

Chapter 1: PIAAC Assessment Design

Irwin Kirsch and Kentaro Yamamoto, ETS

The heart of any large-scale comparative survey is the assessment design. This chapter provides an overview of both the Field Test and Main Study designs. These designs were complex because PIAAC measured four domains – literacy, numeracy, reading components and problem solving in technology-rich environments – across two modes of administration – paper-and-pencil and computer delivered – while also offering participating countries both core and optional components. As the intent of PIAAC was to have its results linked to previous international adult assessments, these designs assumed that 60 percent of the literacy and numeracy tasks would come from ALL and IALS. New items were also developed for the literacy and numeracy domains and new measures developed for reading components and problem solving in technology-rich environments based on their respective frameworks.

The assessment designs assumed approximately 30-40 minutes of administration time for the BQ and JRA and 60 minutes for the direct assessment. The JRA items collected information on skill use at work, while the BQ collected contextual information about respondents, including their demographic characteristics, educational background, labor market experiences, and skill use outside of work. The JRA and background items were collected and processed through the use of a CAPI system. The target population ranged from 16 to 65 years of age.

1.1 Field Test goals and design

Field Tests are an integral part of any large-scale assessment and must be designed to yield adequate information relating to four key areas: survey operations, instrument quality, computer-delivery platform, and scaling and psychometric characteristics. Standardized procedures and quality mechanisms were embedded into various phases of PIAAC including survey development, implementation, and analysis and reporting of the data. The outcomes of the Field Test were used to assemble the final instruments for the Main Study and to modify or refine any of the operational issues detailed in the “standards and guidelines” document that improved the overall quality of the assessment.

1.1.1 Operational goals

Operation includes an examination of the efficiency and accuracy of data collection procedures, response rates for various subpopulations of interest, efficiency and accuracy of data processing

including recoding, and data transmission. In particular, the following issues related to field operations needed to be examined:

- Review sample characteristics in terms of responses to BQ
- Review response rates by key background variables
- Evaluate coding of nonresponse interviews
- Identify and fix operational difficulties
- Summarize administration time for BQ as well as cognitive items
- Evaluate efficacy of scoring of paper-and-pencil items
- Evaluate efficacy of data capture
- Evaluate operational issues associated with International Standard Classification of Education coding and other BQ variables
- Evaluate efficacy and accuracy of data transmission
- Review and approve quality assurance mechanisms

1.1.2 Instrumentation

In addition to survey operations, the Field Test needed to provide quality information relating to the survey instruments, including adequacy of the scoring procedures, examination of translation and adaptation, and an evaluation of the scaling and analytic procedures that were used. In particular, the Field Test needed to address the following issues related to instrumentation:

- Review accuracy and comparability of survey instruments, including translation and scoring guides and all related manuals
- Evaluate the timing and flow of questions in the BQ
- Evaluate appropriateness of questions across participating countries
- Examine response distribution in all categories of BQ

1.1.3 Computer-delivery platform

PIAAC represents an innovation in large-scale assessment methodology in that the assessment was also computer based. PIAAC was the first large-scale assessment delivered on a laptop computer to respondents in their homes. An integrated computer-delivery platform was used to integrate the CAPI tool to be used for the administration of the BQ and the JRA with the tool that delivered the cognitive

instruments. In its turn, the integrated PIAAC system needed to work in conjunction with the survey management systems of the organizations administering the survey in countries. Thus, in addition to looking at the instruments and survey operations, the Field Test also addressed the following issues related to the computer-delivery platform:

- Test and evaluate the functioning of the cognitive portion of the delivery platform, particularly response capturing and automatic scoring
- Test and evaluate the functioning of the CAPI system, particularly the flow of questions and efficiency of the system in capturing information
- Evaluate the accuracy of the interviewer's instructions
- Test the effectiveness of the system during the interview
- Verify the integration of the PIAAC platform with national survey management systems

1.1.4 Scaling and psychometric characteristics

The Field Test design allowed us to evaluate the psychometric characteristics of items and scales, including the evaluation of the equivalence of item parameters among linking items from IALS and ALL to PIAAC, and the equivalence of item parameters between paper-and-pencil and computer formats. In the case of PIAAC, the Field Test was also an opportunity to examine the role of computer familiarity and to determine the standards for branching respondents. In this regard, the Field Test provided initial IRT parameters that were used to construct the adaptive testing algorithm that were then implemented in the Main Study. In particular, the Field Test addressed the following issues associated with respect to IRT scaling and psychometric characteristics:

- Examine equivalence of item characteristics among the literacy and numeracy items common to IALS and ALL on the paper-and-pencil version
- Examine equivalence of item characteristics of literacy and numeracy items common to paper-and-pencil and computer-based formats
- Examine equivalence of item characteristics across languages within a country
- Examine equivalence of item characteristics across countries
- Identify tasks among the literacy, numeracy and problem-solving items that could be assembled into a core assessment
- Examine the expected proportions of subsamples routed to the different formats and to the different stages of the computer-delivered testlets based on preliminary background information and the core.

- Evaluate the overall psychometric characteristics and quality of the Field Test items to guide the selection of items for the Main Study

Because this was the first cycle of PIAAC, the Field Test was also viewed as a “dress rehearsal” of all newly developed aspects of the survey. In terms of sampling procedures, the Field Test did not need to be a full probability sample. However, critical aspects of sampling (such as sampling individuals within households), as well as other aspects of the overall sampling plan (such as descriptions of the sampling frames), and sampling guidelines had to be tested in this phase of the project. All quality control forms and procedures were also developed and tested. Finally, even though weights were not required for the Field Test, the weighting process was evaluated using the Field Test data.

1.2 The Field Test design: An integrated approach

This central Field Test design provided good item-level information on the full range of direct assessment measures included in PIAAC and was extremely useful in addressing other operational and psychometric issues identified above. The BQ and a core set of questions focusing on ICT helped to ensure that respondents who reported no familiarity with computers were routed to the paper-and-pencil version of the assessment. In order to link the paper-and-pencil and the computer-delivery formats, the remaining adults (the majority of adults in each country who are expected to pass the core) were randomly assigned to either the paper-and-pencil or computer-delivered branches of the Field Test (see Figure 1.1).

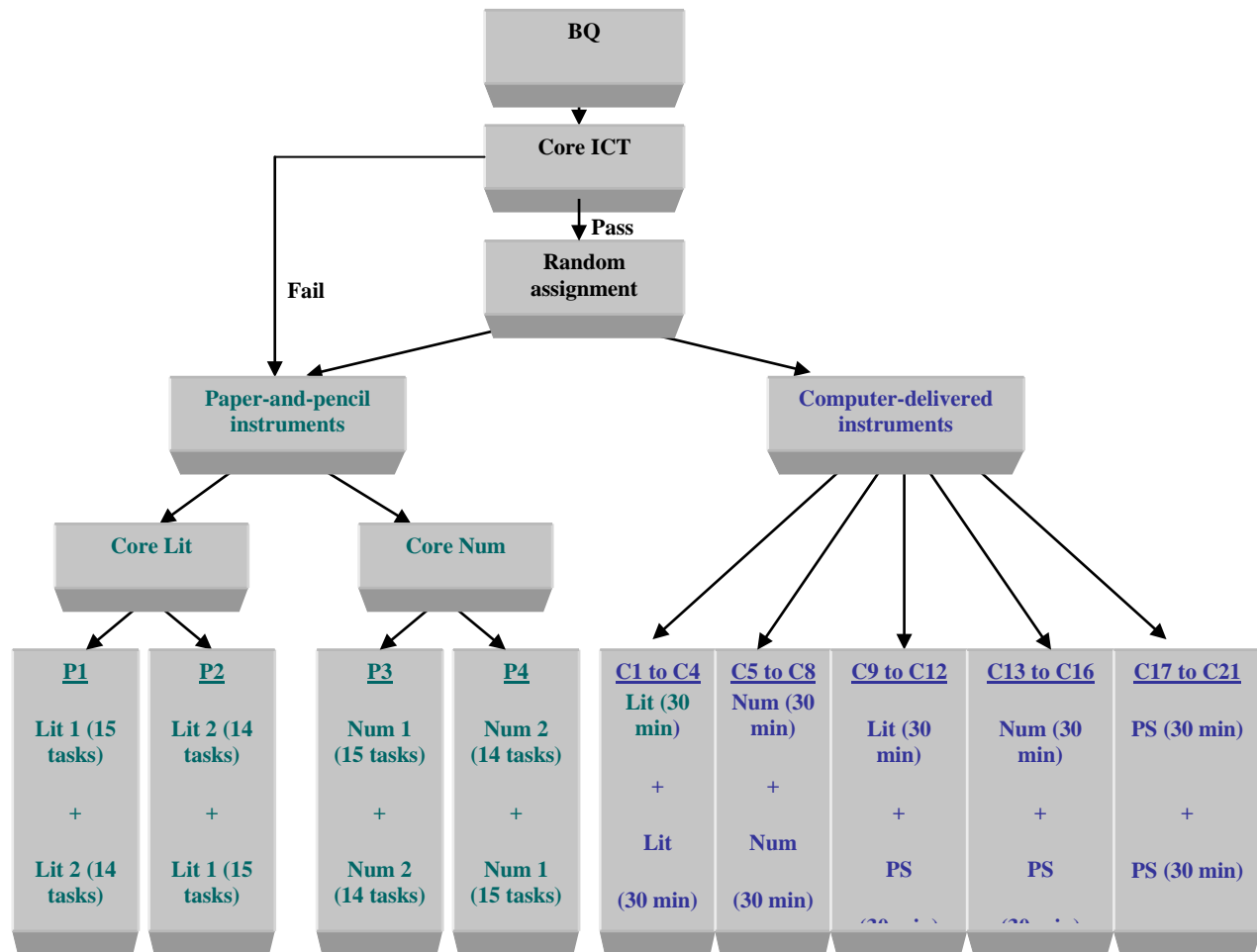
The BQ, including the JRA, was delivered by the interviewer using a computer-assisted format with respondents taking one of three variants, 30-40 minute variants (a 20-minute core set of items and one of three, 10-minute subsets) that were administered along with the cognitive instruments. The paper-and-pencil branch of the direct assessment was composed of a 10-minute core of either literacy or numeracy skills with six tasks each. This was followed by a pair of 20-minute clusters of literacy or numeracy, totaling 29 tasks, and a final 10-minute cluster of component skills. Four paper booklets were designed (details in Annex A1). Thus, each of the four direct assessment Field Test booklets was estimated to take 60 minutes.

In contrast to the paper-and-pencil branch of the Field Test design, the computer-delivered branch included 21 testlets that were 60 minutes long, consisting of a pair of 30-minutes blocks of items in each testlet¹ (as shown in Figure 1.4). As reflected in this design, each of the computer-delivered testlets contained only literacy tasks, only numeracy tasks, both literacy and problem-solving tasks, both numeracy and problem-solving tasks, or only problem-solving tasks. Overall, for the Field Test, there were 13 blocks that are 30 minutes long, grouped to form the 21 testlets: four blocks of literacy tasks (L1-L4), four blocks of numeracy tasks (N1-N4) and five blocks of problem solving tasks (PS1-PS5), as illustrated in Annex A2. The administration of these 21 testlets followed the administration of the BQ, including the ICT core as described above.

¹ The CBA comprised intact clusters of items that were grouped following a predetermined format. These groupings were not visible to users but are still called testlets for reference.

In this design, the direct assessment time was 60 minutes, on average, and each item was expected to be answered by a minimum of 150 adults based on an estimate of 1,500 respondents per country/per language (i.e., completed cases): 1,100 for the computer-delivered test and 400 for the paper-and-pencil test.

Figure 1.1: Paper-and-pencil Field Test assessment design, integrated



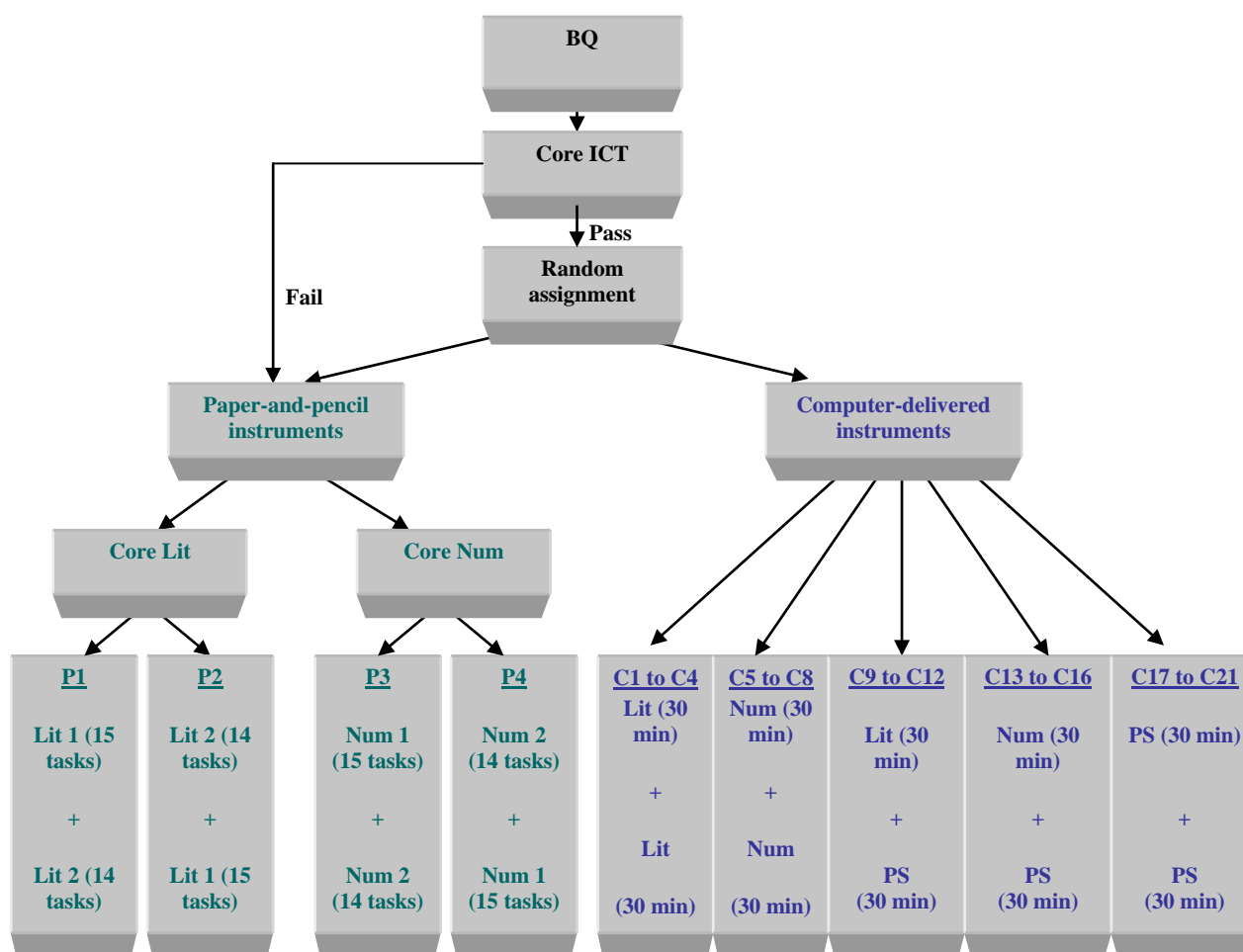
1.3 The role of international options in the Field Test assessment design

Some of the domains that were tested in the direct assessment were identified as international options. Which options were chosen by each of the participating countries had an impact on the Field Test and Main Study designs as well as on the required sample size.

1.3.1 Reading component skills as an international option

A country's decision not to assess reading components (one of the international options) had minimal impact on the overall Field Test design, as shown in Figure 1.2. Countries choosing not to include the reading components measures saved about 10 minutes in the overall assessment time and were able to reduce their sample size by a total of 100 adults. Under this design, assessment time was estimated to be 50 minutes, each item was expected to be answered by 150 adults, and the design was based on an estimate sample of 1,400 respondents per country/per language (i.e., completed cases): 1,100 who respond to the computer-delivered instruments and 300 who respond to the paper-and-pencil booklets.

Figure 1.2: Paper-and-pencil Field Test assessment design, without reading components

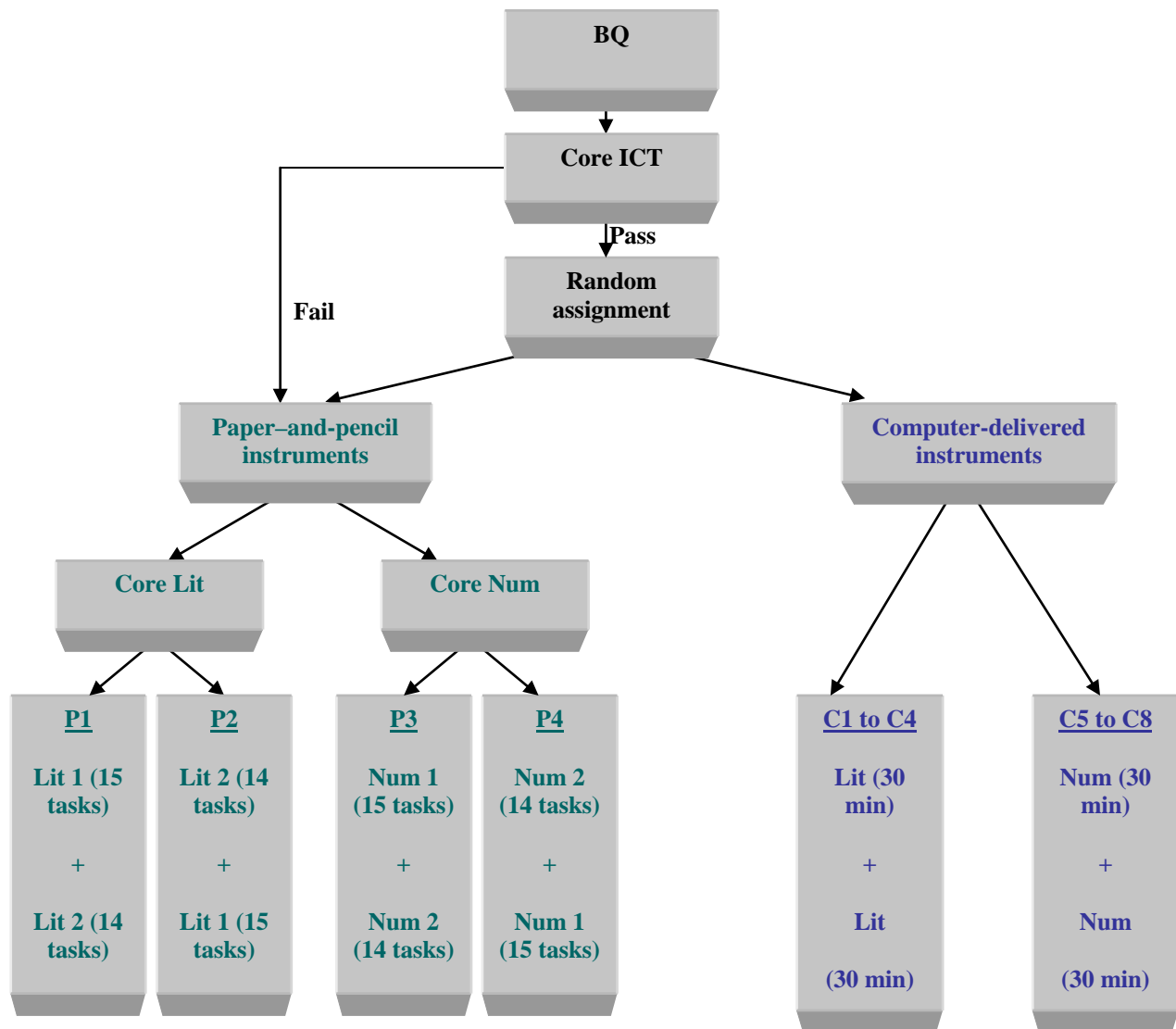


1.3.2 Problem solving in technology-rich environments as an international option

The international option to include reading components but not to assess problem solving had a significant impact on both the sample size needed for the Field Test as well as on the number of

computer-based booklets. This is shown in Figure 1.3. To compensate for the lack of covariance information, the number of respondents per item was increased but the overall sample size reduced by some 300 completed cases. In this design, assessment time per individual remained at 60 minutes, each item was answered by 200 adults, and was based on an estimate of 1,200 respondents per country/per language (i.e., completed cases): 800 who responded to the computer-delivered measures and 400 who responded to the paper-and-pencil items.

Figure 1.3: Paper-and-pencil Field Test assessment design, without problem solving



1.4 Item development needs

The item development requirements and goals for the literacy and numeracy domains are shown in Table 1.1. Overall, the Main Study required 24 items in each domain for the paper-and-pencil

assessment and 48 items for the computer-delivered measures in each of the two domains. Of these, some 19 paper-and-pencil and 29 computer-delivered items were needed in each domain to serve as linking items. Linking items refer to items selected from IALS and ALL that were used to establish the link between PIAAC and these previous studies and between paper-and-pencil and computer-delivered formats. In order to meet these goals for each domain, it was necessary to develop and assess a larger pool of items for the Field Test.

The Field Test item pool required a total of 35 paper-and-pencil literacy and 35 paper-and-pencil numeracy items. The computer version needed 72 items for each domain. Of these, 42 were used to evaluate their utility as linking items for the computer-delivered measures while a subset of 25 was used to evaluate their utility for linking the paper-and-pencil and computer-delivered formats.

Table 1.1: Literacy and numeracy development item needs for PIAAC

Literacy or numeracy item development needs	Field Test		Main Study	
	Link	New	Link	New
Paper-based	25	10	19	5
Computer-based	42	30	29	19

As a new construct and domain for adult surveys, the assessment of problem solving in technology-rich environments involved scenarios of varying levels of complexities. Scenarios were designed to take between an average of five to 15 minutes to complete. Overall, 150 minutes of testing material was developed for the Field Test (approximately 16 scenarios of varying lengths) with some 75 minutes of problem solving in technology-rich environment tasks selected for inclusion in the Main Study (approximately eight scenarios of varying lengths). The scenarios finally selected for the Main Study were organized into a pair of 25-minute blocks.

Reading component measures also were constructed according to the framework developed by the literacy expert group. These measures focused on speed and accuracy and were assessed in a limited amount of time. A total of 20 minutes was allotted for the Main Study to measure several of these skills with final measures assembled from 40 minutes worth of Field Test data.

1.5 Main Study goals and design

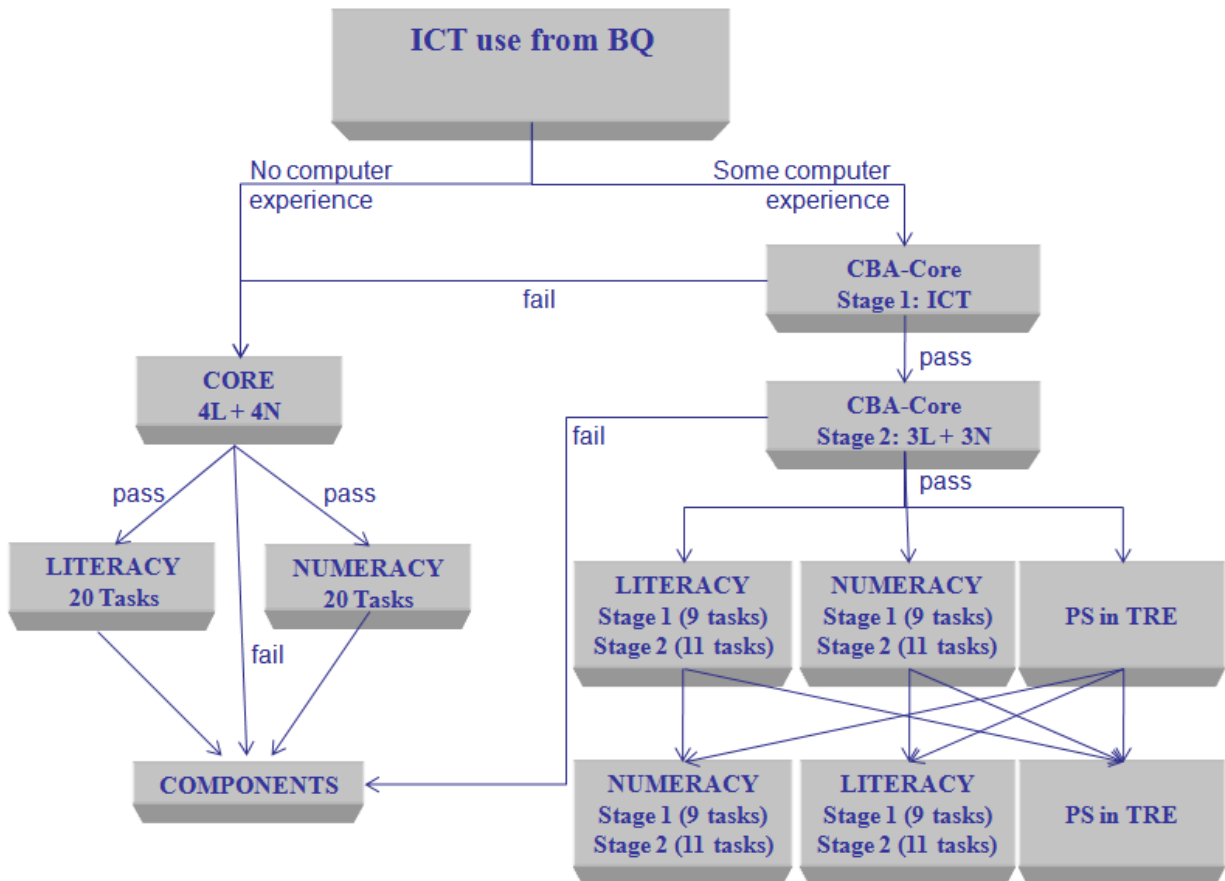
The assessment design for the Main Study served two primary goals, to: 1) provide good measurement of all the domains included in PIAAC and 2) provide a baseline for assessing trends or changes over time in future rounds of PIAAC or similar assessments. The PIAAC assessment design for the Main Study was based on an assumption of 60 minutes of testing time, on average, for the direct assessment. As PIAAC was not a timed assessment, some respondents were expected to take longer to complete the survey.

