δ, τ) used for construction and scoring of the Behavioral Competencies module were drawn from an existing pool that was previously developed across a series of studies within the United States. Although statement content was translated for each of the personality statements using the same approach as that used by Education & Skills Online cognitive content, parameter estimates were not obtained for each locale. Consequently, whether the statement operation is invariant across the different language and country combinations is unknown and scores from this module should be interpreted with caution. The Field Tests conducted in 2013 and 2017 were able to confirm that the higher-order structure of the five-factor model holds across the participating countries.

In the Behavioral Competencies score report, test takers receive their percentile rank for each personality trait. The percentile ranks are based on the international data obtained during the Education & Skills Online 2013 Field Test from the participating countries. Our analysis after

the 2017 Field Test indicated that the percentile estimates from the 2017 Field Test countries do not differ significantly from the percentile estimates calculated from the 2013 Field Test countries' data. Examining the percentile scores for each of the 13 traits is appropriate for group-level comparisons, such as examining score distributions across groups of test takers within an organization or a country. At the individual level, scores may not be as reliable and should not be used for high-stakes decision-making.

The 13 personality facets measured by the Behavioral Competencies module and their descriptions in the context of workplace behaviors are provided in Table 8.6, organized by the Big 5 domains.

Table 8.6: Behavioral Competencies scale descriptions

Big 5 Domain	BPC Scale	Description
Agreeableness	Collaboration	Describes individuals who are viewed as trusting and cooperative. People high in collaboration are often easy to get along with and usually work well on teams.
	Generosity	Describes individuals who are willing to offer their time and resources in support of others. People high in generosity tend to be helpful to others at work
Conscientiousness	Diligence	Describes behaviors associated with working towards objectives. Individuals who are high in diligence are goal orientated tend to be described as hard working, ambitious and confident
	Organization	Includes behaviors associated with maintaining a sense of order as well as an the ability to plan work tasks and work activities
	Dependability	Behaviors related to a sense of personal responsibility. Individuals who are high in dependability tend to be reliable and make every effort to keep promises
	Self-Discipline	Indicates an ability to be patient, cautious and level-headed. People who are high in self-discipline tend to maintain control at work
Extroversion	Assertiveness	Indicates an ability to take charge at work. People who are assertive are often described as direct, decisive and "natural leaders."
	Friendliness	Indicates an interest in social interactions. People high in friendliness are often interested in meeting new people at work and using this skill for the betterment of the organization
Emotional Stability	Stability	Describes individuals who are relaxed and worry free. People high in stability often works well with changing work priorities and manage stress well
	Optimism	Describes individuals who have a positive outlook and cope well with setbacks. People who are optimistic tend to incorporate feedback well at work.
Openness to Experience	Inquisitiveness	Describes behaviors that relate to being perceptive and curious at work. People high in inquisitiveness tend to be interested in learning more by attending workshops at work.
	Creativity	Behaviors that are inventive and imaginative. People high in creativity tend to be innovators at work.
	Intellectual Orientation	Indicative of an ability to process information and make decisions quickly. People high in intellectual orientation are often viewed as knowledgeable by others

References

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology,* 92(2), 410–424.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49, 71-75.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (Technical Report No. 1311). Fort Belvoir, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Flynn, F. J. (2005). Having an open mind: The impact of Openness to Experience on interracial attitudes and impression formation. *Journal of Personality and Social Psychology, 88*(5), 816-826.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216.

Heckman, J. J., & Kautz, T. (2013). *Fostering and measuring skills: Interventions that improve character and cognition* (NBER Report No. w19656). Cambridge, MA: National Bureau of Economic Research.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Hendriks, A. A. J., Perugini, M., Angleitner, A., Ostendorf, F., Johnson, J. A., De Fruyt, F. … Ruisel, I. (2003). The five-factor personality inventory: Cross-cultural generalizability across 13 countries. *European Journal of Personality, 17*(5), 347-373.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765-780.

McCrae, R. R., & Costa Jr., P. T. (1999). A five-factor theory of personality. Handbook of personality: *Theory and Research, 2*, 139-153.

McCrae, R. R., & Terracciano, A. (2005). The five-factor model and its correlates in individuals and cultures.

Neal, A., Yeo, G., Koy, A., & Xiao, T. (2012). Predicting the form and direction of work role performance from the Big 5 model of personality traits. *Journal of Organizational Behavior, 33*(2), 175-192.

Organisation for Economic Co-operation and Development (2011). *How's life?: Measuring well-being*. Paris, France: Author. doi:10.1787/9789264121164-en

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin, 132*(1), 1.

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). *O*NET® interest profiler short form psychometric characteristics: Summary*. Raleigh, NC: National Center for O*NET Development.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703-742.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology, 38*, 227-242. doi:10.1177/0022022106297301

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

World Health Organization (2015). *Global database on body mass index*. Retrieved August 5, 2015 from http://apps.who.int/bmi/index.jsp?introPage=intro_3.html