The Main Study design was implemented using the design illustrated below, where L represents literacy tasks, N represents numeracy tasks and PSTRE represents tasks involving problem solving in technology-rich environments. Among other things, the BQ asked about the respondent's computer experiences, which was essential to branch respondents to either the paper-and-pencil or CBAs at the conclusion of the BQ. Respondents with no computer experience based on BQ questions G_04 and/or the H_04a were routed to the paper branch, as were respondents refusing to take the test on the computer. The remainder of respondents were routed to the computer branch of the survey.

As shown in the figure above, the Main Study had the following characteristics:

- The paper-delivered branch of the assessment included a 10-minute core assessment of literacy and numeracy skills. Respondents who performed at or above a minimum standard were randomly assigned to a 30-minute cluster of literacy or numeracy items, followed by a 20-minute assessment of component skills. The relatively small proportion of respondents who performed poorly on the paper-and-pencil core tasks skipped the literacy and numeracy items and were routed directly to the reading component skills measures.
- The computer-delivered branch of the assessment first directed respondents to the CBA Core section, which was composed of two stages taking approximately five minutes each. Poor performance on either stage of the computer-based CBA Core section resulted in switching over to the appropriate sections of the paper-and-pencil instruments. Respondents who failed CBA Core Stage 1 (which contained ICT related tasks) were directed to begin the paper-based Core section and proceed with the process outlined in the above bullet. Respondents who passed CBA Core Stage 1 but failed CBA Core Stage 2 (which contains six cognitive items) were then administered only the reading components tasks. Respondents who performed well on the both CBA Core sections were routed to one of three possible outcomes (each taking approximately 50 minutes): 50% of respondents received a combination of literacy and numeracy tasks, 33% received problem solving combined with either literacy or numeracy, and 17% received only problem-solving sections.

Figure 1.4: Integrated Main Study assessment design



It is also important to note that PIAAC was the first international comparative survey to include multistage adaptive testing as part of the Main Study. The Main Study CBA for literacy and numeracy, represented by each numeracy or literacy block in Figure 1.4, was organized according to the design shown here in Table 1.2. As noted here, the literacy and numeracy modules each consisted of two stages. Each stage contained a number of testlets varying in difficulty. In each stage, only one testlet was delivered to a respondent. Within each of these modules, a respondent took 20 items (nine items in Stage 1; 11 in Stage 2). Thus, respondents taking literacy in Module 1 and numeracy in Module 2 (or vice versa) answered 40 items. Each module was designed to take an average of 30 minutes.

Problem solving in technology-rich environments (PSTRE) is unique because of the nature of the domain. It was organized as two fixed sets of tasks: seven in Module 1 and seven in Module 2. These were also designed to take an average of 30 minutes.

Table 1.2: Design of the Main Study CBA instruments for literacy and numeracy in the integrated design

STAGE 1							
(18 unique tasks – 9 tasks per testlet. Each respondent takes 1 testlet)							
	Block A1	Block B1	Block C1	Block D1			
Testlet 1-1	4 tasks	5 tasks					
Testlet 1-2		"	4 tasks				
Testlet 1-3			"	5 tasks			
STAGE 2							
(31 unique tasks – 11 tasks per testlet. Each respondent takes 1 testlet)							
	Block A2	Block B2	Block C2	Block D2	Block E2	Block F2	Block G2
Testlet 2-1	6 tasks	5 tasks					
Testlet 2-2		"	3 tasks	3 tasks			
Testlet 2-3				"	3 tasks	5 tasks	
Testlet 2-4						"	6 tasks

However, due to the diversity of the participants' country, language, and educational backgrounds, a deterministic assignment of stages would likely have resulted in certain subpopulations being exposed to only a small percentage of items created for the assessment. To help mitigate the potential impact of such a situation, a set of conditional probability tables of item exposure rates for specified subpopulations was developed. By adjusting these parameter values, a balance between the adaptiveness of the assessment and the predetermined item exposure rates for the given subpopulations was achieved.

Choice of first module: For the computer branch, the selection of a domain (literacy, numeracy or problem solving) for the first module was random. The choice was determined by a random number between 0 and 1 that was generated by the system. A literacy module was chosen if the random number was less than 0.3333333, a numeracy module was chosen if the number was equal to or greater than 0.3333333 and less than 0.6666666, and a problem-solving module if the random number was equal to or greater than 0.6666666.

In problem solving, all respondents took a problem-solving orientation followed by the same set of tasks. In literacy and numeracy, because of the adaptive design, respondents also received the associated orientation but were then assigned to one of the three testlets in Stage 1.

Choice of Stage 1 testlet within literacy and numeracy: The literacy and numeracy testlets in Stage 1 varied in difficulty. There were three levels of testlets: easy (Testlet 1), medium (Testlet 2) and difficult (Testlet 3). Three variables determined which testlet was chosen for a respondent:

- Education level (EdLevel3) from the BQ: Levels were low, medium or high
- Native versus nonnative speaker: The respondent was considered a native speaker if his or her first language was one of the assessment languages
- CBA-Core Stage 2 score: Passing scores between 3 and 6

These three variables were organized in a matrix that results in two threshold numbers. The following matrix provides an example, using Stage 1 selection as explained below in Table 1.3.

Table 1.3: Example of matrix design for Stage 1 selection of literacy and numeracy testlets

EdLevel3:		Low		Low		Medium		Medium		High	
Native Speaker:		No		Yes		No		Yes		Both	
Threshold:		I	II	I	II	I	II	I	II	I	II
CBA-Core Stage 2 Score	0	0.900	0.950	0.872	0.922	0.850	0.900	0.822	0.872	0.800	0.850
	1	0.738	0.945	0.710	0.917	0.688	0.895	0.660	0.867	0.638	0.845
	2	0.607	0.924	0.579	0.896	0.557	0.874	0.529	0.846	0.507	0.824
	3	0.505	0.887	0.477	0.859	0.455	0.837	0.427	0.809	0.405	0.787
	4	0.433	0.834	0.405	0.806	0.383	0.784	0.355	0.756	0.333	0.734
	5	0.392	0.765	0.364	0.737	0.342	0.715	0.314	0.687	0.292	0.665
	6	0.380	0.680	0.352	0.652	0.330	0.630	0.302	0.602	0.280	0.580

As shown in the matrix above, if a respondent had a high education level, was a native speaker, and scored high on the CBA-Core Stage 2 (for a total score of 6), he or she would be assigned 0.280 and 0.580 as thresholds. Then a random number between 0 and 1 was generated. This respondent received the easier testlet if the random number was less than 0.280; the medium test if equal to or greater than 0.280 and less than 0.580; and the difficult test if equal to or greater than 0.580. This process ensured that respondents who were native speakers, highly educated, and performed well on the core were most likely to receive the most difficult testlet at the first stage compared to other testlets. However, there was some probability they would receive one of the other easier testlets.

Choice of second testlet for literacy and numeracy module (1): The four literacy and numeracy testlets in Stage 2 also varied in difficulty, with Testlet 1 being the easiest and Testlet 4 the most difficult. For this scenario, three thresholds were defined because there was one more category than in Stage 1. Thus, the test assignment for Stage 2 depended on the following three variables as shown in Table 1.4:

Table 1.4: Example of matrix design for Stage 2 selection of literacy and numeracy testlets

EdLevel3:		Low			Low			Medium			Medium			High		
Native Speaker:		No			Yes			No			Yes			Both		
Threshold:		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
CBA-Core Stage 2 + Testlet 1 Score	0	0.800	0.900	1.000	0.775	0.875	0.975	0.750	0.850	0.950	0.725	0.825	0.925	0.700	0.800	0.900
	1	0.735	0.871	0.998	0.710	0.846	0.973	0.685	0.821	0.948	0.660	0.796	0.923	0.635	0.771	0.898
	2	0.673	0.841	0.993	0.648	0.816	0.968	0.623	0.791	0.943	0.598	0.766	0.918	0.573	0.741	0.893
	3	0.616	0.812	0.986	0.591	0.787	0.961	0.566	0.762	0.936	0.541	0.737	0.911	0.516	0.712	0.886
	4	0.563	0.783	0.977	0.538	0.758	0.952	0.513	0.733	0.927	0.488	0.708	0.902	0.463	0.683	0.877
	5	0.513	0.753	0.965	0.488	0.728	0.940	0.463	0.703	0.915	0.438	0.678	0.890	0.413	0.653	0.865
	6	0.468	0.724	0.951	0.443	0.699	0.926	0.418	0.674	0.901	0.393	0.649	0.876	0.368	0.624	0.851
	7	0.427	0.695	0.934	0.402	0.670	0.909	0.377	0.645	0.884	0.352	0.620	0.859	0.327	0.595	0.834
	8	0.389	0.665	0.915	0.364	0.640	0.890	0.339	0.615	0.865	0.314	0.590	0.840	0.289	0.565	0.815
	9	0.356	0.636	0.894	0.331	0.611	0.869	0.306	0.586	0.844	0.281	0.561	0.819	0.256	0.536	0.794
	10	0.327	0.607	0.870	0.302	0.582	0.845	0.277	0.557	0.820	0.252	0.532	0.795	0.227	0.507	0.770
	11	0.301	0.577	0.844	0.276	0.552	0.819	0.251	0.527	0.794	0.226	0.502	0.769	0.201	0.477	0.744
	12	0.280	0.548	0.815	0.255	0.523	0.790	0.230	0.498	0.765	0.205	0.473	0.740	0.180	0.448	0.715
	13	0.263	0.519	0.784	0.238	0.494	0.759	0.213	0.469	0.734	0.188	0.444	0.709	0.163	0.419	0.684
	14	0.249	0.489	0.751	0.224	0.464	0.726	0.199	0.439	0.701	0.174	0.414	0.676	0.149	0.389	0.651
	15	0.240	0.460	0.715	0.215	0.435	0.690	0.190	0.410	0.665	0.165	0.385	0.640	0.140	0.360	0.615

- Education level (EdLevel3) from the BQ: Levels were low, medium or high
- Native versus nonnative speaker: The respondent was considered a native speaker if his or her first language was one of the assessment languages

- CBA-Core Stage 2 score plus Stage 1 score: CBA-Core Stage 2 passing scores were between 3 and 6 while the results of Stage 1 were between 0 and 9

These three variables are also organized in a matrix that resulted in three threshold numbers (see matrix below as an example). However, there are now three different matrices, depending on which testlet (easy, medium or difficult) the respondent came from in Stage 1. The appropriate matrix was chosen and the variables were compared with the matrix. This resulted in three threshold numbers for the respondent.

Again, if a respondent had a high education level, was a native speaker, and scored high on the CBA-Core Stage 2 (for example a total score of 6) and had the highest score in Stage 1 (a 9), he or she would be assigned thresholds of 0.140, 0.360 and 0.615. Then a random number between 0 and 1 was generated. Thus, this respondent would have received Testlet 1 (easiest) if the random number was less than 0.140, Testlet 2 if equal to or greater than 0.140 and less than 0.360, Testlet 3 if equal to or greater than 0.360 and less than 0.615, or Testlet 4 (most difficult) if equal to or greater than 0.615.

Choice of second module: After completing Module 1 (either the two testlets for literacy or numeracy or the problem-solving module), the respondent proceeded to Module 2. The selection between Module 1 and Module 2 was also based on random probabilities. Thus, a random number between 0 and 1 was generated again.

- If the respondent completed Literacy as Module 1, he or she was assigned Numeracy as Module 2 (starting with numeracy orientation) if the random number was less than 0.75. Otherwise he or she continued with Problem Solving as Module 2 (starting with PS orientation).
- If the respondent completed Numeracy as Module 1, he or she was assigned Literacy as Module 2 (starting with literacy orientation) if the random number was less than 0.75. Otherwise he or she continued with Problem Solving as Module 2 (starting with PS orientation).
- If the respondent completed Problem Solving as Module 1, he or she was assigned Literacy Module 2 (starting with the literacy orientation) if the random number was less than 0.25, Numeracy Module 2 (starting with the numeracy orientation) if the random number was equal to or greater than 0.25 but less than 0.50, or Problem Solving Module 2 if the random number was equal to or greater than 0.50 (without the PS orientation, which he or she would have already received in Module 1).

After completing the paper or computer branches, the interview continued to the Exit Module, where the interviewer thanked the respondent for participating and provided an incentive, if applicable. The interviewer then continued to the case finalization by answering a set of general questions about the circumstances under which the interview took place, called ZZ-questions.

1.6 Summary and conclusions

This document describes and illustrates the goals and assessment design for both the Field Test and Main Study. The multiple goals of the Field Test illustrate its importance in successfully implementing the Main Study. It was intended to address help evaluate four key areas – operational, platform, instrumentation and scaling and psychometric characteristics. The fact that the results of PIAAC were being linked to previous assessments while being implemented in both paper and computer mode – while also including multistage adaptive testing – added to the importance of the Field Test. Information generated during the Field Test was used to help establish the adaptive portion of the Main Study.

The integrated design included the four cognitive domains as specified in the original terms of reference. As the OECD and the participating countries identified reading components and problem solving in technology-rich environments as international options, alternative designs were also illustrated and described in this chapter. Within the four domains and two formats of PIAAC, the described designs brought innovative aspects and important benefits to the overall goal of producing outcomes that are both valid and comparable across countries.

The Field Test data were used to not only evaluate the procedures and quality of the platform and instruments but to serve to establish the feasibility of linking over time and across modes. The design and data from the Main Study not only expands the range of what can be measured in adult surveys but also how they are measured. More importantly, this information in combination with that gained from the BQ and JRA module described elsewhere in this report provides policymakers and others with a rich source of information to understand the distributions of human capital in their country and the connections between these skills and important social, educational and labor market outcomes. The information from the Main Study was also used to adjudicate the quality of each country's data. This information was shared with the OECD Secretariat, the Board of Participating Countries and all National Project Managers.

ANNEX A1. PAPER-AND-PENCIL INSTRUMENTS

Field test, paper-based instruments in the integrated design where P1-P4 present paper booklets, CL represent the core literacy cluster, CN represent the core numeracy cluster, L1-L2 represent literacy clusters, and N1-N2 represent numeracy clusters.

Paper-based instruments	Clusters			
	Core (10 minutes)	1 (20 minutes)	2 (20 minutes)	3 (10 minutes)
P1	CL (6 Lit tasks)	L1 (15 Lit tasks)	L2 (14 Lit tasks)	Components A
P2	CL (6 Lit tasks)	L2 (14 Lit tasks)	L1 (15 Lit tasks)	Components B
P3	CN (6 Num tasks)	N1 (15 Num tasks)	N2 (14 Num tasks)	Components C
P4	CN (6 Num tasks)	N2 (14 Num tasks)	N1 (15 Num tasks)	Components D

ANNEX A2. COMPUTER-BASED INSTRUMENTS

Field test, computer-based instruments with the assessment of reading components where C1-C21 represent computer booklets, L1-L4 represent literacy clusters, N1-N4 represent numeracy clusters, and PS1-PS5 represent problem solving clusters

Computer-based instruments	Cluster 1 (30 min)	Cluster 2 (30 min)
C1	L1	L2
C2	L2	L3
C3	L3	L4
C4	L4	L1
C5	N1	N2
C6	N2	N3
C7	N3	N4
C8	N4	N1
C9	L1	PS1
C10	L2	PS2
C11	L3	PS3
C12	L4	PS4
C13	N1	PS2
C14	N2	PS3
C15	N3	PS4
C16	N4	PS5
C17	PS1	PS2
C18	PS2	PS3
C19	PS3	PS4
C20	PS4	PS5
C21	PS5	PS1

Chapter 2: The Development of the PIAAC Cognitive Instruments

Mary Louise Lennon and Claudia Tamassia, ETS

2.1 Introduction

As the first computer-based, large-scale assessment of adult skills, PIAAC was designed to reflect the changing nature of information, its role in society and its impact on people's lives. As a result, the cognitive instruments developed for PIAAC differed from those in earlier adult assessments in several important ways.

- For the first time, this assessment addressed literacy in digital environments. As a computer-based assessment, PIAAC was able to include tasks that required respondents to use electronic texts including Web pages, emails and discussion boards. These stimulus materials included hypertext and multiple screens of information and simulated real-life literacy demands presented by digital media.
- The definition of numeracy in PIAAC was broadened from that used in earlier assessments and included the ability to access, use, interpret, and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life. The inclusion of “engage” in the definition signaled that not only cognitive skills but also dispositional elements, that is, beliefs and attitudes, are necessary to effectively meet the demands of numeracy in everyday life.
- PIAAC also included a new domain: problem solving in technology-rich environments (PSTRE). This was the first attempt to assess such a construct on a large scale and as a single dimension. PSTRE included computer-based simulation tasks designed to measure the ability to analyze various requirements of a task, define goals and plans, and monitor progress until task purposes were achieved. The focus was not on computer skills per se, but rather on the cognitive skills required to access and make use of computer-based information to solve problems.
- Finally, PIAAC included a reading components domain, which included measures of vocabulary knowledge, sentence processing and passage comprehension. The inclusion of this domain provided more information about the skills of individuals with low levels of literacy proficiency than had been available from previous international assessments. This was important because to have a full picture of literacy in any society, it is necessary to have information about adults with lower skill levels as it is these individuals who are at greatest risk of negative social, economic and labor market outcomes.

While PIAAC introduced significant new elements to the assessment of adult skills in an international context, key aspects of previous surveys were employed as well. In particular, like the earlier assessments to which PIAAC was linked, this development work was based on frameworks that defined the assessment constructs for each domain as well as features of the tasks designed to measure those constructs.

2.2 Defining the domains: The PIAAC cognitive frameworks

The frameworks for each of the three cognitive domains – literacy (including reading components), numeracy, and problem solving in technology-rich environments – were developed using the same process and methodology. Following Messick’s (1994) construct-centered approach, the expert group for each domain defined the construct to be measured, the performances or behaviors expected to reveal that construct, and the task characteristics to be used in building assessment tasks to elicit those behaviors. The overall goal of this process, which included the steps described below, was to explicitly lay out the inferences and assumptions about what was to be measured and how the results would be interpreted and reported.

1. Defining the domain

Each expert group began by developing a working definition of the domain and the assumptions underlying that definition. Such a definition is an important step in developing an assessment framework as it sets the boundaries for what will and will not be measured.

2. Organizing the domain

Once the definition was developed, the experts described the kinds of tasks that represent the skills and abilities included under that definition. Those tasks were then categorized to inform test design and, ultimately, score reporting. The goal of this step was to develop a coherent representation of the domain that would permit policymakers and others to summarize and report information in useful ways.

3. Identifying task characteristics

Step 3 involved identifying a set of key characteristics, or task models, that formed the basis for constructing the assessment tasks. These models defined characteristics of the stimulus materials to be used as well as characteristics of the tasks presented to respondents. Examples of task characteristics used in PIAAC include contexts, material or text types, and task types, which include the cognitive processes or strategies required to complete a given task.

4. Identifying and operationalizing variables

In order to use the task characteristics in designing the assessment and, later, in interpreting the results, the variables associated with each task characteristic needed to be defined. These definitions are typically based on existing literature and on experience with building and conducting other large-scale assessments.

This information allowed item developers to categorize stimulus materials as well as the items they constructed so they could be used in reporting results.

As an example, the literacy framework provided further definition of three key task characteristics in that domain: context, text and task type. “Contexts” were defined to include work and occupation, personal uses (home and family, health and safety, etc.), community and citizenship, and education and training. The expert group specified that “texts” could be classified according to medium (print or digital), format, and text type (description, narration, exposition, etc.), and “task types” were defined to include tasks that required respondents to access and identify information, integrate and interpret texts, and evaluate and reflect on information.

Additional steps that follow the Main Study data collection include work to validate the variables that were used to develop the assessment tasks. This includes data analysis to determine which of the variables account for large percentages of the variance in the distribution of tasks and thereby contribute most towards understanding task difficulty and predicting performance. The goal of this analysis is to provide empirical evidence that a set of variables can be identified that summarizes some of the skills and strategies that are involved in accomplishing various kinds of tasks. Finally, an interpretative scheme is built that uses the validated variables to explain task difficulty and examinee performance. The definition of the proficiency levels for each scale, described in greater detail in Chapter 22, is an example of such an interpretative scheme. For previous large-scale literacy assessments, including IALS and ALL, developing these interpretations has provided a useful means for exploring the progression of information-processing demands across each of the scales and for defining what scores along a particular scale mean. In this way, the interpretative scheme contributes to the construct validity of inferences based on scores from the measure on which it is based (Messick, 1989).

The following sections summarize key aspects of the frameworks for the cognitive domains assessed in PIAAC: literacy, reading components, numeracy and problem solving in technology rich environments. The complete framework documents can be accessed at the OECD site at <http://www.oecd.org/site/piaac/publications.htm>.

2.2.1 Literacy

2.2.1.1 Definition of the domain

In PIAAC, literacy was defined as understanding, evaluating, using and engaging with written texts to participate in society, to achieve one’s goals, and to develop one’s knowledge and potential.

2.2.1.2 Categorizing texts (task characteristics)

A number of variables were used to categorize texts in the PIAAC literacy assessment, including the following:

- **Medium**

Texts were distinguished as either digital (electronic) texts or print texts. A text that could be reproduced in print exactly as it appears on a screen was considered to be a *print* text.

That is, merely being displayed on a computer screen was not a sufficient condition for classification as a digital text. Texts that could not be reproduced in print with all of their features intact were considered *digital* texts.

- **Format**

Texts were also classified as either continuous or noncontinuous, with those containing both elements classified as “mixed.” Continuous texts are made up of sentences formed into paragraphs. Examples include newspaper articles, brochures, manuals, email and many Web pages. Noncontinuous texts, or matrix documents, include tables, graphs, charts and forms.

- **Type**

Text types (rhetorical stances) constitute ways of organizing continuous texts in terms of their content and the purpose of the author. Six types of rhetorical stances were identified for PIAAC including: description, narration, exposition, argumentation, instruction and records.

- **Social context**

The context in which reading takes place may influence the motivation to read and the manner in which texts are interpreted. Therefore, the expert group specified that stimulus materials for the assessment should be drawn from a range of contexts, including: work and occupation, personal (home and family, health and safety, consumer economics, and leisure and recreation), community and citizenship, and education and training

2.2.1.3 Aspects of tasks

Literacy tasks in the PIAAC assessment were designed to address three broad cognitive strategies identified as necessary for achieving a full understanding of texts:

- *access and identify tasks* require respondents to locate information in a text,
- *integrate and interpret tasks* involve relating parts of one or more texts to each other, and
- *evaluate and reflect tasks* require the respondent to draw on knowledge, ideas or values external to the text to evaluate aspects including accuracy, reliability and timeliness.

2.2.1.4 Factors that affect task difficulty

Finally, the Literacy Expert Group defined a number of key factors for item developers to keep in mind as tasks were developed along the continuum from easier to harder.

- **Transparency of information**

One factor affecting task difficulty is the transparency of information in the text as it relates to the presented task or question. A question that explicitly refers to literal information in a text is generally easier to process and therefore tends to be an easier task along the Literacy scale.

- **Degree of complexity in making inferences**

Complexity of inferences can be impacted by the extent to which respondents need to recognize paraphrased information, make high-level text inferences, and employ extra-textual inferences.

- **Semantic and syntactic complexity**

Tasks requiring the reader to identify concrete information such as persons, things or places tend to be easier than those involving abstract properties, such as goals, conditions and purposes. The grammatical structure of the question posed or the stimulus text can also make a task more or less complex. For instance, negative phrases are more complex than affirmative phrases. The presence of subordinate clauses is an example of another feature that can increase the complexity of syntactic processing.

- **Amount of information needed**

The amount of text that must be processed plays a role in the difficulty of any task. The more information a respondent needs from the text to complete the task, the more difficult that task will be.

- **Prominence of the information**

Task difficulty can also be impacted by the location of relevant information in a text. It is easier to access information in a prominent location such as in the first or last sentence of a paragraph, in a main, rather than subordinate, clause, or at the top or bottom of a list.

- **Competing information**

Task difficulty can be impacted by the amount of potentially relevant information the reader has to sift through to access information needed to complete that task. For example, if a text includes telephone, fax and mobile numbers, it will be more difficult for the reader to find the fax number than if the text includes only the fax number.

- **Text features**

The degree to which the reader has to construct relationships among parts of the text affects difficulty. For example, tasks that require respondents to sort out anaphoric references or which include text where cohesion signals are absent tend to be more difficult.

2.2.1.5 Item development goals for literacy

As part of its work, the Literacy Expert Group was asked to define overall item development targets across the three defined task characteristics of text type, context, and process. For text type, the goal was that 70-80% of the items would be based on print texts and 20-30% on digital texts. The higher percentage of print texts was dictated in large measure by the number of linking items required by the PIAAC assessment design, as those items were developed for paper-based assessments. Both the print and digital categories included continuous and noncontinuous texts.

To ensure a range of contexts in the assessment tasks, the overall targets were to have 15% of items in the work context, 40% in personal, 30% in community, and 15% in education. In terms of task aspects, the framework goals included 40% of the items in the access and identify category, 45% in integrate and interpret, and 15% in evaluate and reflect.

Reading components

In previous assessments of adult literacy, the information gathered on the reading abilities of adults with poor skills was often insufficient to gain a proper understanding of their difficulties due to the small number of items at low difficulty levels. To redress this problem, the literacy

framework for PIAAC included a component test intended to provide more information about the abilities of those with low levels of literacy.

The components assessment framework was based on the principle that comprehension – the process of constructing meaning when reading – is built on knowledge of how a given language is represented in its writing system and through component print-reading skills. Evidence of an individual’s level of print-reading skills can be captured in tasks that examine a reader’s ability and efficiency in processing the elements of the written language, including letters/characters, words, sentences, and larger, continuous segments of text.

A second guiding principle is that the assessment of component skills aims to evaluate the extent to which adults can apply their existing language and comprehension skills to the processing of printed texts. The components tasks were not designed to separately assess the level of language skills in the target writing system and the literacy skills assessed in the main literacy survey. Nonnative speakers of the language of the assessment who have only basic oral vocabulary, syntactic/grammatical and linguistic comprehension skills were expected to show poor performance on component reading tasks. As a consequence, low levels of proficiency in the language of the assessment were not differentiated from low literacy skills in the component tasks.

A third guiding principle is that the levels of proficiency, efficiency and integration of component skills are indicative of the levels of reading development and learning potential. As skills and knowledge accumulate, the ease of processing familiar, text-based print increases. Component efficiency is typically indexed by assessing speed or rate of processing, as well as accuracy. For PIAAC, although the reading components assessment was the one domain assessed only in paper-and-pencil form, interviewers timed respondents and recorded that information as part of the measure of efficiency.

It was also assumed that the set of component items administered in each country reflected the linguistic characteristics of the language of assessment. As the relationship of the language to the writing system was anticipated to be very different in different languages, the nature of the items used to assess the components was adapted based on consideration of those differences in order to best ensure comparability across languages. Countries were provided with very specific adaptation guidelines and training on how to adapt the reading components measures for their language(s) of assessment. As was true for the other domains, trained verifiers reviewed these adaptations and provided feedback to countries as needed.

The PIAAC components assessment included tests of vocabulary, sentence processing, and basic passage comprehension. In skilled reading, these components are integrated to support literacy performance. During acquisition, even by adults, these components may be measured separately, with different profiles having implications for learning, instruction, and policy.

2.2.2 Numeracy

2.2.2.1 Definition of the domain

PIAAC defined numeracy as *the ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life.*

Because numeracy is a broad, multifaceted construct referring to a complex competency, the definition of numeracy was coupled with a more detailed definition of *numerate behavior* and with further specification of the facets of numerate behavior. The expert group felt this was necessary for the operationalization of the construct of numeracy in PIAAC and to broaden the understanding of key terms appearing in the definition itself. The definition of numerate behavior adopted for PIAAC was as follows, with key facets or task characteristics associated with numerate behavior shown in Table 2.1.

Numerate behavior involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways.

Table 2.1: Numerate behavior – key facets and their components

Numerate behavior involves managing a situation or solving a problem...	
1. in a real context:	<ul style="list-style-type: none"> – everyday life – work – society – further learning
2. by responding:	<ul style="list-style-type: none"> – identify, locate or access – act upon and use: order, count, estimate, compute, measure, model – interpret – evaluate/analyze – communicate
3. to mathematical content/information/ideas:	<ul style="list-style-type: none"> – quantity and number – dimension and shape – pattern, relationships, change – data and chance
4. represented in multiple ways:	<ul style="list-style-type: none"> – objects and pictures – numbers and mathematical symbols – formulae – diagrams and maps, graphs, tables – texts – technology-based displays
5. Numerate behavior is founded on the activation of several enabling factors and processes:	<ul style="list-style-type: none"> – mathematical knowledge and conceptual understanding – adaptive reasoning and mathematical problem-solving skills – literacy skills – beliefs and attitudes – numeracy-related practices and experience – context/world knowledge

2.2.2.2 Principles for assessing numeracy in PIAAC

The development of the numeracy assessment for PIAAC was based on a number of general principles or guidelines, as listed below:

- **Items should cover as many aspects as possible within each of the four facets of the numeracy competency.**
Items should require the activation of a broad range of skills and knowledge included in the construct of numeracy.
- **Items should aspire to maximal authenticity and cultural appropriateness.**
Tasks should be derived from real-life stimuli and pertain to a range of contexts or situations (i.e., everyday life, work, society, further learning) that can be expected to be of importance or relevance in the countries participating in PIAAC. Item content and questions should appear purposeful to respondents across cultures.
- **Items should have a free-response format, to the extent feasible within the computer platform used for administering the direct assessments.**
Items should be structured to include a stimulus (e.g., a picture, drawing, visual display) and one or more questions, the answers to which the respondent communicates via the modes available within the test platform, primarily: numeric entry, click, highlight a region of the stimulus, or use of various pull-down menus.
- **Items should spread over different levels of ability**
Items should span the range of ability levels anticipated among PIAAC participants, from low-skilled individuals to those with advanced competencies.
- **Items should represent the different response types**
Items should require the range of available response types. It was recognized that certain types of numeracy responses, especially those requiring the use of interpretation, evaluation, analysis and communication, could receive only partial coverage in the first cycle of PIAAC due to the constraints of automatic scoring.
- **Items should vary in the degree to which the task is embedded in text**
Some items should use relatively rich texts while others should use little or no text. This distribution aims to reflect the different levels of text involvement in real-world numeracy tasks, as well as minimize overlap with the literacy assessment.
- **Items should be efficient**
To allow for coverage of many key facets of the numeracy competency, a large number of diverse stimuli and questions should be included. However, given testing-time constraints, the use of short tasks is necessary, precluding items that can simulate extended problem-solving processes or require a lengthy open-ended response.

- **Items should be adaptable to unit systems across participating countries**
Items should be designed in a way that their underlying mathematical demands are as consistent as possible across countries, regardless of language and mathematical conventions. After being translated, items should retain equivalency with respect to their mathematical or cognitive demands.

2.2.2.3 Item development goals for numeracy

As was the case for literacy, part of the development work for the expert group included defining item development goals across the key facets of numeracy as defined in the framework. For response, or process, facets the goals included 50% of items in the act upon and use category, 10% in identify, locate or access, and 40% in interpret and evaluate. The framework specified that tasks should be based on real-life stimuli appropriate to a range of contexts or situations (i.e., everyday life, work, societal, further learning) without outlining specific proportions in each category. For mathematical content, development goals included a distribution of 25% of the items relating to data and chance, 25% dimension and shape, 20% pattern, relationships and change, and 30% quantity and change.

2.2.3 Problem solving in technology-rich environments (PSTRE)

2.2.3.1 Definition of the domain

PSTRE was broadly defined as *using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks.*

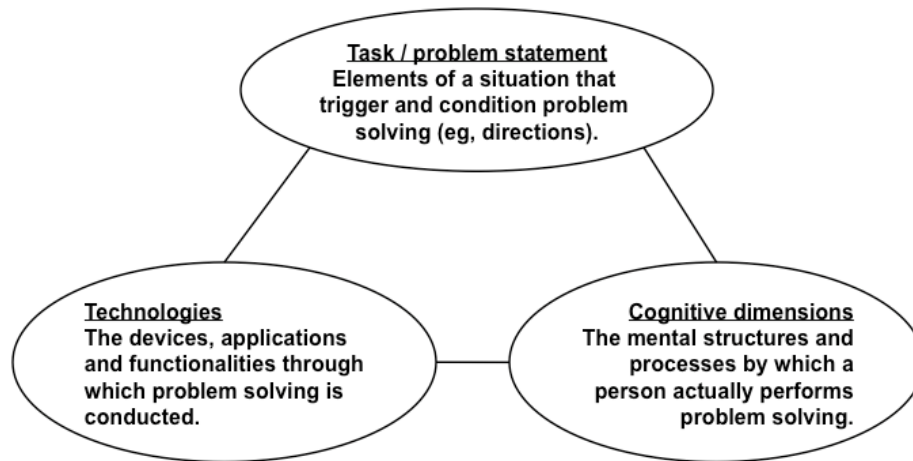
PIAAC represented the first attempt to assess PSTRE on a large scale and as a single dimension. This presented challenges in terms of the definition of tasks and the practical collection of data. Furthermore, digital technologies continue to evolve at a rapid pace, as do the personal, social and work-related uses of these technologies. While setting the stage for further rounds of assessment, the framework took into consideration issues of feasibility as well as the evolution of technology and its uses. In light of these challenges and constraints, the definition went on to further specify the scope of this first assessment of PSTRE for PIAAC:

The first PIAAC problem-solving survey focuses on the abilities to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, and accessing and making use of information through computers and computer networks.

2.2.3.2 Core dimensions of problem solving in technology-rich environments

The domain of PSTRE was conceived along three dimensions, as shown in Figure 2.1.

Figure 2.1: Core dimensions of problem solving in technology-rich environments



“Cognitive dimensions” include the mental structures and processes involved when a person solves a problem. These include setting goals and monitoring progress; planning; accessing and evaluating information; and making use of information by selecting, organizing and transforming information.

“Technologies” are the devices, applications and functionalities through which problem solving is conducted. These include hardware devices (laptop computers in the case of PIAAC); simulated software applications; commands and functions; and representations (text, graphics, etc.).

“Tasks” are the circumstances that trigger a person's awareness and understanding of the problem and determine the actions needed to be taken in order to solve the problem. Ordinarily, a wide range of conditions can initiate problem solving. For instance, a computer user may realize that his or her mailbox is crowded and that a new schema is needed for classifying emails. Alternatively, he or she may be faced with a complex issue (such as finding out more about a medical treatment) and decide to look for relevant information on the Web. In test-taking contexts, tasks are more explicitly assigned to respondents. They include the question and task instructions presented to respondents, as well as the specific materials and time constraints associated with the test.

Dimensions of the tasks being assessed in PIAAC PSTRE included:

- *Task purposes and contexts, including personal, work/occupation, and civic*
- *The intrinsic complexity of the problem*

Intrinsic complexity is related to a set of more specific variables: the minimum number of steps or actions required to solve the problem; the number of options at each phase; the diversity of operators and the complexity of mental reasoning and/or computation; the probability of impasses or unexpected outcomes; the number of constraints to be satisfied; and the amount of composition or transformation needed to communicate a solution.

- *The explicitness of the problem statement and task directions given to the respondent*
This dimension ranges from well-defined, explicit problem statements to implicit and ill-defined problem statements. A problem situation that requires the respondent to select operators and subgoals or define the successful achievement of a goal makes the problem more difficult.

2.2.3.3 PSTRE in relation to other domains of PIAAC

The constructs of literacy, numeracy and PSTRE rely on the same “core” cognitive processes. For example, tasks in all three domains require both an ability to decode printed symbols and a minimal working memory capacity. PSTRE also assessed a set of competencies distinct from those defined in the other two constructs.

The assessment of PSTRE in PIAAC focused on goal setting, monitoring and planning in technology-rich environments, and assessment tasks emphasized the problem-finding and problem-shaping processes typically found in these environments. Tasks included selecting an appropriate software application; deciding on one among several possible strategies; making use of adequate functionalities in a context-sensitive manner; interpreting ill-structured texts; and using online forms.

Respondents needed to complete problem-solving tasks in environments that involved multiple and complex sources of information. Some of the tasks required respondents to use and shift across multiple environments. PSTRE therefore assessed decision making with respect to the use of information sources (for example, choosing which environment to use or deciding whether or not to go to another website.) Evaluation was included as a critical underlying part of problem solving. Additionally, the selection of appropriate devices or tools took a prominent role in this domain.

In terms of processing information, problem solving is a specific construct in that it focuses on:

- the evaluation of sources in terms of reliability and the adequacy of information relative to the problem statement, as opposed to mere topical relevance, which is more applicable for literacy
- the integration of information across sources, especially in cases where the sources provide inconsistent information

PSTRE tasks sought to minimize the numeracy and literacy demands placed on respondents in order to increase the specificity and validity of the construct.

2.2.3.4 PSTRE and ICT competence

What differentiates the problem-solving domain from the general ICT domain? ICT skills may be broadly defined as “the interest, attitude, and ability of individuals to appropriately use digital technology and communication tools” (Lennon, et al., 2003). As is true for literacy and numeracy skills, ICT skills underlie PSTRE. However, the PSTRE construct aimed to encompass more than the purely instrumental skills related to the knowledge and use of digital technologies.

The cognitive dimensions of problem solving were considered the central object of the assessment, with the use of ICT as secondary.

2.2.3.5 Item development goals for PSTRE

Like literacy and numeracy, the PSTRE Expert Group defined targets for the distribution of items across the categories defined in the PSTRE framework. Based on the development of 25 tasks to be considered for the Field Test, goals included the distribution shown in Table 2.2. Additionally, the distribution across contexts was recommended to be 40% personal, 30% occupational and 30% civic. Finally, the task dimensions of intrinsic complexity and explicitness of the problem definition were specified as development variables as they were expected to influence the difficulty of items in the problem solving assessment.

Table 2.2: Distribution of PSTRE tasks as a function of environment and cognitive dimensions

Cognitive dimensions	Web environment	Spreadsheet environment	Email environment	Multiple environments
Goal setting and monitoring progress	2	1	1	1
Planning	2	2	2	4
Accessing and evaluating information	3	0	0	0
Selecting, organizing and transforming information	2	1	3	1
Totals	9	4	6	6

2.3 Developing the cognitive instruments

2.3.1 Overview

For each of the cognitive domains, test developers worked closely with the expert group to ensure that the instruments reflected the frameworks. All items were also submitted for country review to receive input on cultural and linguistic appropriateness as well as item content. In the case of literacy, developers from Australia and the United States attended each of the expert group meetings and the experts reviewed items throughout the development process. ETS developed the reading components tasks and the Literacy Expert Group reviewed those items as well. For numeracy, the expert group itself assumed primary responsibility for developing the PIAAC items. Test developers reviewed those items to ensure consistency in instructions, response modes and presentation across domains. ETS was primarily responsible for developing the PSTRE tasks and developers met with that expert group to receive input and reviews throughout the development process.

Two core requirements for PIAAC had important implications for development of the cognitive instruments. First, because the domains of literacy and numeracy had been measured in previous

large-scale international surveys, it was a requirement that PIAAC link back to the ALL and IALS. As a result, sets of linking items needed to be selected for literacy and numeracy that fit the requirements of the PIAAC assessment design. As described in the following section, transitioning those paper-based linking items to PIAAC's computer-based delivery mode required considerations related to display and response mode issues as part the development process.

A second requirement was that all items be scored by computer. This was a necessary feature in order to implement adaptive testing in PIAAC. Developers thus had to define response modes that could be computer scored across languages for each of the cognitive domains and for both linking and new items in the assessment. The PIAAC design called for the continued use of open-ended response items both to maintain the real-life focus of the assessment and to maintain the psychometric link between PIAAC and prior surveys. While those prior paper-and-pencil surveys allowed respondents to write responses ranging from a word or two to several sentences, the use of automated scoring for such responses was not possible for PIAAC given that the assessment was to be delivered in 33 languages.

The Consortium therefore relied on evidence from previous ETS work on a derivative computer-based test for individuals to define a set of computer-scoreable, open-ended response modes. This work had shown that item parameters for paper-and-pencil items were not impacted when those items were adapted to allow respondents to click on responses, type numeric answers, and highlight responses in text. Development therefore proceeded on the assumption that linking items could be adapted to employ these response modes and still maintain item parameters from previous assessments, an assumption that was ultimately supported by the Field Test data.

Additionally, each of these three response modes required only basic computer skills – an important consideration given that the test needed to be accessible to adults with a range of computer experience. The three are described in more detail below.

- **Clicking items**

These items required respondents to click on graphical elements, cells in a table, links on a Web page, or radio buttons or check boxes to answer. Respondents could select and change their answers while working within each unit. In terms of scoring, one or more correct responses were defined for each item. This response mode had an advantage in that, in general, click areas remained consistent across languages and therefore scoring did not require much adaptation across different national versions of the items.

- **Numeric entry items**

For these items, respondents answered by typing a numeric response using the number keys, decimal point (period or comma as appropriate across participating countries) and space key. In this response mode, all other keys on the keyboard were locked and not available for use to prevent respondents from including text in their responses that could not be scored.

Numeric entry items were scored automatically based on the definition of correct numeric response(s) included in the scoring rule. One scoring rule employed a number match. In this case, a response was correct as long as it represented the correct numerical value, regardless of how that number was represented. For example, if a correct response was 4,

responses such as 12/3 or 2*2 would receive a correct score. The second type of numeric scoring rule required an exact match. That is, instead of checking for numerical equivalence, the system checked for character equivalence. In this case, a response of 229 would be scored differently from responses such as 229.0 or 229.00. As described in more detail later in this section, guidelines were provided to allow countries to adapt numbers and number formats in order to present respondents with realistic numerical values in the context of presented tasks.

- **Highlighting items**

These items allowed respondents to highlight one or more words, phrases and sentences in a text to answer questions. Defining the scoring rubrics for these items was most challenging as responses were language dependent. For each response, developers defined a minimum correct response, as well as a maximum correct response. They based those judgments on ETS's previous work to develop open-ended, computer-scoreable items as well as experience in scoring paper-based responses. In previous paper-based assessments, respondents were given credit for correct answers when they underlined or circled information in the stimulus instead of writing an answer on a response line. Existing rules for what constituted a correct response in those situations thus helped guide the development of rules for highlighted responses in PIAAC. As this was the most language-dependent response mode, countries were actively involved in implementing and testing the minimum/maximum rules for their national versions of these item types.

In terms of the scope of item development for the cognitive instruments, the PIAAC assessment design specified the number of items to be developed for the Field Test and subsequently used in the Main Study. The Field Test and Main Study needs for literacy and numeracy, the two domains with linking items, are shown in Table 2.3. The Main Study design included 24 items for the paper-and-pencil version (19 linking items and five new items) and 48 items for the computer-based version (29 linking and 19 new) for each domain. To reach these goals for the Main Study, the Field Test design specified 35 paper-and-pencil items (25 linking and 10 new) and 72 computer-based items (42 linking and 30 new). Note that for both domains, the Main Study design additionally specified that a set of 18 linking items was to be used in both the paper-and-pencil and computer-based versions of the instruments.

Table 2.3: Literacy and numeracy item needs for PIAAC

	Field Test		Main Study	
	Linking	New	Linking	New
Paper Version	25	10	19	5
Computer Version	42	30	29	19

Reading components tasks were developed according the framework for this domain. These measures focused on speed and accuracy and several measures were assessed in a defined amount of time. A total of 20 minutes was allotted in the Main Study to measure these skills, with final measures assembled from 40 minutes worth of Field Test items.

The assessment of PSTRE involved scenarios of varying complexity and length, designed to take between five and 15 minutes to complete. Overall, 14 units were used in the Field Test. Several of those units included multiple parts, or tasks, so a total of 24 tasks were included. Two 25-minute blocks were included in the Main Study. Block 1 had five units, with seven associated tasks, and Block 2 had six units, also with seven tasks.

2.3.1.1 Selecting and adapting linking items

The assessment design for the PIAAC Main Study required that 60 percent of the literacy and numeracy items be taken from, and therefore link back to, previous surveys. In the case of literacy, items from both IALS and ALL were reviewed as potential linking items for PIAAC. As numeracy was not a domain in IALS, all numeracy linking items were selected from the ALL survey. The following aspects were taken into consideration when selecting linking items for inclusion in PIAAC.

- **Item quality**

To be eligible for inclusion in PIAAC, items needed strong statistics from previous assessments. That is, developers were looking for items with good item parameters and items with no history of differential item functioning or translation problems.

- **Distribution according to the dimensions of the frameworks**

Items were reviewed and reclassified according to the PIAAC frameworks and, to the extent possible, selected to reflect the distributions recommended by the expert groups.

- **Distribution across levels of difficulty**

The difficulty of items was taken into consideration in an effort to be sure items reflected the five levels used to report results for both previous studies and PIAAC.

- **Cultural appropriateness**

Countries were asked to review the selection of linking items to identify any of particular concern in terms of their appropriateness across the range of cultures among PIAAC participating countries.

An additional critical consideration for PIAAC was the suitability of these linking items for computer delivery. All of these items had been developed for paper-and-pencil assessments with open-ended responses that were human scored. For PIAAC, items needed to be computer scored, so selected items needed to be answerable using the response modes of clicking, numeric entry and highlighting. In addition, the stimulus materials for selected items needed to be adaptable to onscreen presentation keeping the same formatting as that used on paper.

The Literacy and Numeracy Expert Groups met in 2008 to review and provide input regarding the selection of linking items for the Field Test as well as to discuss issues associated with moving these items to the computer.

2.3.2 Developing new items

New items were developed to reflect the PIAAC frameworks and take advantage of the computer-based nature of the assessment. For example, new literacy items were designed to assess skills and knowledge associated with digital texts. Literacy and numeracy development also needed to complement the set of items selecting as linking items. As a new domain, PSTRE included only newly developed items. For all domains, the new item development process involved countries, the PIAAC Consortium, and expert groups.

2.3.2.1 National submissions

Countries were invited to participate in the process of developing new items for PIAAC. As is the case with any large-scale international survey, it was important that the pool of tasks for PIAAC reflected the range of contexts and experiences of respondents across participating countries. One way to better ensure this was to solicit national submissions once countries had been introduced to the PIAAC frameworks. The request for literacy and numeracy item submissions was issued in 2008 during the first meeting of the NPMs. The Consortium developed a document that provided: i) a general overview of the item development task, including a description of the scope of work, ii) a summary of the development process to be followed, iii) procedures for submission and review of items, iv) the item development timeline, and v) sample items that illustrated the kinds of items to be developed.

Due to the tight development schedule, countries had three months to develop and submit items. To facilitate country participation, the Consortium accepted item submissions in six languages including English, French, Spanish, German, Japanese and Italian. Additionally, to better integrate submissions into the development process, countries were encouraged to submit items progressively as they were developed, rather than as a single submission close to the deadline.

In preparing materials for submission, national item developers were asked to provide the following information about each item:

- information about the source of the item (original, or from a book or other source)
- information about any copyright considerations for the stimulus materials (who holds the copyright, who had been contacted to seek permission to use the material, and copyright permission when it was obtained). Countries were responsible for obtaining copyright information for any submitted material.
- the classification of each item according to categories in the relevant domain framework

Countries were also encouraged to submit additional stimulus materials without associated items. Wherever possible, the Consortium developed items based on these stimuli in order to ensure a mix of materials that reflected the diversity of cultural contexts represented across participating countries.

Literacy submissions were received from Austria, Estonia, France, Italy and Japan. In Numeracy, submissions were received from Austria, Estonia, Finland, France, Hungary, Italy, Japan and Korea. Submissions received from countries were reviewed and evaluated in terms of their fit to

the PIAAC frameworks and contribution to the item pool. This process was documented and summarized in a detailed report that was shared with countries.

Because PSTRE was a new domain with complex development demands, the Consortium did not expect countries to submit fully drafted tasks. Instead, it asked countries were asked to submit ideas for tasks which illustrated common adult uses of technology in problem-solving contexts or where the appropriate use of technological functions (such as a “compare” function on a shopping site or “sort” function in a spreadsheet) facilitated solving a problem. Additionally, countries were encouraged to provide examples of Web sites and other technology environments that they viewed as representative of materials used by adults in their home, community and work environments.

2.3.2.2 Item development

New assessment materials for the Field Test were developed based both on materials submitted by countries and materials developed by the contractors. The development period extended from early 2008, with the first meeting of the expert groups, to early 2009, when the expert groups finalized the selection of the Field Test item pool.

As previously mentioned, in the case of literacy, the PIAAC contractors, including item writers in Australia and the United States, developed the new items. The process differed for numeracy, where the expert group itself drafted all the new items. To accommodate this work, several additional expert group meetings were held. In August 2008, the numeracy expert group met in Dublin, Ireland, and developed approximately 60 items. In November 2008, the group met again in Frankfurt, Germany, to review countries’ comments on the first batch of materials, consider how best to implement the suggested changes, and review other available items. As a new domain, PSTRE also had a higher level of involvement from experts with two additional meetings. The PSTRE Expert Group met in August 2008 in Poitiers, France, to review an initial set of draft tasks. A second meeting, held in Amsterdam in December 2008, included programmers as well as item developers so that features of the simulated technology environments could be discussed and agreed to along with content for specific tasks.

For each domain, stimulus materials were selected based on specifications provided in the framework for that domain. To the extent possible, stimuli for the PIAAC assessment were taken from real-world materials such as newspaper and magazine articles, advertisements, books, forms, and Web pages that adults ages 16-64 would encounter in a range of everyday life contexts. Given the international context of the assessment, care was taken to select materials appropriate across cultures and languages. Soliciting materials from participating countries and having all countries review the stimulus materials were important steps to better ensure this diversity.

It was also important to ensure that stimulus materials would not become too easily dated. Those that contained dates or references to contemporary individuals or events – particularly if such information was central to completing tasks associated with those materials – could become dated by the time the assessment was administered. Such materials were also avoided as they would become increasingly problematic in future testing cycles if they were needed as linking items.

Tasks for PSTRE were situated in simulated computer environments including a browser, email system, spreadsheet and word processor. While these did not replicate the full functionality of real-life environments, they included many key functions. For example, the email environment allowed respondents to reply, reply to all, forward, send and move emails to folders. In the browser environment, respondents could navigate using the back, forward and home buttons and they could bookmark pages for later reference. Presenting the PSTRE tasks in these simulation environments allowed the computer to capture a variety of process information. For any given task, collected information included time spent, actions taken (e.g., clicking and typing responses or selections from drop-down menus such as “file” and “edit”) and the sequence in which actions were completed. This information provided direct evidence of the processes and strategies respondents used to complete assigned tasks and therefore allowed for better inferences about their knowledge and skills related to PSTRE.

2.3.2.3 Item reviews

As an additional step to better ensure that the new items reflected the range of contexts and experiences of respondents across participating countries and to obtain input about item content, all participating countries reviewed the PIAAC item pool at several stages. Guidelines were developed for the review process which specified that the materials were to be reviewed in relation to:

- coding based on the task characteristic categories in the frameworks
- the overall appropriateness of each item. Items were to be classified into one of three categories: acceptable as is, acceptable with modifications, or unacceptable. For the second category, countries were asked to specify revisions that would make the item acceptable. They were also asked to specify the reason or reasons why they rated any items as unacceptable.
- cultural concerns
- translation concerns

Countries were given an opportunity to review draft items before developers finalized them with input from the expert groups. Reviews were conducted in three batches as described below:

- A first batch of new tasks was released on 21 October 2008 with comments due on 7 November. This batch included: i) four item sets for reading components; ii) 16 literacy units with 105 tasks; iii) 20 numeracy units with 48 tasks; and iv) 11 PSTRE scenarios.
- A second batch of new tasks was sent to countries for review on 17 December 2008 with comments due on 20 January 2009. This included a set of 31 new numeracy tasks and six new literacy tasks.
- A third batch was released on 15 January 2009 with comments due on 29 January 2009. This last batch included seven tasks for PSTRE.

2.3.3 Additional supporting materials

The development process for PIAAC cognitive instruments included several sets of materials beyond the items themselves. These included a set of detailed guidelines to assist countries in translating items and scoring guides so that national instruments would remain comparable with the international masters. Equally important was a set of interactive tutorials that introduced respondents to the PIAAC instruments, ensuring that all participants approached the survey with the same information about how to navigate through the assessment and provide their responses.

2.3.3.1 Translation/adaptation guidelines and scoring guides

To support the work of countries in translating and adapting items, implementing computer-based scoring, and translating scoring guides for the paper-based items, the Consortium developed translation and adaptation guidelines as well as master scoring guides for participating countries. These materials also supported the linguistic quality control process, described in Chapter 4, that was designed to help ensure that instruments across countries were comparable and that consistent scoring procedures were implemented.

A sample set of guidelines for one of the Field Test items is shown in Figure 2.2. The guidelines specified linguistic considerations for translation (e.g., maintaining a literal match between wording in the question and stimulus) and defined the correct response for both the paper-and-pencil and computer versions (minimum and maximum) of the item.














Figure 2.2: Sample translation and scoring guidelines

Item Notes: Translation must maintain literal match between keywords “gym bench” and in question and in table heading under “Muscle building.”

“Muscle” appears in question and four places in the stimulus.

		English Paper and Pencil (same version as ALL)	English Computer
Directions		Use the exercise equipment chart on the opposite page to answer questions x through y.	Look at the exercise equipment chart. Click on the chart to answer the question below.
Question		Which muscles will benefit most if you use the gym bench?	Which muscles will benefit most if you use the gym bench?
Answer	Abdominal (muscles)	<u>Minimum correct response:</u> Clicks on “abdominal muscles” cell <u>Maximum correct response (See illustration below):</u> Muscle building Gym bench Image of gym bench Very good (intersection of abdominal muscles row and gym bench column) Abdominal muscles	

Maximum correct response

Effects on...	Cardio-Training					Muscle Building							
	Exercise bicycle	Rowing machine	Stepper	Tread-mill	Air trainer	Dumb-bells, weights	Elastic	Gym bench	Muscle-building bench	Multi-trainer	AB trimmer	AB shaper	AB roller
													
Arm strength	Ineff-ective	Good	Average	Ineff-ective	Good	Very good	Very good	Good	Good	Good	Very good	Good	Good
Leg strength	Good	Very good	Average	Very good	Good	Ineff-ective	Good	Average	Good	Good	Ineff-ective	Good	Good
Abdominal muscles	Average	Very good	Good	Good	Average	Ineff-ective	Good	Very good	Good	Average	Very good	Very good	Very good
Overall muscle building	Ineff-ective	Very good	Ineff-ective	Average	Ineff-ective	Average	Good	Good	Good	Average	Good	Good	Good
Heart/arteries	Very good	Good	Very good	Very good	Good	Ineff-ective	Average	Average	Average	Good	Average	Average	Average
Flexibility	Ineff-ective	Good	Ineff-ective	Ineff-ective	Average	Average	Average	Good	Ineff-ective	Ineff-ective	Average	Good	Good
Joints	Good	Very good	Good	Good	Good	Good	Average	Average	Good	Good	Average	Average	Average
Slimming	Good	Average	Very good	Good	Good	Ineff-ective	Average	Good	Average	Average	Good	Good	Good
Dangers	None	Back	None	Legs		It is best to learn to use these types of apparatus properly before you make a major effort							

For numeracy, specific guidelines were provided to guide countries in the adaptation of numbers and number formats. For example, two options were provided to address the challenge of consumer-related items that involved currency. The first was for countries to keep the numbers the same but change the currency sign. This was the option of choice for adapting U.S. dollars to euros as the two are close in value. For currencies where simply changing the currency sign would result in unrealistic numbers, a second option was provided. Guidelines specified that in this case, numerical values could be changed by multiplying or dividing them by powers of 10 (and only powers of 10). This restriction was intended to allow countries some flexibility while maintaining similar cognitive demands across national versions.

2.3.3.2 Tutorials

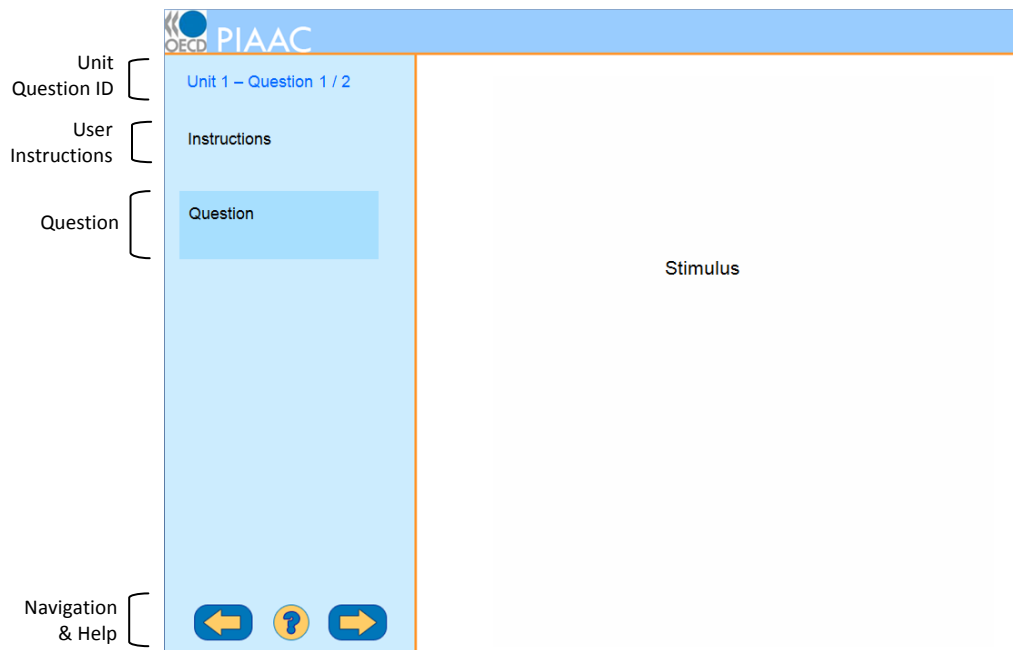
As part of the development process, developers designed a set of tutorials to introduce respondents to the design and layout of the testing screen, familiarize them with the available response modes in each domain, demonstrate the navigation and help functions, and, in the case of PSTRE, define the tools and functionality in the simulated environments. For the Main Study, these tutorials were designed to be relatively brief, about five minutes per domain, in order to reduce respondent burden in terms of the time required to complete the full assessment.

2.3.3.3 PIAAC testing screen

While not a material per se, an additional step in the development of the cognitive instruments was the design of the testing screen for the cognitive items. An important goal was to develop a design which allowed a consistent display and interface across the cognitive domains. PIAAC used a vertically split screen as shown in Figure 2.3. For all domains, the stimulus material was presented on the right and the task information on the left.

Navigation and help icons were located at the lower left. Labels for these icons displayed when the respondent held the cursor over them, allowing translations of various length to display. The user instructions both identified the stimulus and provided information about the required response mode. For example, “Look at the exercise equipment chart. Click on the chart to answer the question below.”

Figure 2.3: PIAAC testing screen



The design presented a number of important advantages.

- The vertical split facilitated the display of paper-and-pencil linking items being moved to the computer. Splitting the screen vertically allowed a display area for stimulus materials that was taller than it was wide. Because this more closely mirrors the width-to-height ratio of paper, this was an advantage for displaying paper-based linking items.
- By not extending the full width of the screen, stimulus text could be formatted with more natural line lengths, improving readability.
- More vertical height accommodated displays across a variety of languages.

2.3.4 Preparation of final Field Test instruments

2.3.4.1 International master

The Consortium finalized and released the master versions of the Field Test items to countries for translation and adaptation according to the timeline shown below. Each round of released items included the items themselves, translation and adaptation guidelines for the items, Verification Follow-up Forms (used for monitoring and documenting the translation/adaptation process), and scoring guides.

- Linking items for numeracy and literacy in computer-based format were released for translation in two rounds: 16 January 2009 and 5 February 2009.
- New numeracy items were released on 6 April 2009. The scoring guides for the paper-based numeracy items were released on 28 April 2009.

- Reading components items were released on 6 April 2009.
- New literacy tasks were released on 9 April 2009. The scoring guides for the paper-based literacy items were released on 28 April 2009.
- PSTRE scenario were released in batches with five scenarios were released on 29 April 2009, five scenarios on 3 May 2009, and four on 30 May 2009.

Master versions of the Field Test paper booklets were also released to countries in the spring of 2009. The assessment design for the Field Test required four sets of paper booklets including: two literacy booklets, two numeracy booklets, and four reading components booklets. The assembly of all paper-based booklets, including instructions for administration and scoring sheets, occurred during this period.

Finally, once the master versions of the computer-based units were tested and finalized, these were assembled into computer blocks following the Field Test design that required four numeracy and four literacy blocks, each with 18 tasks. These blocks were organized in a way that ensured a balanced distribution across important aspects of the frameworks and known or estimated difficulty levels and assembled by the Consortium.

2.3.4.2 National versions

Countries developed their own national versions of the Field Test assessment materials following the translation, adaptation and verification processes developed for PIAAC between April and June 2009. Layout checks were conducted by both the Consortium and countries to identify any display issues requiring modification. Such revisions were prompted by issues including text that did not fit within a table cell due to longer word lengths in some languages, and so on. The Consortium manually fixed layout issues on a case-by-case basis and submitted them 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 highlighting areas – were defined by the national centers and verified as part of the quality assurance process.

2.3.4.3 Scoring testing

The Consortium tested the automatic scoring for the international version of the literacy and numeracy units prior to distributing the national versions. Two sources of error were observed during international testing: i) errors at the level of item editing, that is, the scoring information was specified incorrectly by the item editor (specification error), and, ii) errors at the level of technology, that is, the software did not work accurately (implementation error). All detected errors were fixed, and the scoring procedures of affected units were retested until no further errors were found.

Countries were responsible for testing their national versions based on scripts provided by the Consortium. Scoring testing at the national level was especially important when the correct response included translated and/or adapted textual and numerical information. The testing was done manually, that is, the tester completed each item multiple times, responding to items

correctly and incorrectly as specified in the script. That script included the expected scoring result for each response so the tester could compare the observed and expected scoring result. Discrepancies were documented and reported to the Consortium for debugging, with testing iterations continuing until all problems were corrected.

2.3.5 Moving from the Field Test to Main Study instruments

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

- **Item analysis**

Items were evaluated based on their statistical performance in the Field Test, looking at performance within and across countries as well as across modes (i.e., computer and paper). The purposes of the Field Test analyses were to ensure that items were reliable, valid and comparable across countries and that common scales could be developed across countries and assessments.

- **Item selection**

Based on the Field Test data, developers recommended a draft set of Main Study items for each domain in December 2010. These items were reviewed by the expert groups who, in partnership with developers, finalized the set of items. The recommended set was then presented at a meeting of the NPMs as well as the BPC for their approval.

One challenge for the Main Study selection process was the need to fit the final set of items within the testlets that made up the adaptive design. As shown in Table 1.2 in Chapter 1, the design for the computer-based adaptive instrument included two stages, divided into a total of seven testlets. To accommodate this design, developers needed to look at the difficulty level of items available for the Main Study and determine the appropriate testlets and blocks for the items. For literacy, the fact that items existed as units, or 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.

- **Item corrections**

Countries reviewed the set of items selected for the Main Study, looking for any errors in translation or implementation identified by the Field Test data or during the final national check of those items. Errors were corrected and the final version reviewed and approved for implementation for the Main Study.

The set of items for the Main Study was balanced in terms of construct representation, based on the overall distributions recommendations in the framework. A total of 58 items was selected for literacy and numeracy, with the distribution across linking and new paper and computer versions shown in Table 2.4 below.

Table 2.4: Literacy and numeracy items in the PIAAC Main Study

	Literacy		Numeracy	
	Linking	New	Linking	New
Paper-based	18	6	19	6
Computer-based	30 (including computer versions of the 18 above linking items)	22	28 (including computer versions of 14 of the above linking items)	22 (including computer versions of 3 of the above linking items)

The distribution of these items based on the task characteristics defined in each domain framework is detailed below.

2.3.5.1 Literacy

The distribution of the literacy items included in the Main Study by task characteristics is presented in Tables 2.5-2.7 below.

Table 2.5: Distribution of literacy items by medium

	Final item set		Framework goal
	Number	%	%
Print-based texts	36	62	70-80
Digital texts	22	38	20-35
Total	58	100	100

Note: Each category includes continuous, noncontinuous and combined texts.

Table 2.6: Distribution of literacy items by context

	Final item set		Framework goal
	Number	%	%
Work	10	17	15
Personal	29	50	40
Community	13	23	30
Education	6	10	15
Total	58	100	100

Table 2.7: Distribution of literacy items by task aspects

	Final item set		Framework goal
	Number	%	%
Access and identify	32	55	40
Integrate and interpret	17	29	45
Evaluate and reflect	9	16	15
Total	58	100	100

2.3.5.2 Numeracy

The distribution of the numeracy items included in the PIAAC survey by task characteristics is presented in Tables 2.8-2.10 below.

Table 2.8: Distribution of numeracy items by response (process)

	Final item set		Framework goal
	Number	%	Number
Act upon, use	34	61	50
Identify, locate or access	3	5	10
Interpret, evaluate	19	34	40
Total	56	100	100

Note: Each category includes continuous, noncontinuous and combined texts.

Table 2.9: Distribution of numeracy items by context

	Final item set	
	Number	%
Everyday life	25	45
Work-related	13	23
Society and community	14	25
Further learning	4	7
Total	56	100

Table 2.10: Distribution of numeracy items by mathematical content

	Final item set		Framework goal
	Number	%	%
Data and chance	12	21	25
Dimension and shape	16	29	25
Pattern, relationships and change	15	27	20
Quantity and change	13	23	30
Total	56	100	100

2.3.5.3 Problem solving in technology-rich environments

Fourteen PSTRE tasks were included in the Main Study. These included both short and long scenarios.

The distribution of the PSTRE assessment items included in the Main Study by task characteristics is presented in Tables 2.11-2.13 below.

Table 2.11: Distribution of PSTRE tasks by cognitive dimensions

	Number*
Setting goals and monitoring progress	4
Planning	7
Acquiring and evaluating information	8
Using information	6

*Some tasks address more than one cognitive dimension so total is more than 14

Table 2.12: Distribution of PSTRE tasks by technology dimension

	Number*
Web	7
Spreadsheet	4
Email	9

*Some tasks involve more than one technology environment so total is more than 14

Table 2.13: Distribution of PSTRE tasks by context

	Number
Personal	8
Work/Occupation	4
Civic	2

2.4 Conclusion

The decision to deliver PIAAC as a computer-based assessment presented both opportunities and challenges for the development of the cognitive instruments. Computer delivery allowed the inclusion of technology-based texts and environments, reflecting the range of materials that many adults encounter in their everyday lives. It also allowed adaptive testing, more reliable computer-based scoring, and the ability to collect a broader range of performance data including timing and process information. One significant challenge was that, in keeping with the open-ended response format used in IALS and ALL, developers needed to define response modes that could allow a reasonable range of open-ended responses while still being computer scored.

The three expert groups considered the implications of computer delivery in their frameworks for literacy, numeracy and PSTRE. Those frameworks defined the general outlines of the assessment instrument in each domain, specifying the task characteristics to be manipulated by test developers and outlining the relative proportion of items to be developed based on the key variables associated with those task characteristics.

Instrument development for the literacy and numeracy domains included selecting linking items from previous large-scale assessments and developing new items. The selection process for linking items involved considering how response modes for items could be adapted to open-ended, computer-scored formats as well as evaluating display and formatting issues for stimulus materials. New items for literacy, numeracy and PSTRE were developed with input from participating countries that included item submissions and a detailed review process. Additionally, developers worked closely with the expert groups who reviewed and, particularly in the case of numeracy, developed items for inclusion in PIAAC. This collaborative endeavor, with input from individuals with a range of expertise and perspectives, resulted in a set of innovative cognitive instruments that provided important information about the skills and knowledge of adults across participating countries.

References

- Lennon, M., Kirsch, I., Von Davier, M., Wagner, M., & Yamamoto, K. (2003). *Feasibility study for the PISA ICT literacy assessment*. Princeton, NJ: Educational Testing Service.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, D.C.: American Council on Education.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(1), 13–23.

Chapter 3: The Development of the PIAAC Background Questionnaires

Jim Allen and Rolf van der Velden, ROA; Susanne Helmschrott, Silke Martin, Natascha Massing, Beatrice Rammstedt and Anouk Zabal, GESIS; and Matthias von Davier, ETS

3.1 Introduction

This chapter documents the work done by the Consortium to develop, test and refine the PIAAC BQ. It starts in Section 3.2 by describing the conceptual framework that provides the underpinning for the BQ, outlining the main policy questions that the PIAAC project seeks to answer, and providing the theoretical underpinnings of the concepts that are needed to answer these policy questions and, hence, represented in the BQ. Section 3.3 briefly explains the rationale underpinning the JRA module in the BQ, which was developed separately from the main master BQ by a different team of experts. Section 3.4 deals with the development and validation of the BQ, including an outline of the decision-making process for selection of items in the BQ, a brief summary of two rounds of cognitive testing that were conducted, and a report on the analysis that was conducted of the data from the Field Test with a view to refining and shortening the BQ for the Main Study. In Section 3.5, a brief outline will be given of the content of the BQ, including an overview of the structure, and a brief description of the national adaptations and extensions that were made. Finally, in Section 3.6 we explain how the BQ was implemented on the TAO platform, in particular in terms of the use of instructions to interviewers, help buttons and consistency checks that allowed the BQ to be administered in a coherent and standardized way across the participating countries.

3.2 The PIAAC conceptual framework for the BQ

The policy questions

The PIAAC project seeks to answer the following policy questions:

- A. How are skills distributed?
- B. Why are skills important?
- C. What factors are related to skill acquisition and decline?

How are skills distributed?

Human capital is considered the driving force of economic growth. Investments in skills are vital to keep up with technological change (the so-called Skill-Biased Technological Change) as well

as other changes resulting from market or organizational developments (e.g., the introduction of High Performance Workplace Practices). Policymakers have an interest in monitoring the stock of human capital in their country and identifying the different levels among relevant subgroups. PIAAC assesses the stock of human capital in a society by providing a descriptive analysis of the distribution of skills proficiencies and skills use in the adult population. The survey enables countries to answer questions such as:

- How does the adult population in a country compare to that of other countries in terms of average levels of skills proficiency and skill use? What share of the adult population has low proficiencies of relevant skills?
- In terms of equity, how are skills distributed among relevant subgroups, such as gender, age group, region or migration status? Are certain subgroups particularly vulnerable to low skills proficiencies?
- How are skills proficiencies distributed across sectors of industry? Are there certain sectors of industry that are characterized by particularly low levels of skills proficiency? How do the skill levels of these sectors compare to those in other countries?
- How are skills proficiencies distributed across different levels of schooling when benchmarked against other countries? Are there population subgroups that appear to be underserved by the current education system? Is there an underdevelopment of skills at particular levels of education? What are the skill levels of early leavers from education?
- Who is participating in adult learning of various types? To what extent are particular population subgroups excluded from adult learning systems?

Why are skills important?

There is little interest from a policy point of view for any investment in skills if it has no relation to relevant outcomes. Other services are competing with education and training for a share of budgets, so the case for returns to educational investment needs to be made on a secure and sophisticated evidence base. Moreover governments and the public make education accountable to show the effects of their efforts. For that reason, one of the key goals of the BQ is to provide indicators that can be used to show if differences in skill matter economically and socially. The most obvious area in which policymakers are interested is how skill levels are related to economic outcomes of individuals. Cognitive skills are thought to be a key determinant of an individual's productivity, and therefore it is not surprising that cognitive skills are related to economic success. There is a large body of evidence showing that higher cognitive skills are indeed associated with better labor market outcomes. Relevant questions are:

- How are skills related to individual employment opportunities and job security?
- How are skills related to earnings and other indicators of labor market success?
- Do low skill proficiencies form a barrier to individuals entering the labor force?

- Are low-skilled people more affected by job insecurity? Is there a minimum level of skills needed to be employable?
- How do skills affect the relation between education and training on the one hand and economic outcomes on the other? Can skills compensate for low educational qualifications?

Apart from economic outcomes, other areas are of interest as well, such as the relation among skills, health status and civic participation. Adverse outcomes in such areas place large burdens on governments, businesses, and individuals, including both the direct expenditure of resources (such as government spending on health care) and indirect costs (such as the value of goods and services workers do not produce while ill).

Relevant questions are:

- To what extent is literacy related to health status of individuals, various subgroups, and the overall population?
- To what extent do individuals with low skills appear to be less engaged in the broader society (voluntary work, social trust)?
- How do individuals with low skill levels cope with their everyday reading and numeracy demands? To what extent do these coping mechanisms make these individuals reliant on others? To what extent does the engagement of migration groups or linguistic minorities appear to be inhibited by their lack of skill in the language of the test?
- Do high-skilled people have a higher involvement in civic activities? What is the relation between skills and the level of social trust?

What factors are related to skill acquisition and decline?

Under the assumption that skills matter economically and socially, policymakers have an interest in knowing what factors are related to higher skill levels. Of course, the prime focus is to assess the effects of factors directly affected by policy, such as the provision of formal and non-formal organized learning activities like education and training. But it is also relevant to compare the efficiency of these skill production routes with the efficiency of others not directly under the control of policymakers, such as the informal learning activities in which people can engage. Assessing the overall relation among education and training and skill levels is only a first step in unraveling the determinants of skills acquisition. We can assume that not all education and training activities have the same impact on skills development. Nor can we assume that the impact is the same for all relevant subgroups. Policymakers have an interest in seeing which characteristics of education and training are most strongly related to higher skill levels in the population and which subgroups appear to profit most from which type of intervention. Finally, we need to be aware that skills can be acquired, but also can be lost. Preventing skill decline is probably just as important as promoting skill acquisition, but the underlying factors affecting these processes may be quite different, and it is important to have good insight in both processes.

For these reasons, the survey was designed to enable countries to answer questions like:

- What is the relation between education and training and the skill development of people? Are these relations different from those with other learning activities that people engage in to develop their skills, such as informal on-the-job learning?
- Are the effects of education and training the same for each subgroup? Are there subgroups that appear to profit from the investment in education and training?
- What is the relation between underinvestment in work-related training and adult skill levels? How are characteristics of the work environment related to skill levels? Is informal learning an on-the-job a substitute for work-related training?
- How do processes of skill acquisition and decline vary with age? What are the factors related to skill decline? Are these the same factors as are related to skill acquisition?

Theoretical background

In this section we describe the main theoretical elements of the conceptual framework and, where relevant, indicate the items that have been included in the BQ to reflect these elements. The purpose of this part was to provide a solid theoretical basis for the policy questions formulated in the previous section. It also served as a guideline for the selection of relevant concepts and the translation of those concepts into specific questions in the BQ. This framework also served as a guideline for the analysis and interpretation of the data in the Field Test, where it was used to derive predictions on how particular sets of variables were expected to behave. Its main function in the Main Study is as a basis for deriving hypotheses pertaining to the policy questions outlined in the previous section.

The presentation of the theoretical framework will be divided into three parts, roughly corresponding to the three types of policy questions described above. We start with a brief overview of the literature on the nature of key skills. Although the direct assessment (DA) as such falls outside the scope of the development of the BQ, the *raison d'être* of the BQ is to provide the context information needed for analyzing and interpreting the results of the DA. As a consequence, it is essential to proceed with a solid understanding of what is being measured in the DA and, equally important, what is not being measured. We then summarize the literature pertaining skills acquisition and decline. The theoretical discussion is concluded with a review of the literature on outcomes of skills.

What are key skills?

As noted above, policymakers have a strong interest in knowing how skills are distributed across countries as well as across different subgroups within countries, such as age, gender, ethnicity, regions, sector of industry, and levels and fields of education. If we want to answer these questions, it is important to first take a step back and reflect on what is being compared. Below is a brief overview of the literature on so-called key skills, of which the skills measured in PIAAC form an important subset.

The quest for key skills

The last few decades have seen an increased awareness of human capital as one of the driving forces of economic development. Policymakers have realized the importance of investing in education and training as a way of improving the existing stock of skills. This has resulted in an accompanying need to monitor and assess the stock of human capital. What soon became clear is that education as such is a poor indicator of the stock of human capital. Individuals with the same nominal level and type of education can differ markedly in their command of various skills. Likewise countries that have more or less comparable levels of educational attainment can nevertheless differ substantially in the level of skills that are acquired in education. This has been shown in studies like ALL and the Programme for International Student Assessment (PISA).

As the emphasis shifts from educational qualifications towards skill measurement, the question naturally arises as to what skills should be measured. It seems clear that in order to perform even the most basic tasks, many discrete skills are required. Determining which skills should be measured is a complex and difficult task, which is compounded by the fact that people not only make use of generic skills such as the ability to communicate or the ability to learn, but also of a large number of highly specific skills pertaining to particular tasks, situations and objects.

In order to introduce some order in the understanding of the diversity of human skills, many scholars have engaged in a quest for so-called core skills or key competencies. A major project in this respect was the DeSeCo (Definition and Selection of Competencies) project. This project was initiated by the OECD to provide an overarching framework for international skills assessments. Competencies are defined in this project as “the ability to successfully meet complex demands in a particular context through the mobilization of psychosocial prerequisites (including both cognitive and noncognitive aspects)” (Rychen & Salganik, 2003, p. 43). The basic difference between this view and earlier concepts of skills is the holistic nature of the concept of competence. It refers not only to a range of cognitive and noncognitive skills and other prerequisites that need to be in place in order to perform in a competent way, but also to the notion of “orchestration,” which is defined as the ability to use these constituent elements in a meaningful and deliberately arranged way.

Although the theoretical framework provided by the DeSeCo project injects some welcomed theoretical rigor into the discussion of skills measurement, it does not in itself directly give rise to clear recommendations as to the competencies to be measured. The best way to conceive of this overarching framework is to see that it indicates the main underlying competencies that give skills their significance.

Binkley et al. (2003) developed a framework that provides more detailed guidance for the development of skills measurements. This work concentrated on two strands of research: what skills are necessary in the workplace, and cognitive functioning. From the first strand, a list of six skill areas was extracted that seemed to underlie many of the most important skills: communication (speaking, listening, reading and writing), mathematical, problem solving, intrapersonal (motivation, metacognition), interpersonal (teamwork, leadership) and technology. From the strand of psychological theory, four core domains of intelligence were extracted: practical abilities, crystallized analytical abilities, fluid analytical abilities, and creative abilities (the ability to cope with novelty). As the authors point out, the two strands are not mutually exclusive, but rather represent different aspects of skill. The workplace skills provide the context

within which each of the four core intelligence domains are expressed; conversely, each category of workplace skill can involve four distinct types of thinking.

The choice of direct assessments in ALL was based not only on these theoretical notions but on practical considerations such as an established tradition of measurement where assessments are sufficiently compact to be used in a household survey. As a consequence, ALL concentrated on only part of the matrix formed by the intersection of the two strands of research, in particular the more generic aspects of the communication and mathematical skill areas. PIAAC builds on the direct assessments in ALL, extending these to the area of problem solving in technology-rich environments, which contains elements of the problem solving and technology skill areas. Although it is not possible to draw any sharp dividing line, the three domains of direct assessments in PIAAC differ in the extent to which they relate to the four types of thinking derived from psychological theory. Because the developmental pattern throughout life is thought to be quite different for the different types of thinking, this has important implications for the manner in which the different skills can typically be expected to be acquired and in some cases eventually lost. We will return to this point below.

To the extent that the skills measured in the direct assessments are shown to be related to important economic and social outcomes (see below), the pragmatic restriction to those skill aspects that lend themselves well to a survey approach need not seriously diminish the value of the information gathered. It is important, however, to keep in mind that we are dealing with a subset of the skills possessed by the individuals participating in the survey. The intrapersonal and interpersonal skill areas are not included in the direct assessment, but as will be outlined below, these are covered to some extent by items included in the BQ. Arguably the most conspicuous omission is in the area of specific skills used by individuals in their chosen line of work.

The importance of professional expertise

Even though employers often list generic cognitive skills and personal traits skills as the most important ones required in the workplace, professional expertise is a condition sine qua non for success in many occupations. For example, nobody would doubt that in order to become a good medical doctor, architect or car mechanic, one needs to acquire the domain-specific knowledge and skills that make up the professional domains of these occupations. The German psychologist Weinert formulated this as follows: “Over the last decades, the cognitive sciences have convincingly demonstrated that context-specific skills and knowledge play a crucial role in solving difficult tasks. But generally, key competencies cannot adequately compensate for a lack of content-specific competencies” (Weinert, 2001, p. 53).

There is, however, a plethora of specific professional skills. It is not possible to measure professional expertise directly in the PIAAC assessment, simply because there is no common assessment instrument that allows all different types of professional skills to be measured in a meaningful way for large populations. The absence of direct measures of specific skills underscores the importance of obtaining information on the occupation of working respondents, based on the answers to questions D1a and D1b in the BQ. As the differences among occupations in the skills measured in the direct assessments is likely to be at least matched and probably eclipsed by differences in level and type of specific skills, the residual occupation-level variance in economic outcomes should provide a rough indication of the economic importance of specific skills relative to the generic skills measured.

Although no direct assessment of occupation-specific skills is included in the PIAAC survey, measures of skill use in some more generic work-related areas, as well as in the domains covered by the direct assessment and in the area of interpersonal skills, have been developed in a separate module based on the JRA. This module is described in Section 3.3.

Current investments in education and training

From a descriptive point of view, it is important that PIAAC provides accurate information on current levels of education and training. Access to lifelong learning by different groups remains a crucial issue for governments of the OECD member countries. Formal education (B_Q01-B_Q10), formal training (B_Q12-B_Q20), and informal training (D_Q13a-c) all contribute to the stock of human capital, and countries will display different profiles in how the human capital stock is built up. PIAAC will provide a snapshot of human capital investments by the incidence and intensity of training during the previous 12-month period. From a policy viewpoint it is important to not only obtain an indication of the volume of investments, but in the case of adult education and training, to have information on financing of such investments. A large part of adult education and training efforts are paid for by employers. Since most training received by individuals also benefits other employers (externalities of training) this typically leads to too little work-related training being provided because part of the returns are captured by outside parties (competing organizations and the individual). From a policy perspective, this could warrant some interventions in the training market to balance out a potential source of underinvestment in training. In addition, knowledge on current investments in learning can contribute to the formation of policies designed to provide more equitable or effective inducements to encourage participation among those most in need of further learning. This refers both to differences across different skill levels (Are low-skilled individuals investing enough in their human capital?) and across key reporting categories as specified below. The questionnaire contains indicators of whether the training was followed in working hours (B_Q15b, to assess the level of investment by employers in training in terms of opportunity costs), whether the respondent's employer contributed to the costs of training (B_Q16, to assess the level of direct investment in training by employees, employers and other actors), and (reasons for) nonparticipation in learning activities in which the respondent would have preferred to engage (B_Q26a-b).

When analyzing training, it is necessary to be able to distinguish different categories of training. At the most general level, it is important to distinguish work-related from non-work-related training (B_Q14a). Work-related training is usually expected to have some effect on performance, which is presumably expected to be based on increased skill levels, and to result in productivity and possibly wage gains. Training that has been undertaken for other reasons may also increase certain skills but would not necessarily lead to productivity increases at work.

Reporting categories

For reasons of effectively addressing skill deficiencies, but also from the point of view of social equity, it is important to have a good picture of where the deficiencies are most concentrated. Are there population subgroups that appear to be underskilled? To answer these questions, we need to know how skills are distributed among relevant subgroups, as defined, for example, by gender (A_N01a), age (A_Q01a), socioeconomic background (J_Q06b, J_Q07b) or migration status (J_Q04, J_Q06a, J_Q07a). These so-called reporting categories are important both from a

point of view of equity and efficiency: If skill gaps lead to social and/or economic exclusion, this is not only detrimental to the well-being of the groups involved, but also to the functioning of the economy and society. Because the reasons for skill gaps are likely to be systematically different for different “at risk” groups, the policy measures undertaken are likely to be group-specific. Age is also important because both skills acquisition and skills decline are related to age, leading to typical age profiles of skills and skill-related outcomes.

Region (collected through the Case Management system) is an important reporting category as well because of strong regional differences in level of economic development in some countries. It may be that certain regions are being held back by particularly low levels of skills proficiency, or conversely, that regions can be identified where skill demand is particularly low. In addition, because policy is often formulated and/or implemented at the regional level, it is crucial to have access to outcomes at that level.

Occupation (D_Q01, E_Q01), sector of industry (D_Q02, E_Q02) and firm size (D_Q06, E_Q06) are needed to detect areas in which skill gaps exist and to assess the extent to which training investments are taking place to reduce these gaps. This and similar information form the basis for directing possible policy interventions to those groups where intervention is most needed.

Because highest level of education (B_Q01) is assumed to be one of the strongest predictors of skills (see below), and because this is differentially distributed across countries, a breakdown by this variable is needed for even the most elementary understanding of the results. In addition it is important to know how access to the education system is distributed across different subgroups that are “at risk” from the point of view of skills proficiencies.

Determinants of skills acquisition and decline

As was the case for defining and measuring skills themselves, there is not just one but several strands of research pertaining to how individuals acquire and in some cases lose skills over their lifetime. One prominent strand is that of the economics of education. Since the pioneering work by scholars such as Becker (1964) and Schultz (1963), economists have looked at education, training and other activities undertaken by individuals to improve their level of knowledge and skills as investments in human capital that are expected to yield returns in the labor market. A second major strand is that of sociological research that points to the social environment affecting school choice and educational attainment. The third strand is educational research, in which scholars have tried to uncover those features of education that are particularly effective in promoting learning. Fourth, a conceptually related but empirically largely distinct area concentrates on how people continue to learn after leaving initial education. An important focus of this strand of research is on courses, workshops and other forms of training in which employees participate, but in recent years the focus has increasingly broadened to include features of the job or organization that promote informal learning. Finally, this focus on lifelong learning has led to increased attention to the fact that individuals not only acquire skills over their lifetime but are also confronted with skill loss and a general decline in the ability to acquire and retain new knowledge and skills. In this section we will look at each of these strands of research in turn.

Education as an investment

In economics, education and learning are treated as an investment. From this point of view, people are expected to invest in education and learning when the costs are smaller than the future benefits. Not everybody is equally likely to invest in the same amount of education. People differ in the degree in which they enjoy education or learning and in the degree to which they value the potential benefits of education. Due to heterogeneity in preferences, there will also be heterogeneity in the decision to learn. Borghans et al. (2007) provide a model for investments in education and learning that capture a wide range of potential differences between individuals.

First, people differ in their capacity to acquire skills. The costs of education are lower for people who acquire skills more easily because they learn faster. The capacity to learn depends not only on innate cognitive abilities but also on personal traits. For example, someone who is easily distracted from a task will need more time to learn. Second, people differ in preferences. They might differ in how they value learning, working and leisure. They might differ in how much they value a high income or other potential benefits of education, and they might differ in how they value future benefits compared to current benefits (time preference, the discount rate) and how they account for risks in outcomes (risk aversion). Third, people might face constraints in their choices. Credit constraints can influence the decision to attend school, but also a lack of facilities for education and less favorable family conditions can be treated as such constraints. Finally, the decision to invest in education will depend on information available at the time of investment. If people don't know about the benefits of education, it is unlikely they will invest.

The main reason it is important to take account of factors expected to influence willingness to invest in education is that they may have a direct impact on skill levels distinct from the indirect effect via the increased level of investment in education. If such factors are not taken into account, estimates of the effect of education on skill levels will be biased. The BQ covers some, but not all, of these factors. The questionnaire contains no direct indicators of innate learning abilities. It does, however, include a number of control variables that are related to this concept, in particular the family background in terms of parents' education (J_Q06b and J_Q07b). Learning strategies (I_Q04) are included as they may affect individuals' ability to learn.

The social environment

The constraints facing different social groups have been extensively studied by sociologists, who have a long tradition of research looking at the social barriers to education and training. While gender inequality in initial education has vanished and actually turned into an advantage for girls in many Western countries, it still persists in occupational careers and later access to training. The sex of the respondent is therefore a key reporting category for PIAAC. Inequality in access to education related to the family background both in terms of socioeconomic status and migration status is more persistent.

Part of these differences relates to differences in school performance and learning abilities, the so-called primary effects of social stratification (Boudon, 1974). These may be caused both by differences in innate abilities and socialization processes. The cultural capital of the family (Bourdieu, 1984) in particular provides a powerful predictor of the school performance. But even with the same school performance, students from different family backgrounds make systematically different choices in education (the secondary effects of social stratification), and

given the number of choices that have to be made during the educational career, the cumulative effect of these choices might even overwhelm the primary effects. These differences in choices relate to differences in social cost-benefit analyses. The social costs and benefits involved in obtaining education are different for students from different social backgrounds. Following an educational career that is different from the one that is common in the family induces social costs, while the social benefits may be lower. The BQ includes indicators of gender (A_N01), parents' education (J_Q06b, J_Q07b), migration status (J_Q04a-c, J_Q06a, J_Q07a), cultural capital in parental home (J_Q08), and language used in parental and current home (J_Q05a1-2, J_Q05b).

Effective learning and instruction

Following a certain type of education or training path does not automatically imply that all students are likely to acquire the same set of skills. Educational research has shown that there is considerable variation among educational systems, schools, study programs and teachers in how much skills students acquire during education or training. A large part of the effect of education on skill development is likely to be indirect, as students are turned into more effective or less effective learners for life. In other words, different characteristics of education may affect both the direct acquisition of skills as measured in the direct assessments, as well as the ability to acquire these skills after leaving education. Without providing too much detail, we can note a number of interesting approaches here:

- Situated learning theories (Glaser, 1991) emphasize that competencies and competence development are context-specific. They stress the importance of coherence and context-relevance (e.g., real-life experiments, simulation and practical work experience) in the design of the curricula in order to develop expertise.
- Active learning theories reject the traditional naïve model of the teacher as the expert, imparting his or her knowledge directly to the student. “Powerful learning environments” (De Corte, 1990) and active instructional methods like problem-based learning and project-oriented education are thought to foster the development of generic competencies like problem solving and metacognitive abilities.
- In addition to these innovative ways of learning based on elaborate theories on how individuals actually learn, educational research has traditionally stressed “time on task” as one of the most important factors affecting student outcomes. That is, the actual time students spent on education (within the classroom and through self-study) is a good predictor of the learning outcomes net of other factors.

Although it is not practicable to describe the educational environments respondents have been exposed to, it does make sense to include indicators of respondents' learning strategies, which may in part be a result of such exposure. As Peschar (2003) has remarked, such strategies can be seen as important prerequisites for learning throughout one's life. Self-regulated learning theories point to the relevance of metacognitive abilities and information-processing strategies of students (Kolb, 1984). Learning styles differ among students, ranging from memorizing and rather atomistic ways of learning towards a more constructivist approach in which concepts and theories are actively incorporated in a coherent body of knowledge. Although such attitudes are

likely to be heavily influenced by one's family background, either directly through genes or indirectly through early socialization, there is evidence that such attitudes and strategies can be influenced by education. Question I_Q04a-m contains indicators of learning strategies. Although the list of items has been strongly based on previous international comparative research, the question in its current form is new.

Among the characteristics of the educational career, the achieved level of education (B_Q01a) is, of course, the most important concept affecting skill levels. More years of schooling are expected to have a positive impact on the skills proficiencies. Based on the information of national experts, all reported national categories in the achieved level of education are converted into the nominal years of schooling needed to achieve that particular level of education (see Appendix 5). Moreover, the particular field of education (B_Q01b) followed will also affect skill levels: Graduates from certain fields of education will have higher scores in the literacy domain; others will probably have higher scores in the numeracy domain.

Other relevant characteristics of the educational career that may affect the skills development are the type of pathway in secondary education (whether a general or school-based vocational (B_Q01a). Based on the information of national experts, we determined for all relevant reported national categories in International Standard Classification of Education (ISCED) Levels 2 to 4 whether the types of pathway in secondary education was general or vocational (see Appendix 5). It is also important to identify whether the education has been completed outside the host country (in the case of migrants) in order to identify any negative effect on literacy skills. The BQ therefore contains information on where the highest qualification was obtained (B_Q01a2).

Training and informal learning

People do not only learn during initial education but later in life. In the human capital literature, many studies have analyzed the effects of workplace training participation on workers' wages (see Bassanini, Booth, Brunello, De Paola, & Leuven, (2005) for an overview). Several studies have found high returns on workers' participation in training. Brunello (2004) found that having recently attended training increases a worker's income by about 12 percent.

However, one may wonder whether it is really the participation in formal training that makes the difference. Borghans, Golsteyn and de Grip (2006) show that employees spend much more time on informal learning activities than on formal learning. They also found that when employers stimulate workers' participation in formal courses, these workers will also spend more time on informal learning in the workplace. As many of the studies on the effect of formal training do not measure the time spent on informal learning, all the benefits of the knowledge and skill acquisition of the workers are attributed to their participation in formal training. It is important that PIAAC not only looks at the incidence of formal training but also explores various kinds of informal learning, as they contribute highly to skills acquisition.

Arrow (1962) emphasized the importance of unstructured workplace learning, not from the perspective of the individual worker but that of the firm. He found that informal learning is a more or less automatic byproduct of the regular production process of a firm, which he labeled "learning by doing." Furthermore, job characteristics might also affect post-initial schooling. Employees with mainly monotonous tasks are expected to attend less formal training than those

in jobs with more complex tasks. Jobs that require problem solving and learning new things probably include high training incidence and informal learning as well.

Human resources practices and job characteristics are the major work characteristics that determine the opportunities for workers to attend training and learn in an informal way. Although these opportunities are often necessary for actual training behavior, a workplace characterized by these training opportunities might not be sufficient. Workers' characteristics will probably determine whether the learning opportunities at work are fully exploited. Personal characteristics such as age, gender and level of schooling are found to be important determinants of post-initial schooling behavior (Bassanini et al., 2005).

In addition to measures of participation in education at the time of the survey and over one's lifetime (B_Q01 to B_Q10), the BQ contains questions on recent investments in training (B_Q12 to B_Q20), including the main reason for participating in training (B_Q14b), crucial for analyzing the effects of training, informal training by supervisors, colleagues, etc. (D_Q13a), learning by doing (D_Q13b), keeping up to date with new products or services (D_Q13c) and work autonomy (D_Q11a-d).

Regardless of the specifics of the training and learning practices applied in the organization in which individuals work, the amount of work experience acquired can be expected to have a strong effect on skills development. In wage estimations, work experience is generally assumed to be positively related to productivity, but the returns are assumed to diminish with further experience. In terms of skill acquisition, this is consistent with the notion that one is likely to be most exposed to situations from which one can learn something new early in one's career. As the career develops, the chance that one will be exposed to new stimuli is decreased. This pattern is likely to be reinforced by typical patterns of brain development over the lifecycle, which predicts a steady decline in learning and retention abilities from young adulthood onward. We will return to this point below.

Skill acquisition is not only dependent on total experience, but also on the specific way in which this experience has been acquired. In addition to total work experience, the number and timing of changes of employer and/or career breaks is therefore also important. There is probably a certain minimum time one would need to remain with a given employer to have a reasonable chance of learning new things, and the returns to tenure in most jobs are likely to remain positive for at least a few years (although probably not in very low-level routine jobs, see below). Because the new experiences one can expect to be exposed to when working for a given employer are likely to diminish over time, we would expect a certain number of job changes over the career to have a positive effect on learning. Lengthy career breaks comprise periods during which the exposure to work-relevant experiences is likely to be limited.

In addition to these direct effects of work experience on learning, there may be indirect effects when work history is interpreted by potential employers as a signal of productivity and learning potential. In that case, a career characterized by frequent changes and/or lengthy interruptions may affect the willingness of potential employers to hire an individual and to invest in his or her human capital. Lengthy periods of unemployment – that is, seeking work without success – may additionally exert a negative effect on individual motivation.

The questionnaire contains a number of questions related to the above-mentioned aspects of employment history. Question C_Q09 allows us to establish the total number of years of work experience (if any) the respondent has acquired in his or her lifetime. Question C_Q10a provides information on the number of different employers worked for in the last five years.

Skill loss

The increased focus on lifelong learning in recent years has led to increased attention to the fact that individuals not only acquire skills over their lifetime but are also confronted with skill loss and a general decline in the ability to acquire and retain new knowledge and skills. The single-most important finding of IALS and ALL was that skill loss was large enough to offset all of the expected gains from increasing educational quality and quantity. Until now, only scattered studies on different aspects of skills obsolescence have been published. Most of these studies were published in periods in which unemployment was high. This increased the focus on the adverse impact of skills obsolescence for the workers involved. It is interesting that in the recent policy debates on skills obsolescence and “lifelong learning,” the main focus has been on the waste of valuable human resources and on the nonoptimal performance of workers with inadequate skills. This brings skills obsolescence to the heart of the economic challenge the western economies face: in realizing the transformation towards a knowledge-based society with an aging population.

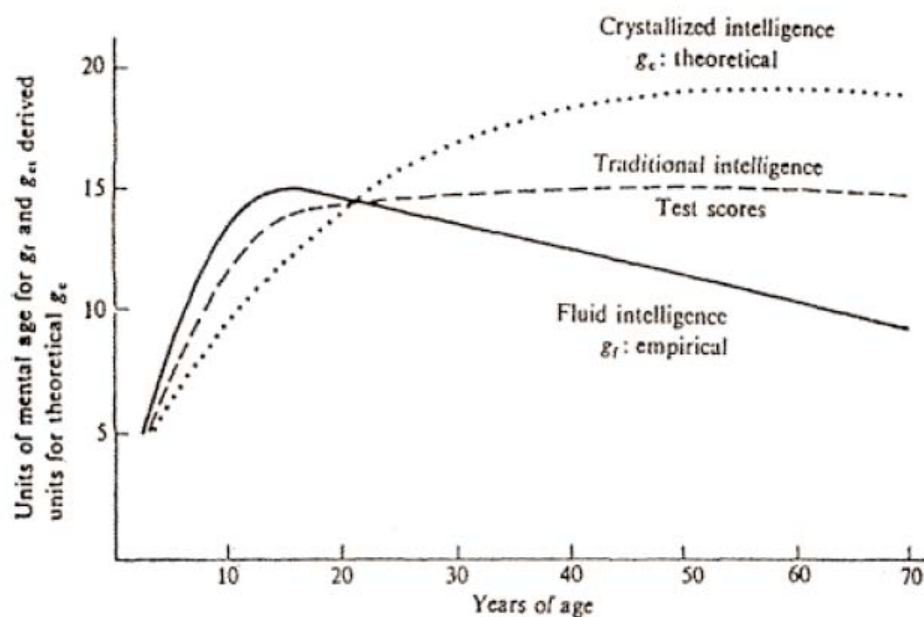
From a cognitive and neuropsychological perspective, higher order brain functions follow a steep developmental pattern and reach a plateau of optimal functioning in young adulthood. Such processes and changes therein can be measured on a behavioral level using dedicated neurocognitive instruments which tap the efficiency within specific neuropsychological domains, such as language, intelligence, memory, attention and speed of information processing.

Optimal neurocognitive development is dependent on a complex interplay of factors, with genetics, socioeconomic status, educational achievement, adequate nutrition, and uncompromised mental and physical health being the strongest predictors of developmental success. Researchers have coined the term “brain reserve capacity” (BRC) to indicate the neurobiological constraints which determine maximum processing capacity of higher order brain functions. This concept has proven its validity in, for example, predicting individual cognitive aging trajectories later in life. Important proxy measures of BRC include educational level and occupational achievement.

On a population level, most cognitive abilities such as memory function, information processing speed and attention capacity tend to decline with advancing age. Adequate preservation of cognitive abilities is of primary importance to older people, as cognitive decline can result in a loss of productivity among those still working, and a loss of independence and autonomy for retired people. Large individual differences exist in the offset and rate of decline of specific cognitive functions. We drew attention above to the theoretical distinction drawn in psychological research between “fluid” and “crystallized” abilities. The former refers to functions that involve controlled and effortful processing of novel information (cognitive mechanics), and the latter to the representation of learned skills and access to knowledge (cognitive pragmatics). Fluid abilities are far more sensitive to aging (Figure 3.1), and both cognitive domains show different developmental patterns across the life span. Fluid abilities typically start declining in the mid-20s, while crystallized skills may improve until and beyond

even the age of 70. The distinction between the two is important because the direct assessments in PIAAC will differ in the extent to which they relate to crystallized or fluid abilities. One may hypothesize that numeracy and literacy skills relate more strongly to crystallized abilities, while dynamic problem solving in a technology-rich environment will relate more to fluid abilities. For adults, the decline in fluid abilities is more likely to strongly hamper their working and everyday life than the decline in crystallized abilities.

Figure 3.1: Theoretical representation of ‘crystallized’ and ‘fluid’ abilities over the life span



*Figure 1. A theoretical description of life span curves of intellectual abilities. From *Intelligence: Its structure, growth and action* (p. 206) by R. B. Cattell, 1987, Amsterdam: North-Holland. Copyright 1987 by Elsevier Science Publishers. Reprinted with permission.*

The two most prominent symptoms of “usual” cognitive aging in daily life are a gradual reduction in memory retrieval and information processing speed. Stored information remains relatively intact, but access and retrieval becomes increasingly difficult for older individuals. Another feature that has received considerable interest in research is the reduced ability of older individuals to suppress or inhibit irrelevant information, making decision processes more complicated, and therefore slower.

Still, cognitive aging is not merely a predestined process which ultimately leads to pathological states, such as a cognitive disorder like dementia. The ability to learn new skills is still present in older individuals, but – on average – more time is needed to develop the same level of mastery as for younger persons. Recent advances in cognitive neuroscience have convincingly demonstrated that healthy brains show considerable capacity to compensate for reduced integrity of functional networks or to reorganize existing networks to adapt to changing task demands. The importance of adequate and continued exposure to environmental stimuli during the lifetime is now considered pivotal for optimal conservation of cognitive abilities in old age (conceptualized in the “use it or lose it” paradigm).

Empirical findings suggest that complex intellectual activity increases cognition of older workers (Schooler et al., 1999). Skill investments made during working life might improve people's capacity to continue learning and adapting to new environments. Other factors that are conjectured to affect the development of cognitive ability at later stages in life include occupation, leisure activities, lifestyle and social interaction.

Building partly on such insights from cognitive and neuropsychology, De Grip and Van Loo (2002) developed a typology of different types of skills obsolescence. First, the depreciation of human capital may simply be caused by the wear of skills, resulting from the natural aging process. Physically or mentally challenging working conditions may accelerate the wear of a worker's skills. Large epidemiological studies have shown that health-related factors are involved in the enhanced cognitive decline seen with increasing age. In addition, several chronic diseases have been associated with a reduced cognitive capacity in both epidemiological surveys and clinical case-controlled studies.

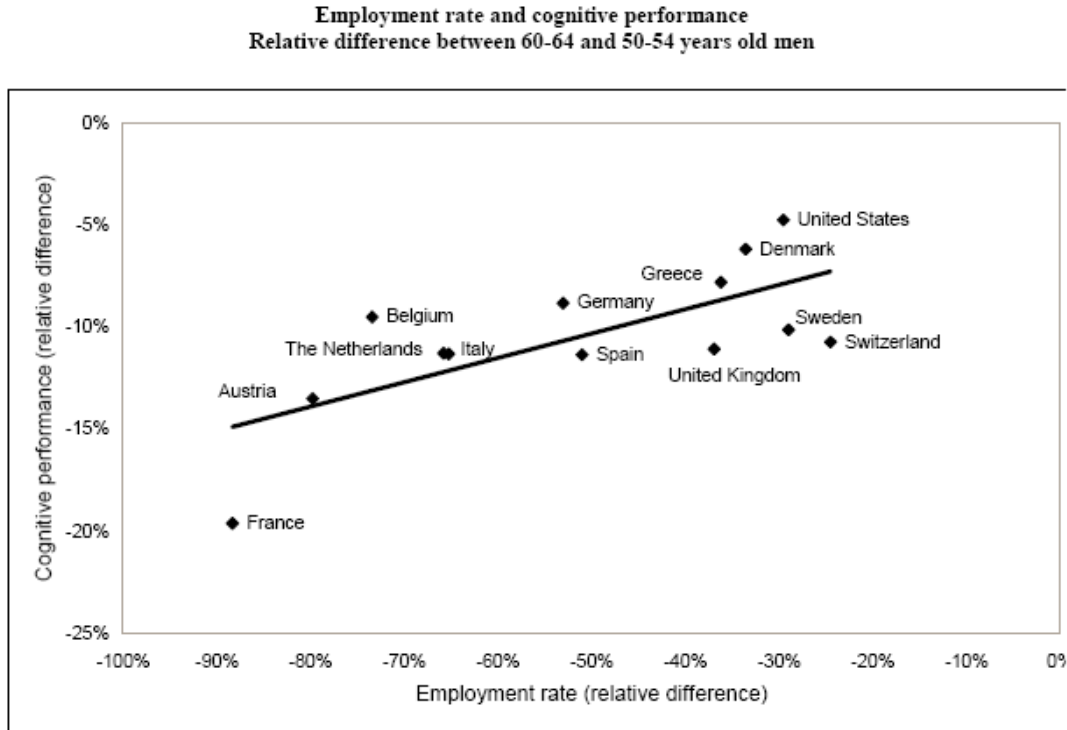
The second category of technical skills obsolescence concerns the atrophy of skills due to the lack or insufficient use of them. This atrophy could result from unemployment and career interruptions, or from employees working below their attained level of education. Arthur et al. (1998) conclude, on the basis of a meta-analysis from the psychological literature on skill decay and retention, that there is substantial skills obsolescence when they are not practiced or used. De Grip et al. (2008) show that job-worker mismatches induce cognitive decline with respect to immediate and delayed recall abilities, cognitive flexibility and verbal fluency. Also, as a result of specialization, certain knowledge and skills acquired during initial education may get lost. Apart from these two factors related to the personal characteristics of the worker, skills obsolescence may also occur as a result of changes in the demand for skills, due to, e.g., technological or organizational developments in the production process.

The BQ enables insight as to some of the possible causes of skills obsolescence, such as age (A_Q01a-b), health (I_Q08), unemployment (C_D05), working below one's level (D_Q12a-b), long tenure (D_Q5a1-2) and sector of industry (D_Q02a-b).

Institutional factors

There is a need to study whether policy and institutions can affect the process of cognitive decline. It is well established that early retirement decisions are largely driven by institutions. Gruber and Wise (2004), for example, show there is a very strong cross-country relationship between retirement rates and government policy. If keeping workers active can postpone cognitive decline, there is an important role for policies that increase labor market participation of older workers. Using data from the Survey on Health, Ageing and Retirement on cognitive skills of the population aged 50 and over, Adam et al. (2006) show that relative average cognitive skills among older workers are on average higher in countries in which – as a consequence of national institutions – participation rates of older workers are also higher (see Figure 3.2).

Figure 3.2: Employment rate and cognitive performance



Source: S. Adam, E. Bonsang, S. Germain and S. Perelman (2007), “Retirement and cognitive reserve: A stochastic frontier approach applied to survey data,” CREPP DP 2007/04, University of Liège.

Even though it is extremely important to better understand how the process of cognitive decline can be stopped and whether there is scope and need for policy intervention, the study of the determinants of cognitive decline is still in its infancy. Much can be learned from relating differences across countries to cross-country differences in policies, regulations and institutions. PIAAC offers a unique opportunity to gain such insights as it provides detailed data of the distribution of skills across age. By linking this type of data to information from other data sources on institutional factors, we can at least explore how these relations look at the aggregate level of countries.

Skills and outcomes

We remarked above that the policy relevance of measuring skills is strongly dependent on their effect on relevant outcomes. In addition to economic outcomes such as employment opportunities and rewards in the labor market, it is important to take account of outcomes in other areas that may also be influenced by skills, such as health status, voluntary work, and social trust.

Skills and labor market outcomes

Cognitive skills are a key determinant of an individual’s productivity, and therefore it is not surprising that cognitive skills are related to economic success. There is a large body of evidence

showing that higher cognitive skills are associated with better labor market outcomes (e.g., Heckman et al., 2006). The most basic of economic outcomes is an individual's current labor status, which is constructed using several questions in the questionnaire (C_D05). A distinction can be drawn between those who participate in the labor force and those who do not. The former category can be divided in turn into those who are employed and those who are unemployed (that is currently not working but available for and actively seeking work). There are several reasons why an individual might fall into the latter category – for example, study, household duties, or sickness or disability. To provide a broader indication of respondents' current situation, in question C_Q07 respondents are asked to report their own self-declared main labor status.

For those currently or recently in work, several important labor market outcomes are included in the questionnaire, including working hours (D_Q10), individual earnings (D_Q16-18), job security (D_Q09), occupational status (D_Q01a-b), and the quality of the match between education and work (D_Q12a-c)

One of the interesting questions in this respect regards the precise role of education and skills in producing these outcomes. There are rivaling hypotheses on this point. Very often the strong relation between education and labor market outcomes is explained in terms of human capital theory (Becker, 1964), which claims that people with more years of schooling earn more because the competencies they acquired in education have made them more productive. While this is probably true to some extent, at least in the aggregate, it tells only part of the story. Scholars such as Spence (1973) and Arrow (1973) have pointed out that the selection, allocation, and rewarding of individual employees takes place on the basis of signals such as formal qualifications as well as on the basis of productivity. This is usually explained in terms of incomplete information and bounded rationality. The signals form a solution to this problem, as they are assumed to indicate the average productive capacities of the group to which they refer. The labor queue theory (Thurow, 1975) adds an interesting twist, pointing out that many relevant competencies are not even learned in education, but picked up through work experience on the job. According to this theory, education is an indicator of low training costs rather than high productivity. Finally, some scholars have questioned whether education has any effect at all on graduates' ability to perform, pointing out that this relationship is in fact weaker than that between education and reward (Bills, 2003). This has led credentialists such as Collins (1979) to claim that higher education does not lead to superior competencies but is used by “gatekeepers” to legitimize the rationing of access to high-status, highly paid jobs.

In reality, there is probably an element of truth in all these theories. The crucial point then comes down to specifying the contexts under which one or the other mechanism prevails. The mechanisms are likely to differ according to the kind of job or position, labor market segment (private/public, economic sector), and country. In a study like PIAAC, we might expect large differences between the countries in the extent to which skills affect labor market outcomes relative to the effects of educational credentials. There is strong evidence that in countries characterized by a high degree of selectivity, stratification, and standardization, employers are more likely to select and reward employees on the basis of formal educational qualifications than in countries where education is less regulated (Müller & Shavit, 1998).

Many of the control variables that are needed to get unbiased estimates of the effects of skills on economic and social outcomes are comparable to the ones discussed above on the effect of education and training on skills development, although education and training will now be

treated as control variables instead of the predictor of interest. As indicated above, the highest level attained in formal education is one of the strongest predictors of skills. This is not only interesting in its own right, as a skill predictor or reporting category, but will likely be a confounding variable for many of the issues that policymakers are trying to understand in the context of PIAAC. Level of education is also a strong predictor of economic and social outcomes, and although this is often assumed to reflect differences in skill levels between levels of education, the precise causal mechanism is still somewhat controversial (Are the effects all directly attributable to human capital, or do theories of signaling and credentialism also tell part of the story?).

In this respect it is not only important to register highest formal level (which can be translated into number of years of formal schooling), but also the number of additional years of schooling that did not result in a diploma (which can be calculated as the difference between the year in which one last left education without completion (B_Q03c) and the year in which one last successfully completed formal education (B_Q01c). This schooling should lead to additional skills, and if the human capital theory is correct, to better outcomes.

In addition to level of education, labor market studies show large and robust differences in economic outcomes between fields of study in tertiary and secondary vocational education. Arts and humanities and social sciences often perform poorly, while business and engineering studies often do better than average. From a policy point of view, it is important to establish whether these differences are due to differences in the supply of and/or the demand for the skills of the graduates of these programs, to signaling or credentialism, to individual preferences, or to other factors.

The variables related to the number and intensity of received training is not only relevant in predicting skills, but also in predicting economic outcomes. As indicated above for education, the precise mechanism is not known and the estimates of the returns to training are biased by heterogeneous selection into training. For example, some people might get training because they are expected to be promoted instead of the other way around. We have included control variables like firm size (D_Q06a-b) to control for this unobserved heterogeneity. Most of these control variables are the same as the ones we discussed above. Additionally, when estimating effects of education and skills on outcomes, it is important to control for factors relating to household composition (J_Q01), family formation, as indicated by marital/cohabitant status (J_Q02a), and number and age range of children (J_Q03a-d), and job characteristics such as employee/self-employed status (D_Q04), supervisory status (D_Q08a-b) and job tenure (D_Q05a).

Skills and other outcomes

There is good empirical evidence that education not only affects labor market outcomes but is also a strong predictor of outcomes in other life domains. The BQ includes indicators of family formation (J_Q02a, J_Q03a-d), health (I_Q08), voluntary work (I_Q05f), political efficacy (I_Q06a) and social trust (I_Q07a-b). Education not only affects the individual outcomes in these domains but also affects social returns as a result of spillover effects. This is one of the reasons why policymakers are so interested in understanding these broader effects of education, because the social returns in terms of decreased costs for health and crime may well overwhelm the individual economic returns. The OECD recently published a report on the social outcomes of

learning (Schuller and Desjardin, 2007), underpinning this need for investment in education to increase health and civic and social engagement.

As with the effects of education on labor market outcomes, the effects of education on other outcomes are still not completely understood. Broadly, two mechanisms can be distinguished: an effect on skills and an effect on allocation. For the first effect we assume that education directly affects knowledge and skills that are relevant for healthy behavior, civic engagement, and so on. For instance, health programs may increase the knowledge of students in this area, leading to healthier behavior. The second mechanism refers to the role of education in allocating students to particular jobs or roles in society, for example, higher education increases the chance of ending up in healthier jobs or in social networks in which civic engagement is higher. In that case, the role of education is more indirect and it is not certain that investing in education will always have the anticipated effect. This is dependent on whether these outcomes are scarce resources or not. If people have to compete for scarce resources (as in the case of high-level jobs), investment in education changes the relative distribution but not the absolute.

From a policy point of view, it is therefore important to gain further insight into the underlying mechanisms. Moreover it is important to investigate to what extent low skills as a risk factor for social outcomes may be compensated for by other protective factors like job conditions, educational attainment, and so on.

3.3 The development of the JRA

In 2004 the OECD launched an initiative to develop a module in PIAAC on generic work skills requirements as a complement to the direct assessments. This was called the Job Requirements Approach (JRA). In the JRA, workers are asked to indicate the level of skills that is required in their current work in several skill domains. The basic idea of asking workers to report on skill requirements in their job is already older and has been successfully applied in different surveys, such as the British Skills Survey, similar surveys in Italy and Spain, the US O*NET survey, and several international graduate surveys (CHEERS and REFLEX).

The main arguments for developing a separate JRA module for PIAAC were the following:

- The direct assessments in PIAAC are limited to relatively few, albeit crucial, skill domains. Yet there was a widespread feeling, supported by some case studies, that other skills were becoming increasingly relevant in modern workplaces. Important examples were communication skills and the skills needed to work within teams, to work at multiple and flexible tasks, and to work more independently. There was also evidence that some of these skills, like computing skills, were being rewarded in the labor market over and above the returns to the education that people had received (Dickerson & Green, 2004). It was intended that the JRA module would provide a cost-effective way of assessing the relevance of these skills.
- Earlier skills surveys like IALS and ALL were mainly limited to the supply side of skills, that is, the stock of skills of the population. It was felt that some information on the demand side for skills was needed as well, that is, on the utilization of skills in the workplace. Sociological theory makes a distinction between “own skills” (the skills that

individuals have) and “job skills” (the skills defined by jobs), and it was decided to measure some important job skills directly.

In the JRA module, respondents were asked questions about the skills that they use at work. First, the module generated many items describing the generic activities involved in doing the job. The choice of items was informed by theories of skill and the practices of commercial occupational psychology. To reduce the multiple items to a smaller and theoretically meaningful set of generic skills, statistical techniques were used to generate several generic skill indicators from the responses on these items.

In the course of development of the BQ, it became apparent that parts of the JRA module corresponded to a large degree to measures of skill use that are required for analyzing the results of the direct assessment. The subject matter expert groups (SMEGs) in the areas of literacy, numeracy and ICT developed scales that integrated the experiences from ALL with the newly developed insights from JRA. Scales were developed that measure the use of skills both at work and in everyday life (including study) in a similar way. These scales are broadly comparable to what has been measured in ALL, but the scales were adjusted to have better psychometric properties. Items are now included for the three central domains covered by the direct assessments literacy (reading: G_Q01a-h, H_Q01a-h; writing: G_Q02a-d, H_Q02a-d); numeracy (G_Q03a-h, H_Q03a-h); and ICT (G_Q04, G_Q05a-h, G_Q06-8, H_Q04a-b, H_Q05a-h).

In addition to these three central domains covered by the direct assessments, the JRA module contains items pertaining to problem solving (F_Q05a-b) as well as a range of interaction/social skills: cooperation (F_Q01b), influence (F_Q04a-b), managerial skills (F_Q03b), self-direction (F_Q03a, c), horizontal interaction (F_Q02a-c) and client interaction (F2d-e), and physical skills [stamina (F_Q06b) and manual skill (F_Q06c)].

Two assumptions underpin the use of the JRA. First, it is assumed that the individual is well-informed to report about the activities involved in the job he or she is doing. All jobs differ, even within quite narrowly categorized occupations, and one would normally expect the job-holder to know best. Nevertheless, this might not always be true, and where the job-holder has only been in a post for a short time, the assumption might be questioned. In the case of out-of-work respondents, the Field Test has assessed the reliability of respondents’ ability to recall the activities of their most recent job in the previous 12 months. No indications were found that there was a serious recall bias. Second, it is assumed that the individual reports these activities in an unbiased way. This assumption might also be questioned: Individuals might talk up their jobs to boost their self-esteem. However, it is held that they are less likely to do so when reporting their activities than reporting how good they are in the performance of these activities. To minimize bias, the general principle was to ask respondents to report actual behavior, such as frequency of use and proportion of time spent on using different skills, rather than often-used alternatives such as the importance of these skills for the job.

The measures of “job skill” obtained through the JRA module are direct measures of the “own skill” held by respondents. Discrepancies between job holders’ skills and job requirements are possible, however. Some individuals may have an excess supply of some skills and not be using them fully on the job; others may have insufficient skills for the job they are doing but may survive in the short run despite the consequent poor performance. These mismatches are dynamic: They can appear and disappear as both jobs and people change. In the domains that are

also being directly tested, it will be possible to generate indicators of mismatch, where individuals have high levels of own skill and are in jobs where that same skill is used at a low level, or vice versa. There is also a general subjective question on self-perceived skill underutilization (F_Q07a). In several domains, however, there is no specific mismatch indicator available: The only indicator of skill in these domains will be the use of the skills in the job.

3.4 The development and validation of the BQ

3.4.1 The process of questionnaire development

Within the Consortium, the Research Centre for Education and the Labour Market was responsible for the development of the BQ. Advice on the BQ was given by the BQ expert group, consisting of the following members:

1. Prof. Ken Mayhew (chair), Pembroke College, Oxford and director of SKOPE, Research Centre on Skills, Knowledge and Organisational Performance.
2. Dr. Patrice deBroucker, Statistics Canada and member of OECD Network B.
3. Dr. Enrique Fernandez, European Foundation for the Improvement of Living and Working Conditions (Dublin, Ireland).
4. Prof. Francis Green, Professor of Labour Economics and Skills Development, Institute of Education, University of London
5. Prof. Masako Kurosawa, National Graduate Institute for Policy Studies, Japan.
6. Dr. Scott Murray, DataAngel Policy Research Incorporated.
7. Prof Jürgen Schupp, Honorary Professor for Sociology in the Faculty of Political and Social Sciences at the Free University and deputy director of the department Socioeconomic Panel Study at the German Institute for Economic Research DIW in Berlin.
8. Prof. Tom W. Smith, Director of the General Social Survey, National Opinion Research Center, University of Chicago.
9. Prof. Kea Tijdens, University of Amsterdam.
10. Prof. Robert Willis, Research Professor, Population Studies Center, University of Michigan.

Three meetings were held with the BQ expert group:

- 1-2 May 2008: Paris
- 23-24 June 2008: Offenbach (Frankfurt)
- 5-6 December 2010, Princeton, NJ, USA

Based on the discussions with the BQ expert group, several draft versions of the BQ were discussed with the NPMs, the BPC and the OECD.

For the inclusion of concepts and items in the BQ, we adopted the following list of criteria:

- The concepts must have a clearly established relation in the theoretical and empirical literature to skills and other relevant outcomes.
- Items must have good measurement properties in terms of reliability and validity and be able to maintain that over time.
- Items must be comparable across groups and across countries. This posed limits to items that may have been deemed vulnerable to cultural bias.
- Ex-ante harmonization was preferred over ex-post harmonization. National adaptations of questions (other than translation issues) were minimized and were only allowed in cases where it was functional (e.g., in asking about type of education, etc.).
- Wherever possible, items were preferred that were comparable with other international surveys. Most important was the comparability to IALS and ALL, but other international surveys such as the Labor Force Survey (LFS), World Value Survey (WVS) and the European Social Survey (ESS) constituted important markers as well.
- In general we recommended that most questions should be asked to everybody, or at least to a majority of the respondents. Developing items for small subgroups was minimized.

3.4.2 Two rounds of cognitive pre-tests

Rationale

Cognitive pretesting is an important tool for improving the quality and validity of questions (Willis, 2005; Beatty and Willis, 2007): They enable the identification of problems with the draft items, provide valuable insights into how the questions or specific terms are interpreted by respondents, how respondents use the given answer scales, how they recall (relevant) information, and how they make decisions and construct their responses. The results inform the evaluation and modification of survey questions.

As part of the overall validation strategy for the BQ, including the JRA, two subsequent rounds of cognitive pre-tests were therefore carried out with a selected subset of items.

The cognitive pre-tests were carried out in various countries and languages to forward PIAAC's goal of achieving comparability of instrumentation and measures. Countries were chosen to represent a maximum bandwidth of cultural and language diversity. The first round of cognitive pre-tests was conducted from August to October 2008, the second round from October 2008 to January 2009. The pre-testing phase included item selection, translation and adaptation of these items, as well as the development and translation of an interview guide with general specifications for conducting the cognitive interviews including a scripted protocol. After the interviews were completed, both country-specific reports as well an overall report with combined findings and including recommendations were produced for both rounds.

The investigated questions were selected by an expert group identifying those items that (a) had not been tested and extensively used in previous studies and (b) appeared to be problematic in their formulations and/or response options. Given that the JRA items had already been validated in a separate pilot study, the two rounds of pre-tests focused on the feasibility of using the JRA for the recently unemployed.

Methodology

The following section shortly describes the item and country selection, the translation process, the specifications for the administration of the cognitive interviews, and the sample scheme.

Item selection

The item selection was based on version 3.1 of the BQ for the first round, and on version 3.5 of the BQ for the second round of cognitive pre-tests. Items were selected by staff from GESIS – Leibniz Institute for the Social Sciences, by experts in the field of cognitive pre-tests, and by item developers from the Research Centre for Education and the Labour Market (ROA). Items were selected according to criteria such as inclusion of crucial variables, inclusion of items with certain response categories and scales, or inclusion of items that had been identified as potentially problematic.

Due to the process of probing and follow-up probing, an item in the cognitive interviewing context requires much more time than in a standard interview. To reduce respondent burden, it was therefore necessary to limit the number of items and to restrict total interview duration, with 90 minutes recommended as an utmost maximum (Prüfer and Rexroth, 2005). Thus, for each of the two rounds of pre-tests, a total of 30 items were selected. Respondents answered different sets of questions depending on their education and employment status, with a maximum of 20 items per respondent.

Country selection

The first round of cognitive pre-tests was conducted in three PIAAC countries (United States, South Korea, and Germany), the second round in five PIAAC countries (United States, South Korea, Germany, Sweden and Spain). The countries were selected to cover three important linguistic areas and cultural regions: North America, Central Europe and Asia. Furthermore, this selection allowed the English source questionnaire to be pre-tested, thus ensuring that the potential problems identified in the cognitive pre-tests were not only due to translation, but rather to general design issues.

Item translation

Item translation was accomplished via double translation by two independent translators, followed by reconciliation. Problems and questions that arose during the translation process were communicated to the item developers and ambiguities were clarified.

Interviewer guide and techniques

An interviewer guide was developed by the cognitive pre-testing experts at GESIS – Leibniz Institute for the Social Sciences. The protocols in the interviewer guide integrated two techniques

typically used in cognitive pre-tests: *Paraphrasing* and *probing*. The *paraphrasing* technique asks respondents to reformulate the question in their own words. This method provides information on how the item is understood by the respondents and whether this interpretation matches the question intent. The *probing* technique comes into play after the respondent has answered a survey question and focuses on specific issues (e.g., how the item is understood, potential ambiguities or reasons for choosing a specific answer category).

The interviewer guide specified item-by-item instructions on how to conduct the cognitive interview. It included information on probes and additional questions, as well as specifications for the data format. The interviewer guides were translated and used in each country to ensure that the same techniques and procedures were used for specific questions across all countries.

Administration of cognitive pre-tests

The cognitive interviews were carried out face to face and were audio-recorded. Prior to the cognitive pre-test, respondents were informed that the aim of the interview was to evaluate and improve questionnaire items, and not test the respondents. All institutes carrying out the cognitive pre-tests gave monetary incentives for participation.

Sample and quota scheme

The requested sample size was 25 respondents per country for each round of cognitive pre-tests, with a predefined quota scheme. This scheme called for respondents with specific combinations of education, and employment status, with a heterogeneous distribution of age and gender (cp. Table 3.1).

Table 3.1: Quota Scheme for Cognitive Pre-tests (Round 1 and Round 2) (N = 25)

	Lower educational level (ISCED < 3) N = 17		Higher educational level (ISCED ≥ 3) N = 8		
	Not in education	Currently in education	Not in education	Currently in education	Total
Job	6	---	3	---	9
Recent job	6	---	3	---	9
No recent job	3	2	1	1	7
Total	15	2	7	1	25

Results

All respondent-level information was carefully reviewed. This included evaluating the detailed protocol results, concrete responses to the items (e.g., which category on the answer scale), as well as spontaneous respondent reactions. Respondent results were supplemented by interviewer comments. As a result, item-specific recommendations were derived. Collating and merging results from interviewers and respondents from different cultural and linguistic backgrounds greatly enriched the pre-testing findings.

The results of the cognitive pre-tests Round 1 were presented at the NPM Meeting in Lisbon in October 2008 and a report was sent to participating countries, the item developers and the BQ expert group. The results of Round 2 were communicated through a written report in February 2009. The recommendations in the reports were considered for the further development of the BQ.

3.4.3 Analysis of Field Test data

The BQ for the PIAAC project was developed with a view towards supporting the three broad policy questions described above that are central to PIAAC as a whole. First of all, it was designed to provide a clear view of how skills are distributed in the adult population. The second broad policy question underpinning the PIAAC project was to establish why skills are important. The third was the need to determine what factors are related to skill acquisition and decline. It is these policy considerations that have shaped the selection of items for the BQ as used in the Field Test.

The analysis of the data from the Field Test was guided by a number of main goals. Regarding contents, it was primarily aimed at validating the BQ by examining its general feasibility, empirical item and scale properties, quality of the underlying concepts and its operationalization. Regarding length, it was first necessary to assess the average time needed to complete the questionnaire, or subsets of items, in order to estimate by how much the questionnaire needed to be reduced to achieve a practicable questionnaire length for the Main Study. The combination of these two analyses helped identify items that could potentially be removed while making sure that main reporting variables were retained and the BQ still addressed PIAAC's main policy goals.

Moreover, the analysis aimed at discovering irregularities in the country data sets that could reveal potential translation errors or technical problems during the BQ administration.

In order to achieve these goals, the following analyses were conducted:

- A timing analysis assessing the average duration of the administration of the BQ
- An item-based analysis focusing on item nonresponse, item response distribution and response duration
- A scale-based analysis assessing the reliability and functioning of the BQ's multi-item scales both within and across countries
- An analysis of the functioning of the items representing main concepts such as education and training, labor market and other outcomes, and noncognitive skills
- Routing checks of crucial filters and branching rules within the national BQs

All analyses were conducted at an overall (international) level. The timing analyses, the item-based analyses and the routing checks were also run at the country level. Countries included in the international item-based analysis and routing checks were Austria, Chile, Cyprus,¹ the Czech

¹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Republic, Denmark, Estonia, Finland, Flanders (Belgium), England/Northern Ireland (UK), France, Germany, Ireland, Italy, Japan, Korea, the Netherlands, Norway, Poland, Portugal, Spain and Sweden.² At the time the analysis at the overall level was conducted (November 2010), the Field Test had not been completed in Canada, Slovakia, the Russian Federation³ and the United States; they were thus not included in the overall analysis. In addition, national reports giving detailed information about item distributions, durations, routing and potential irregularities were provided to each country.⁴

The analyses of the multi-item scales and the functioning of the main concepts are based on PIAAC Field Test data of 18 countries – those included in the overall analysis, excluding England/Northern Ireland (UK), Flanders (Belgium) and Norway.

Completed interviews and partial completes, were taken into account in the analysis. Across countries, a total of N=81,597 interviews (completes and partial completes) were analyzed.

For all timing analyses, only completed cases were included. In order to eliminate outliers at the item level, the data were trimmed by replacing all item time values beyond +/- 4 times the median of each item per country with the value of +/- 4 times the median. All 20 countries mentioned above with the exception of Spain were included for the timing analysis.⁵

The assessment of the questionnaire length

With respect to the BQ, the main goal of the Field Test was to finalize the instrument to be used in the Main Study, which in practice primarily meant a significant reduction in length. The Field Test intentionally included more items than were to be implemented for the Main Study. This total Field Test questionnaire was estimated to take some 55 to 60 minutes on average. To make the Field Test as realistic as possible in terms of total time of the interview, it was decided to use a random module design. All respondents got a core questionnaire and one of four modules: one with questions on the use of nonliteracy skills at work (section F), one with questions on skill use in reading and writing (first parts of sections G and H), one with questions on skill use in numeracy and ICT (second part of sections G and H), and one module with questions on noncognitive skills and noneconomic outcomes (section I). This approach was thought to bring back the total interview time for the Field Test to some 40 minutes.

As the preparation for the Main Study needed to start quite soon after the data collection for the Field Test, the Consortium, countries and the OECD agreed early in 2010 on a two-phase process for revising and adapting materials for the Main Study BQ. Phase I took place prior to

² In **Estonia**, approximately 15% of the interviews had been administered in Russian language, and 85% in Estonian. In the analysis, interviews conducted in Russian were not taken into account. In **Portugal**, due to an error in the random assignment of the BQ modules, certain sections of the BQ were omitted from analysis. **Australia** did not share its dataset due to data confidentiality reasons.

³ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

⁴ Including Canada, Slovakia, Russia and the United States but with the exception of Australia.

⁵ There were doubts as to the representativeness in the sample taken from the population for the Field Test in Spain. This gave rise to a lower than average response duration in that country. For this reason, Spain was excluded from the database for the timing analysis.

the analysis of the Field Test data (between January and September 2010), while Phase II utilized results obtained in the analysis of the Field Test data (November and December 2010).

In the case of the BQ, Phase I of the revision process began with the creation of an interim BQ. This revised version of the Field Test BQ was based on recommendations provided by the OECD/BPC, which identified a prioritized list of questions to be deleted and areas of the questionnaire where further reduction in the number of items could be made if supported by results from the Field Test once that data was available. The interim BQ was finalized in June 2010.

Based on the Field Test data, this interim BQ was estimated to take some 45-50 minutes on average, ranging from under 40 minutes for inactive respondents to just over 50 minutes for employed respondents.⁶

The second step of the revision was data driven, which meant that it could only be implemented after the data collection for the Field Test had been conducted and an international data file prepared. The requirement was to reduce the BQ in its final form to a length of 40 minutes for the common core, with a maximum of an additional five minutes allowed for countries to add any national questions they considered necessary for their own purposes. This meant that the interim BQ needed to be cut back another 5-10 minutes. The rationale for dropping items in this interim BQ was based on a thorough analysis of the functioning of individual items as well of the concepts that were made up of individual items (e.g., scales). In the next paragraphs we report the main findings of the Field Test analysis on which we based the decision to further reduce and finalize the interim BQ.

Individual item functioning: item distribution and item nonresponse

Item nonresponse

Item nonresponse was assessed across countries with a focus on (a) questions on the individual's income, as these questions are known to have high nonresponse rates, and (b) questions asking about past behavior (retrospective questions), in order to explore whether the JRA questions could be administered to the currently unemployed estimating the requirements of their past job. Item nonresponse was also investigated at the level of single countries and language groups within countries, as this might indicate potential country-specific translation errors or technical problems during the BQ administration.

For most of the BQ items and across all countries, nonresponse was very low (1% or less per question). However, some items showed higher nonresponse rates, such as the open-ended income questions:⁷ 9% for employees [6% refused (RF), 3% don't know (DK)] and 26% for self-employed (12% RF, 14 % DK). However, the follow-up questions asking those who did not

⁶ The PIAAC BQ is a highly adaptive instrument with a large variety of routings depending on education, labor force status, and other variables. As we used a random module design, it was not possible to simply add up the time spent on the different items in the Field Test. Thus, different methods were used to arrive at a reliable time estimate and the time was calculated for different types of respondents: employed, unemployed and inactive with accompanying assumptions on the share of these people routed into different questions (e.g., the share receiving training).

⁷ (D_Q16a - D_Q18c2).

respond to the open-ended questions to report their income in broad income categories proved to be effective. Indeed, among employees, the total item nonresponse decreased from 9% to 5% and for self-employed from 26% to 11% after having presented the follow-up questions.

In order to find out whether the JRA questions and other job-related questions were more difficult for those currently unemployed (but with work experience in the last year) than for those currently in paid work, the respective nonresponse rates between the two groups were compared. Results show that most job-related retrospective questions did not have increased “don’t know” or refusal rates among the currently unemployed.

Item response distribution

Response distribution was examined in order to identify items or response categories (a) that could potentially be deleted or (b) that reveal irregularities indicating potential translation errors or data entry or coding issues. In addition to an initial visual inspection of the item response distribution and the number of respondents per item, statistical key figures of interval- and ratio-scaled items such as the mean, standard deviation, skewness and kurtosis were analyzed.

Overall, it can be stated that nearly all items behaved in the expected way and only minor issues were discovered.

Most *ordinal and nominal* level items were distributed as expected, that is, they showed sufficient coverage in terms of frequencies across all response categories. No severe floor or ceiling effects were detected for any of these items. For the majority of the *open-ended questions* with manually entered numeric data, improbable values were identified in the answers. However, the total number of respondents with improbable and/or impossible responses was very small. Moreover, some of these outliers seem to be simply caused by technical problems. For the Main Study this led to some revisions in the minimum and maximum values that could be assigned.

Two items showing slight problems in this regard were the intensity of formal qualification and the last learning activity. Results indicated that some respondents may have had difficulties in assessing the time spent on their formal education or learning activities, particularly when judging this time in terms of hours. For example, when asked for the hours spent on their formal education, 10% of respondents said they didn’t know. Other respondents indicated they had spent 24 hours per day on learning activities (1%, n= 57).⁸ Due to these problems as well as the limited predictive power of these items, the Consortium, in consultation with the BQ expert group, simplified this set of questions for the Main Study.

Potential translation/comprehension issues were identified with respect to the interviewer instructions for items regarding the number of learning activities during the past year:⁹ Interviewers were advised to “count related learning activities held on different days as a single episode.” In nearly half of the countries investigated, the maximum number of reported learning activities was 50. This might indicate that respondents and/or interviewers thought the number of *lessons* (in units) was meant, and not the number of *courses*. These observations led to some modifications in the interviewer instructions.

⁸ Formal education: B_Q09a; time spent on learning activities: B_Q19b.

⁹ Question B_Q12b-B_Q12h.

Some **other irregularities in data entry and coding or technical issues** were identified. For example, national adaptations were not coded back into the international core variables for all countries.

Routing

As the PIAAC BQ is a highly adaptive instrument, it contained several routings depending on education, labor force status and other variables, such as computer use.¹⁰ In order to investigate the functionality of the national BQs, various routing checks were conducted in the Field Test analyses. In detail, it was tested whether respondents did indeed arrive at the questions they were intended to receive (and no other questions) as defined by the BQ design. These routing checks focused on the crucial filters and branching rules in the national BQs and were not exhaustive. For each country, two general types of checks were run: a) within section filters, to a large extent focusing on the operative functioning of the derived variables, and b) between section routings, also taking the BQ random modules into account.

The analysis of the routing showed that no systematic routing issues were observed and generally the flow of the BQ worked as intended in all countries. Overall, routing checks within and between sections yielded only a few minor issues and affected only individual cases. Within-routing checks showed, for example, that in 0.1% of the cases, respondents did not receive any computer use question even though at least one of these items was to be received by all respondents. Across-section routing checks revealed only very few incidents where problems with transitions from one into another section occurred. These problems were most likely due to technical issues.

Multi-item scale functioning in sections F, G and H

Four sections of the BQ contain collections of items that can be regarded as multiple indicators of the same construct. Section F contains a collection of items around different types of nonliteracy skills used at work, while sections G and H contain collections of items around literacy skills. Section I contains scales that largely address inter-individual differences in terms of perseverance, learning strategies, locus of control, and others. This section will be discussed in the next paragraph.

These four sections were also the ones that were subject to pseudo-random assignment of respondents (rotation), so that analytic strategies had to take into account that not all variables in these sections have been observed with all other variables. Also, the fact that four randomly assigned rotations were used limited the sample size to about one-fourth of the realized sample size in each country. Taking these limitations into account, the analyses conducted with the Field Test data from these sections took the following two routes:

Exploratory route: Section F contains up to three items per skill domain such as communication, planning, advising, and others. The interrelations between skill domains in section F, and how these skills are related to different occupations, were expected to be of interest for reporting the main test data. Exploratory analyses were carried out using factor analytic techniques

¹⁰Questions G_Q04 (“Do/Did you use a computer in your Job/Last job?”), H_Q04a (“Have you ever used a computer?”), and H_Q04b (“Do you use a computer in your everyday life now/outside work?”)

(summarized below), while other analytic techniques such as latent class analyses and hierarchical multidimensional item response modeling were also explored.

Confirmatory route: Sections G, H, and I contain well defined and larger collections of items around topics and can thus be regarded as psychological or behavioral scales. In order to evaluate the functioning of the scales in these sections, reliability analyses and scale refinement, as well as predictive analyses using a proxy of the respondent's test score (the so-called ETS zlogit score) were conducted.

Exploratory analyses of section F

For the exploratory analysis of section F, data were pooled across countries. Data from 21 countries as available on October 29, 2010 were used in the analysis.

The Consortium ran a factor analysis with subsequent Promax rotation. Given the results obtained from these analyses, it can be expected that profiles of skills based on the items in section F can be formed. The results of these analyses suggested factors that could be referred to as 1) cooperation, 2) advising, selling and negotiation, 3) teaching and presenting, 4) planning, and 5) physical work. It should be possible for these results and factors to be further refined in the future based on within-country analysis, because the Main Study provides sufficient sample size for such analyses.

However, two items did not perform as expected and were therefore recommended to be removed. Item F_Q01a appeared somewhat ambiguous, while item F_Q06a was redundant, covering much the same meaning as F_Q06b (correlation between the two was >0.7), which performed somewhat better in other respects.

Confirmatory analyses of sections G and H:

The items in sections G and H are to a large extent aimed at a parallel assessment of self-reported literacy skills around reading, writing, numeracy and ICT. Each item in these two sections belongs to exactly one of these four skill domains. Therefore, confirmatory 1-factor models (to check item coding is working as expected) and reliability analyses were conducted for four scales each in sections G and H.

The reliability analyses were conducted by country and then aggregated across countries yielding the following results:

- G_Q01 (Skill Use Work – Literacy – Reading).
 - The scale consisted of eight items: G_Q01a, G_Q01b, G_Q01c, G_Q01d, G_Q01e, G_Q01f, G_Q01g and G_Q01h. Across countries, the average Cronbach's Alpha was .82 (SD =0.04). The reliability did not increase when leaving any of the items out.
- G_Q02 (Skill Use Work – Literacy – Writing).
 - The scale consisted of four items: G_Q02a, G_Q02b, G_Q02c, and G_Q02d. Across countries, the average Cronbach's Alpha was .63 (SD =0.07). The reliability increased to a value of .66 when leaving item G_Q02b out (SD = 0.09). In addition,

- the average item-total correlation of item D_Q02b was lower than .3 ($M = .24$, $SD = 0.07$).
- G_Q03 (Skill Use Work – Numeracy).
 - The scale consisted of eight items: G_Q03a, G_Q03b, G_Q03c, G_Q03d, G_Q03e, G_Q03f, G_Q03g and G_Q03h. Across countries, the average Cronbach's Alpha was .83 ($SD = 0.02$). The reliability increased to a value of .85 when leaving out item G_Q03a ($SD = 0.02$). However, the average item-test correlation of item G_Q03a was close to .3 ($M = 0.29$, $SD = 0.06$).
 - G_Q05 (Skill Use Work – ICT – Internet and Computer).
 - The scale consisted of eight items: G_Q05a, G_Q05b, G_Q05c, G_Q05d, G_Q05e, G_Q05f, G_Q05g and G_Q05h. Across countries, the average Cronbach's Alpha was .81 ($SD = 0.03$). The reliability increased to a value of .82 when leaving out item G_Q03g ($SD = 0.03$). In addition, the average item-test correlation of item G_Q05g was smaller than .3 ($M = .25$, $SD = 0.10$).
 - H_Q01 (Skill Use Everyday Life – Literacy – Reading).
 - The scale consisted of eight items: H_Q01a, H_Q01b, H_Q01c, H_Q01d, H_Q01e, H_Q01f, H_Q01g and H_Q01h. Across countries, the average Cronbach's Alpha was .72 ($SD = 0.04$). The reliability increased slightly to a value of .72 when leaving out item H_Q01e ($SD = 0.06$). In addition, the average item-test correlation of item H_Q01e was smaller than .3 ($M = .28$, $SD = 0.07$).
 - H_Q02 (Skill Use Everyday Life – Literacy – Writing).
 - The scale consisted of four items: H_Q02a, H_Q02b, H_Q02c and H_Q02d. Across countries, the average Cronbach's Alpha was .51 ($SD = 0.12$). The reliability did not increase when leaving an item out.
 - H_Q03 (Skill Use Everyday Life – Numeracy).
 - The scale consisted of eight items: H_Q03a, H_Q03b, H_Q03c, H_Q03d, H_Q03e, H_Q03f, H_Q03g and H_Q03h. Across countries, the average Cronbach's Alpha was .84 ($SD = 0.02$). The reliability did not increase when leaving any of the items out.
 - H_Q05 (Skill Use Everyday Life – ICT-Internet and Computer).
 - The scale consisted of eight items: H_Q05a, H_Q05b, H_Q05c, H_Q05d, H_Q05e, H_Q05f, H_Q05g and H_Q05h. Across countries, the average Cronbach's Alpha was .75 ($SD = 0.04$). The reliability increased to a value of .76 when removing item H_Q05h ($SD = .03$). However, the average item-test correlation of item H_Q05h was larger than .3 ($M = .32$, $SD = 0.09$).

Except for the writing skill scales G_Q02 and H_Q02, the reliabilities of the scales in sections G and H were quite satisfactory. Note that the writing scales with four items each were also the shortest scales in the literacy skill-use sections.

Predictive analyses were also conducted on the scales in section G and H. Predictive analyses were conducted by country and then summarized across the 21 countries. The correlations of self-reported skill-use scales with the zlogit score were at a moderate level and consistent across scales as well as countries. Compared to other measures such as the ones collected in section I, the skill use correlations with zlogit were higher. Note that even the least reliable (and shortest) writing skill use scale on average correlated with the zlogit 0.256 for skill use at home and 0.269 for skill use at work. The good consistency of skill use scales in terms of reliability and predictive validity led us to believe that these scales would be among the most valuable predictors of outcomes in modeling and reporting of the Main Study data.

Functioning of concepts

In this part of the analysis, we looked at the functioning of the key concepts in the BQ. We looked at items related to respondents' socioeconomic background, education and training, their labor market outcomes, some possibly relevant noncognitive skills, and some other outcome measures. We used a range of methods of analysis, including univariate (inspection of frequency distributions), bivariate (relation with other relevant indicators), scaling (mutual correlation of sets of items) and multivariate (relation with outcome measures, controlling for other characteristics) methods.

Background, education and training

Socioeconomic background (J_Q06b-e, J_Q07b-e)

The BQ contained five indicators of respondents' socioeconomic background, namely the highest level of education (in three broad categories) ever attained by both parents, the occupational code of both parents when the respondent was age 16, and the number of books in the household when the respondent was age 16 (as indicator of the level of cultural capital in the parental home). With the exception of some possible minor measurement issues in some countries, which were referred back to the countries involved for checking and where necessary correction prior to the Main Study, these variables all performed well in the analyses. They showed plausible frequency distributions and were related in the expected way to each other and to respondents' education, occupation, earnings and skills. This applied not just to bivariate relationships between the indicators of socioeconomic background and these characteristics of respondents, but continued to hold in multivariate analyses with controls for gender, age, field of study of highest completed education, employment status, immigrant status, cohabitation status, parenthood, country of residence and respondents own education and occupation (the latter with the exception of those analyses where these were the dependent variables).

However, the predictive power of parents' education was in almost all cases greater than that of parents' occupation, which added little additional explained variance once parents' education was included in the analyses. The only exception was when the respondent's own occupation was the dependent variable. Understandably, parents' occupation was in this case a better predictor than parents' education, but even here parent's education showed a significant effect.

Taking into account the length of time required for administering the questions on parents' occupation (around 1.5 minutes on average), it was decided that this was a strong candidate to be dropped from the BQ for the Main Study. The number of books in the parental household was a strong predictor of test score proxy and other relevant outcomes.

The recommendation was to retain items on parents' education and number of books in the household at age 16, but drop items on parents' occupation for the Main Study.

Level of education (B_Q01a, B_Q01a3, B_Q02b, B_Q02b3, B_Q03b, B_Q03b3, B_Q05a, B_Q05a3)

The component variables for this set of indicators were the highest completed level, the education level engaged in by those currently in education, and the highest level of education of programs that respondents may have started but failed to complete. All three indicators were asked separately for home country and foreign qualifications, so it was necessary to combine these into a single measure. All three indicators were initially composed of detailed ISCED codes distinguishing 13 levels as well as a category of "No formal qualification or below ISCED 1." For the purposes of most analyses this was recoded into three broad levels ("ISCED 1, 2 and 3C short," "ISCED 3C long, 3A-B and 4" and "ISCED 5 and 6). Again with the exception of some minor country-specific issues, these variables performed well in the analyses. They were plausibly related to each other, as well as to respondents' occupation, earnings and skill level. Being currently engaged in education at a higher level than the highest completed level or having left education at a higher level without completion was associated with higher skill levels even after controlling for highest completed level of education.

Due to the extremely tight timeline available for revising the BQ for the Main Study, the separate items on level of foreign qualification for current, unfinished and recent education were already dropped prior to the analysis of the Field Test data. Because few respondents reported foreign qualifications, the data analysis provided no reason to reverse this decision.

Taking into account the fact that the separate items on level of foreign qualification for current, unfinished and recent education were already dropped prior to the data analysis, the recommendation was to retain the remaining set of items unchanged for the Main Study.

Field of study (B_Q01b, B_Q02c, B_Q05b)

For highest completed education, current education and other education followed in the last 12 months, respondents were asked to report their field of study (ISCED 97 broad fields of education and training, i.e., 1-digit codes). Apart from some minor country-specific issues, these variables all performed well in the analyses. They behaved well in terms of their frequency distributions, which were plausible and similar in all three cases, with a main exception that current and recent education tended less often to be general programs than highest completed education. This latter finding is consistent with the tendency for education to become progressively more specific as the educational career progresses.

In all fields of highest completed education, the most frequent choice of subsequent field of education was the same one. In addition to this relation with the field of study for current or

recent education, the field of study of the highest completed education showed a clear and plausible relation with occupation, economic sector, gender, earnings and skill levels. This result held not only in bivariate analyses but in multivariate analyses that controlled for level of education as well as other relevant indicators such as gender, age, employment status, immigrant status, and country of residence. This confirmed that field of study is a relevant dimension in addition to the level of education.

The recommendation was to retain this full set of items unchanged for the Main Study.

Training participation and intensity (B_Q06-B_Q09, B_Q17-B_Q20, B_Q22-B_Q25)

Component variables for training participation and intensity were the number of training episodes in the last 12 months, hours of training of most recent episode, hours of training of second-most recent episode, proxy total time spent on training (a construct based on the former three variables), and the time spent in the last 12 months on education. These variables are inherently skewed: Most people follow little or no training, but a small number invest heavily in training. The skewedness is accentuated by some apparent measurement difficulties.

Several factors are likely to have contributed to these measurement difficulties. For the number of training episodes, it seemed likely that a small number of respondents reported repeated sessions of the same training episode (for example, a weekly language course) as separate episodes, which resulted in an implausibly large number of reported episodes for a small number of respondents. For hours of training in the two most recent episodes, there were also some implausibly high values, which seemed to be largely – although possibly not entirely – due to the fact that for those who opted to report training in weeks or days as opposed to hours, the final measure was based on answers to two separate questions that then needed to be multiplied with each other to produce the final measure. An error in either answer would therefore be multiplied and will thus result in an even larger error in the final indicator.

This problem was even greater for the proxy for total time spent on training, which was based on the number of training episodes and the time spent on the last two episodes. Because we lacked data on time spent on all but the last two episodes, this indicator was necessarily inaccurate at the individual level, and this problem was compounded by measurement error.

A more “holistic” method of measuring training duration was introduced for the Main Study, and which is believed to have reduced the measurement error, although it may not have removed it altogether because of the inherent difficulty of asking respondents to report the duration of all episodes combined. For the purposes of the analyses reported here, we assumed that high values on all the indicators were likely to be inaccurate, so we removed extreme values prior to analysis. After these adjustments, especially the indicator for training frequency was well behaved. The frequency distributions of all these variables appeared plausible. When related to other relevant indicators especially training frequency was well behaved, showing clear relations with level of education, occupation, earnings and skill level. These relations held up well in multivariate analyses after controlling for other relevant indicators, and training frequency and even training incidence (training yes/no) were also strong predictors of labor force status as well as noneconomic outcomes such as health, civic engagement and social trust. Training duration also showed some effects on other variables, but these effects were generally much weaker and less consistent.

Taking into account measurement issues, the limited predictive power as well as the length of time required for administering the questions on training duration (estimated at three minutes on average for the “holistic” measure of training duration proposed for the Main Study and an additional 1.5 minutes on average for the duration of participation in education in the last year), it was decided that these were strong candidates to be dropped from the BQ for the Main Study.

The recommendation was to retain questions on training frequency, but drop questions on training duration for the Main Study.

Labor market outcomes

Labor force status (C_Q01-C_Q05, C_Q07)

Formal labor force status, which differentiates the statuses “employed,” “unemployed” and “not in the labor force,” is constructed on the basis of answers to a series of questions on whether respondents are currently employed, available for work, waiting to start work, or have taken active steps to find work. There are two versions of this indicator, an Australian version automatically generated while the BQ is administered, and a European version. The difference between these two versions is both conceptually and empirically minor, with the sole difference being whether looking at job advertisements in the newspapers is regarded as an active step or not. There are only marginal differences in the frequency distribution in either case (several tenths of a percent shifting between “unemployed” and “not in the labor force”), and regardless of which version is used, these variables all performed well in the analyses. In both cases the frequency distribution was plausible. There was a clear relation between formal labor force status and subjective employment status (i.e., how respondents see themselves), but these were far from identical. However, the differences between subjective and objective status were plausible, with, for example, a considerable proportion of those who saw themselves as unemployed being formally out of the labor force. Labor force status was also related in a plausible fashion to education and skills. There was no real relation with parents’ education, but this did not seem to indicate a problem with either indicator.

The recommendation was to retain this full set of items unchanged for the Main Study.

Earnings (D_Q16-D_Q18)

The gross earnings of respondents were measured by way of a separate set of questions asked to salaried and self-employed respondents. Respondents who were unable or reluctant to report precise earnings were given the opportunity to report earnings in broad categories. Salaried respondents were given the choice of reporting earnings per hour, day, week, two weeks or year, and were also asked to report any annual payments they received in addition to their regular pay package. Self-employed respondents who had conducted their own business for at least a year were asked to report their gross earnings from their business in the last year, and those who had conducted their business for less than a year were asked to report their earnings for the last month. Here as well, respondents who did not report precise earnings were given the opportunity to report in broad categories.

Based on assumptions on the earnings distribution and taking into account the basis on which salaried employees reported their earnings, the answers to all these questions were combined into

an overall measure of hourly and monthly earnings, with a separate measure for salaried employees and self-employed as well as a combined measure for all respondents in paid employment. A thorough validation of earnings of the self-employed was not really feasible due to idiosyncrasies inherent in earnings from business (for example, many respondents reported zero earnings). The analyses presented here are based on earnings from salaried employment. The complex method of measuring earnings leads to some apparent measurement error for salaried workers. The causes of these problems are familiar from other research, with, for example, some respondents reporting hours worked in the last week rather than in a typical working week, but subsequently reporting typical earnings rather than the earnings corresponding to the reported hours. Because the final earnings indicator adjusts for hours worked, the resulting indicator will be flawed in cases when the reported hours deviate strongly from typical hours. For this reason we removed the top and bottom 2.5% of the distributions prior to the analyses presented here.

As anticipated, the use of broad categories as alternative to precise earnings was the exception rather than rule, but the inclusion of this option significantly reduced item nonresponse on earnings variables. After removal of extreme values, these variables all performed well in the analyses. The earnings distributions were still slightly skewed, but plausible. There were similar distributions in each country, with some variation in kurtosis and skewness. The broad categories worked very well, showing a highly similar distribution to directly reported earnings. Earnings were plausibly related to skills and to investments in training, as well as to respondents own education, and to parents' education and occupation.

The recommendation was to retain this full set of items unchanged for the Main Study.

Noncognitive skills

GRIT and locus of control (I_Q01-I_Q02)

The items under consideration here are related to three broad concepts: GRIT, self-discipline and locus of control. GRIT can be further subdivided into perseverance and consistency of effort, and locus of control into internal and external. None of these sets of variables formed a good scale, but internal locus of control achieved a Cronbach's alpha of 0.66, which is satisfactory for a scale consisting of only three items (the other alphas were: perseverance of effort, 0.53; consistency of effort, 0.54; GRIT combined scale, 0.59; self-discipline, 0.47; and external locus of control, 0.41). Although GRIT showed some relation to level of education and labor market outcomes, neither GRIT nor its subscales were convincingly related to test scores, and the bivariate relation with earnings disappeared in the multivariate analyses. For this reason, the Consortium recommended dropping all these items. Much the same applies to self-discipline, which in multivariate analyses was not related to test scores or economic outcomes. Both internal and external locus of control showed a clear bivariate relation with test scores, although only the effect of external locus of control held up in multivariate analyses. By contrast, internal locus of control showed clear effects in multivariate analyses of labor market outcomes. Closer inspection of the data revealed it was possible to develop a combined measure comprising two internal locus-of-control items and one (reversed) external locus-of-control item (representing roughly the concept of decisiveness) which performed well in multivariate analyses both on outcomes and test scores. However, the Consortium did not think this warranted keeping these items for the Main Study.

The recommendation was to drop all items related to GRIT, self-discipline and internal locus of control.

Time preference (I_Q03a-d)

This set of four items was dropped on the basis of the list of priorities provided by the OECD. The goal of analyzing this set of items was to establish whether that decision was justified, or whether strong reasons existed to reverse that decision. The analyses showed that, although this set of items formed an unreliable scale (Cronbach's alpha = 0.44), this scale performed surprisingly well in the multivariate analyses, showing among other things a strong positive relation to test scores, and for males also a clear relation with employment status and earnings. However, in the view of the Consortium, these results were not sufficient to warrant overturning the earlier decision.

The recommendation was to stand by the original decision to drop these items.

Learning strategy (I_Q04a-m)

This long set of items was intended to represent two related concepts: deep or elaborate learning and surface-rational learning. The results show we could form a reliable scale for deep or elaborate learning consisting of the following items: I_Q04b, I_Q04d, I_Q04h, I_Q04j, I_Q04l, I_Q04m (Cronbach's alpha = 0.78). We could not form a reliable scale for surface-rational learning. As the intention was to at least significantly reduce the number of these items retained for the Main Study, we therefore proposed dropping the remaining items, and evaluating the performance of the deep learning scale in multivariate analyses. These analyses showed mixed results. Importantly, however, it showed a strong positive relation with test scores, and inclusion of this indicator as a control variable resulted in significant changes in the estimated effects of education and training variables on skills. Although there was no really robust relation with other outcomes, on balance the strong relation with test scores and its impact as control variable made this, in our view, a strong candidate to be retained for the Main Study, together with internal locus of control. The recommendation was to retain the reduced set of six items for the Main Study.

Table 3.2 shows the average correlation of the zlogit proxy with scales in section I.

Table 3.2: Average correlation of section I scales with zlogit (proxy of skills)

Item	Average	SD
I_Q01_mean About Yourself - Grit and Self-Discipline	0.015	0.077
I_Q02_mean About Yourself - Locus of Control	0.148	0.087
I_Q03_mean About Yourself - Time Preference	0.219	0.070
I_Q04_mean About Yourself - Learning Strategies	0.145	0.088
I_Q06_mean About Yourself - Political Efficacy	0.211	0.059
I_Q07_mean About Yourself - Social Trust	0.093	0.080
I_Q03_mean About Yourself - Time Preference	0.219	0.070

Table 3.3 summarizes the main results of the effects of noncognitive skill scales in the multivariate analyses:

Table 3.3: Significant effects of section I scales in multivariate analyses.

Scale:	males' labor force status		females' labor force status		males	females	test scores
	unem- ployed	non- active	unem- ployed	non- active	hourly wage	hourly wage	
I_Q01 Grit, subscale perseverance of effort		nnn		nnn			
I_Q01 Grit, subscale consistency of effort							
I_Q01 Grit, combined scale		nnn		nnn			nnn
I_Q01 Self-discipline							
I_Q02 Internal Locus of Control	n	nnn		n	p	p	
I_Q02 External Locus of Control							nn
I_Q03 Time Preference	nnn						ppp
I_Q04 Learning Strategies: deep learning	ppp	p	pp				ppp

ppp/nnn: positive/negative effect significantly different from 0.0 at 1% level

pp/nn: positive/negative effect significantly different from 0.0 at 5% level

p/n: positive/negative effect significantly different from 0.0 at 10% level

Other outcomes

Civic engagement (I_Q05a-h)

In the June revision of the BQ, this set of items was replaced by a single item on voluntary work. This decision was vindicated by an initial inspection of the data, which shows a strong correlation between the two separate items in this topic in the Field Test version of the BQ. Among other things, the data showed that civic engagement is positively related to test scores and that this relationship entirely accounted for the bivariate relation between civic engagement and level of education. There was also a significant relation between civic engagement and immigrant status, labor force status and health status. We therefore believed this was a useful outcome variable that should be retained for the Main Study.

The recommendation was to retain single item on voluntary work for the Main Study.

Political efficacy (I_Q06a-d)

No reliable scale could be formed for this set of items (Cronbach's $\alpha = 0.47$). However, in consultation with representatives of the OECD, it was decided it would be valuable to retain a single item indicator in order to maintain a diversity of noneconomic outcomes. After consulting an expert on this topic, the Consortium recommended keeping the first item (I_Q06a), which was felt to best reflect the meaning of individual political efficacy.

The recommendation was to retain single item on individual political efficacy for the Main Study.

Social trust (I_Q07a-d)

Although no strong scale could be formed, the first two items (I_Q07a and I_Q07b) have worked very well in the past in other surveys and together achieve a Cronbach's α of 0.64. This reduced scale was positively related to test scores, as well as to employment status, training participation, and current participation in education. Unexpectedly, it was also negatively related to deep learning.

The recommendation was to retain reduced scale of two items on social trust for the Main Study.

Health (I_Q08, I_Q09 and I_Q10)

It was decided to drop I_Q09 from the BQ. The remaining items seemed to perform well. However, both the bivariate and multivariate analyses showed that the subjective health question (I_Q08) worked a little better than the objective question (I_Q10). Most importantly, there was a clear relation of subjective health with test scores, but no such relation with objective health. Both health indicators were related to level of education, labor force status, and training participation. On balance, taking into account the clear relation with test scores and also that it has been well validated in earlier research, we felt that the subjective health indicator was preferable to the objective indicator.

The recommendation was to retain the single item on subjective health status for the Main Study.

Summary and conclusions

The most important result that can be reported on the basis of the analyses of the Field Test data is that the Field Test BQ to a very large extent succeeded in collecting the necessary information on respondents across countries. In addition, some decisions to delete variables made during the first phase of the revision process were supported by the Field Test data. For the most part, the items that were deleted in Phase I did not perform as well in certain respects as items that were retained, for example, in terms of high item nonresponse, proportion of the population covered, or performance in data analyses.

Applying the criteria noted in the introduction to the results of the Field Test analyses, the Consortium recommended removing the following list of items for the Main Study version of the BQ:

JRA items

In line with the JRA pilot analyses, most of the items in this section performed well. However, two items did not perform as expected and were therefore recommended to be removed. Item F_Q01a appeared somewhat ambiguous, while item F_Q06a was redundant, covering much the same meaning as F_Q06b (correlation between the two was >0.7), which performed somewhat better in other respects.

Skill use at work and in everyday life

In general the skill use items performed very well. We recommended that most be retained for the Main Study. However, two numeracy items and one ICT item did not perform as expected for both work and everyday life contexts and were recommended to be removed. G_Q03a/HQ03a did not show any consistent relation to skills and did not scale well, especially in the work context (lowest item total correlation among the group of items scaled together). G_Q03e/H_Q03e was part of a redundant item pair together with G_Q03f/H_Q03f (correlation was about 0.7 in both cases). It was decided to retain just one of these two items and drop G_Q03e/H_Q03e. The same was true for G_Q05b/H_Q05b, which covered much the same meaning as item G_Q05c/HQ5c (correlation was above 0.6 in both cases).

Section I

Perseverance and self-discipline (I_Q01a-I_Q01i)

This set of items did not perform well in terms of predictive power (average correlation with “test scores” was 0.015 across countries) and was at least partially redundant with respect to the concept of deep learning strategies.

Surface learning (I_Q04a, I_Q04c, I_Q04e, I_Q04f, I_Q04g, I_Q04i, I_Q04k)

This set of items showed poor scaling properties. However, deep learning formed a good scale and performed better in multivariate analyses.

Political efficacy (I_Q06b-I_Q06d)

This set of items showed poor scaling properties. However, in the interest of retaining a selection of noneconomic outcomes, it was recommended that the first of these items (I_Q06a) be retained because it was considered to be the most appropriate indicator of this variable among the four (highest average item total 0.307 among the four items, and explained $>5\%$ of variance of the “test score” proxy).

Social trust (I_Q07c, I_Q07d)

The first two items of the intended four-item scale performed reasonably well in terms of scaling properties and was recommended for retention. These two items are also the well-researched/established way of measuring social trust. The second two items did not perform as well and were recommended to be dropped from the Main Study instrument.

Disability (I_Q10a-I_Q10b)

We have a subjective overall health measure that performs well, showing a strong relation with “test scores,” among other variables. The specific disability-related measures did not perform as well in comparison (no significant relation with “test scores”) and were recommended to be dropped. In addition, the distribution of responses differed substantially across countries for these variables. These two items also were more time consuming than expected.

Intensity of formal education (items B_Q06-B_Q09a,b)

Intensity of last activity (B_Q17-B_Q20a,b)

There were some measurement problems with these items, and, in particular, with the summary measure for total time spent on formal education as well as nonformal learning activities based on these items. As pointed out earlier, Field Test results indicated that some respondents may have had difficulties in judging the time spent on their formal education or learning activities. In the case of nonformal learning, the question that related to the number of activities a respondent engaged in performed substantially better in multivariate analyses (for example, intensity of training activities explains no additional variance in “test scores” after a simple dummy (training yes/no) has been included). In addition to the measurement considerations, the administration time for these items was excessively long (1.5 minutes in the case of formal education, and an estimated three minutes in the case of nonformal learning for those who take these questions).

Even though these variables have analytical importance, the Consortium proposed to drop them.

- Mother's or female guardian's occupation (J_Q06c-J_Q06e)
- Father's or male guardian's occupation (J_Q07c-J_Q07e)

There was evidence that the question on mother's and father's education performed better in terms of predictive power in multivariate analysis. The occupation variable did not provide substantial incremental predictive power compared to the education variable in these analyses. In addition, the items were quite time consuming (1.5 minutes) and require human coding of responses compared to the education variables. The Consortium therefore proposed to drop these items.

The recommended revisions led to the required reduction in time for the Main Study BQ of some 10 minutes compared to the interim BQ. The expected average interview time for the international core BQ was therefore under 40 minutes – not including any national extensions.

3.5 The content of the Main Study BQ

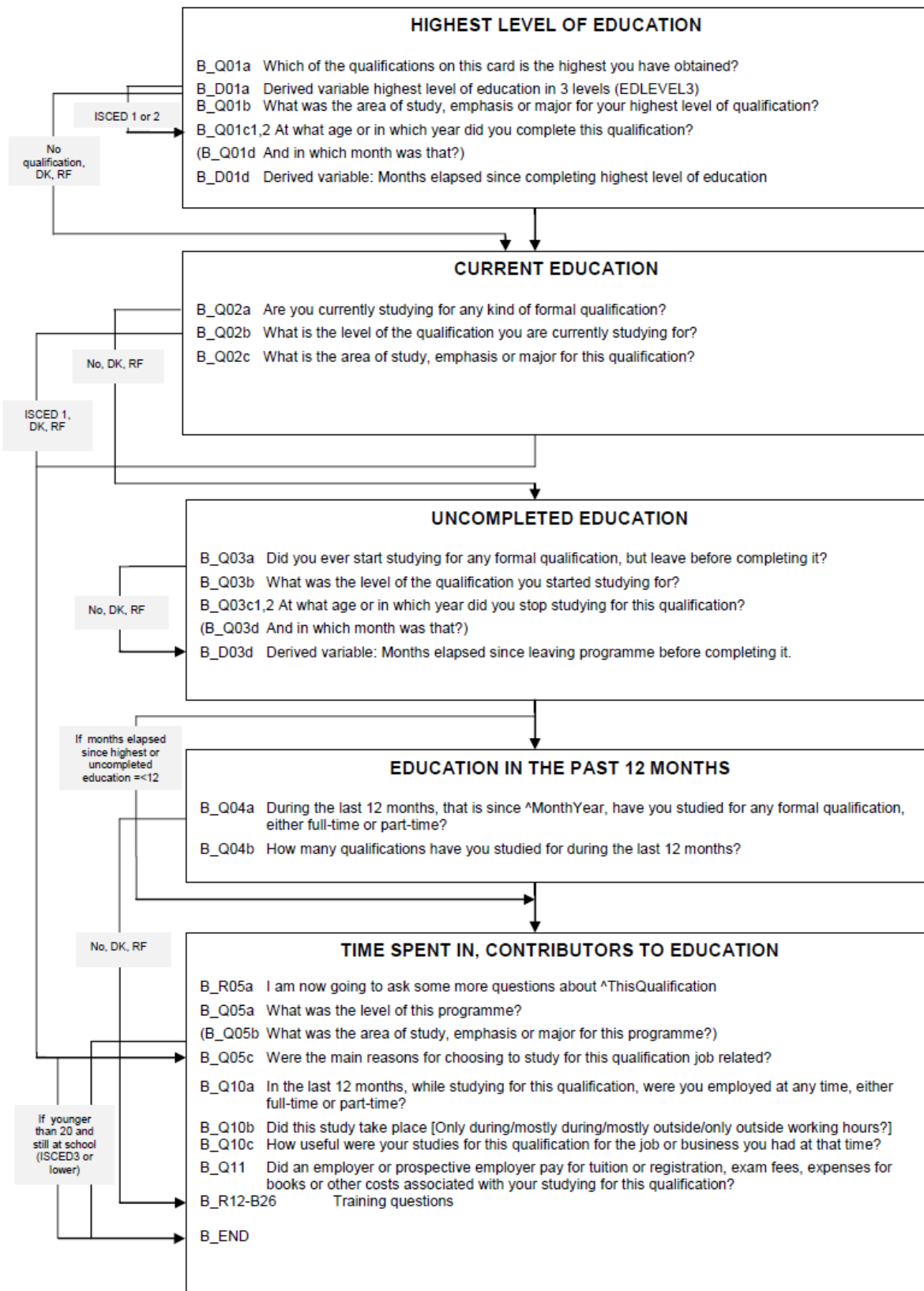
3.5.1 BQ Main Study

As indicated above, based on the analyses of the Field Test data, a final BQ for the Main Study was developed. The basic structure of the Main Study BQ is relatively straightforward, although some sections involve somewhat complex routing depending on, among other things, the educational and labor market status of the respondent. The BQ consists of a total of 10 sections:

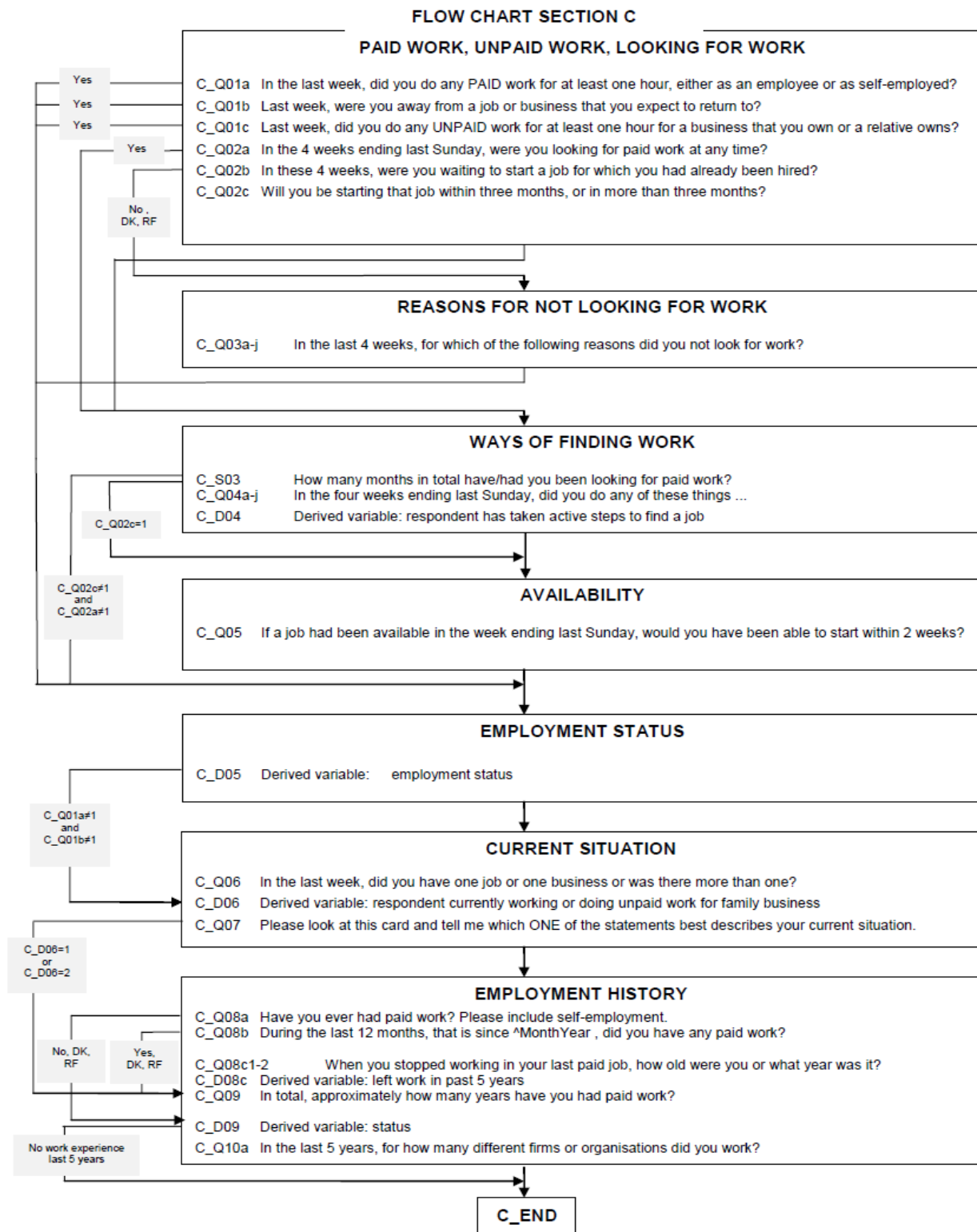
- A. General information (date of birth, gender: all respondents)
- B. Education and training (all respondents)
- C. Current status and work history (all respondents)
- D. Current work (for those currently employed or self-employed)
- E. Last job (for those not currently employed or self-employed, who have worked in last five years)
- F. Skills used at work (JRA Module; for those currently employed or employed in the last 12 months)
- G. Skill use literacy, numeracy and ICT at work (for those currently employed or employed in the last 12 months)
- H. Skill use literacy, numeracy and ICT in everyday life (all respondents)
- I. About yourself (learning strategies, voluntary work, social trust, health: all respondents)
- J. Background information (household composition, migration status, languages, parental education, cultural capital parental home: all respondents)

Sections B and C contain relatively complex routing. The following flow chart indicates the routing for section B.

FLOW CHART SECTION B



The following flow chart indicates the routing in section C.



3.5.2 National extensions

All countries were allowed limited scope to include national extensions they required for their own policy purposes. In order to avoid undue burden on respondents that could negatively affect the data quality, a strict rule was imposed that the total additional time added to the questionnaire in the form of such extensions was not allowed to exceed five minutes. The time estimates used to enforce this restriction took into account the number and type of proposed items to be added by a country.

3.5.3 National adaptations

The major adaptations countries were required to perform in the Main Study BQ were the following:

- Levels of education [highest, current, uncompleted and (other) recent education]: For obvious reasons, it was not feasible to use a standard international classification in order to ascertain the level of education a respondent is currently following or has followed in the past. In order to be comprehensible to respondents, all questions pertaining to level of education needed to be framed in terms of the qualifications currently or formerly available in the country concerned. Countries were required to develop an individual list of qualifications that could be directly matched to the standard list in the Master BQ to the extent that national equivalents for the levels described therein exist or have existed in the past. Countries were required to supply a full conversion scheme from their national levels to the international ISCED levels included in the Master BQ, including a specification of nominal years of schooling corresponding to each level and orientation and, where relevant, the vocational or academic nature of the program. A separate Excel sheet is provided with an overview of the national qualifications used in PIAAC with their conversion into ISCED level and orientation, nominal years of schooling, and vocational/academic.
- Country and language lists: Several questions in the BQ referred to countries or languages. These questions have a two-stage structure – first, a closed list comprising a limited number of countries/languages that are considered most relevant in the country concerned, and second, an open question for those respondents who wished to report a country/language not included in the standard list. Because the relevant countries and languages differ strongly from country to country, each country was required to adapt these items to national needs.
- In section C a block of questions was included that was designed to capture the respondent's job search behavior. Because search channels can differ subtly among countries, countries were asked to inspect the standard list of questions and, if necessary, adapt or add items to correspond to the national institutions and so on that may be involved.
- In sections D and E, several questions were used to ascertain the (last) occupation and (last) economic sector in which the respondent works or had worked in the past. Countries were required to check and, if necessary, adapt these items to the national setting.

- All income questions were asked in two forms: First, respondents were asked to report their income directly in the national currency. For respondents who were unwilling or unable to report directly their precise earnings, the option was made available to report in broad ranges. Countries were required to adapt the amounts and the currency used in these broad ranges based on explicit instructions how these should be derived from national statistics on recent population earnings distributions.
- For several questions throughout the BQ, countries were required to check, and if necessary adapt, the wording of questions to correctly reflect the national setting.
- Wherever adaptations involved some kind of structural change to the BQ – for example, splitting of a single item into multiple items, or the addition or deletion of one or more response categories – countries were required to make any necessary adaptations to routings, derived variables and so forth that make reference to the original items.

3.6 Quality check in the BQs

The BQ contained a number of features designed to assist the interviewers and ensure it was administered in a standardized way across all countries. These features were:

- Instructions given to interviewers. These instructions were designed to provide the interviewer with any relevant information that might be needed in order to pose the question in the correct manner, to indicate when to hand over and take back show cards, to provide support to respondents, and so on.
- Help buttons. In addition to these interviewer instructions, which were always visible to the interviewer but not read out to the respondent, the BQ contained a number of help buttons that the interviewer could consult if needed. These contained such things as additional information that could be provided to the respondent if needed, additional background information on the meaning or intent of questions, and so on.
- Consistency checks. For some items, there were consistency checks built in to the BQ that were triggered when a respondent gave a numeric answer to a question that might be considered to fall outside a plausible range of values. Examples include the age at which a respondent has reported a given event or status or the earnings reported by the respondent.

References

- Adam, D., Bay, C., Bonsang, E., Germain, S., & Perelman S. (2006). *Occupational activities and cognitive reserve: A frontier approach applied to the survey on health, ageing, and retirement in Europe (SHARE)* (CREPP Report No. DP 2006-05). Retrieved from <http://hdl.handle.net/2268/72515>
- Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, 24, 155-173.
- Arrow, K. J. (1973). Higher education as a filter. *Journal of political economy*, 2(33), 193-216.

- Arthur, W. J., Bennett, W. J., Stanush, P. L., & McNelly T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance*, 11(1), 57-101.
- Bassanini, A., Booth, A. L. Booth, Brunello, G., De Paola, M., & Leuven, E. (2005). *Workplace Training in Europe*. (IZA Discussion Paper No. 1640). Retrieved from IZA website: <http://ftp.iza.org/dp1640.pdf>
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71, 287–311.
- Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis with special reference to education*. New York: Columbia University Press.
- Bills, D. (2003). Credentials, signals and screens: Explaining the relationship between schooling and job assignment. *Review of Educational Research*, 73, 441-70.
- Binkley, M. R., Sternberg, R., Jones, S., Nohara, D., Murray, T. S. & Clermont Y. (2003). Moving towards measurement: The overarching conceptual framework for the ALL study. In T. S. Murray, Y. Clermont, & M. Binkley (Eds.), *Measuring adult literacy and life skills: New frameworks for assessment*. Ottawa: Statistics Canada.
- Borghans, L. A., Duckworth, L., Heckman, J. J. & Ter Weel, B. (2007). *The economics and psychology of cognitive and non-cognitive traits* (working paper). Chicago: University of Chicago.
- Borghans, L. A., Golsteyn, B., de Grip, A. (2006). *Meer werken is meer leren: Determinanten van kennisontwikkeling*. [English translation: Working more is learning more: determinants of knowledge development] 's-Hertogenbosch: CINOP.
- Boudon, R. (1974). *Education, opportunity and social inequality*. New York: John Wiley & Sons.
- Bourdieu, P. (1984). *Distinction: A social critique of the judgment of taste*. London: Routledge & Kegan Paul.
- Brunello, G. (2004). On the complementarity between education and training in Europe. In D. Checchi and C. Lucifora (Eds.), *Education, training and labour market policies in Europe*. Retrieved from <http://www.palgraveconnect.com/pc/doi/10.1057/9780230522657>
- Collins, R. (1979). *The credential society: An historical sociology of education and stratification*. New York: Academic Press.
- De Corte, E. (1990). Towards powerful learning environments for the acquisition of problem-solving skills. *European Journal of Psychology of Education*, 5, 5-19.
- De Grip, A., Bosma, A. H., Willems, D., & van Boxtel, M. (in press). Job-worker mismatch and cognitive decline. *Oxford Economic Papers*.
- De Grip, A., & Van Loo, J. (2002). The economics of skills obsolescence: A review. In A. de Grip, J. van Loo and K. Mayhew (Eds.), *The economics of skills obsolescence, research in labour economics*, Vol. 21, (pp. 1-26) Amsterdam/Boston: JAI Press.

- Dickerson, A., & Green, F. (2004). The growth and valuation of computing and other generic skills. *Oxford Economic Papers-New Series*, 56(3), 371-406.
- Glaser, R. (1991). The maturing of the relationship between science of learning and cognition and educational practice. *Learning and Instruction*, 1, 129-144.
- Gruber, J., & Wise, D. (2004). *Social security programs and retirement around the world*. Chicago: University of Chicago Press.
- Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labour market outcomes and social behavior. *Journal Of Labour Economics*, 24(3), 411-482.
- Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.
- Müller, W., & Shavit, Y. (1998). The institutional embeddedness of the stratification process: A comparative study of qualifications and occupations in thirteen countries, In Y. Shavit & W. Müller, *From school to work. A comparative study of educational qualifications and occupational destinations*. Oxford: Clarendon Press.
- Prüfer, P., & Rexroth, M. (2005). Kognitive Interviews. [Cognitive Interviewing]. *ZUMA How-To-Reihe*, 15. Mannheim: ZUMA.
- Rychen, D. S., & Salganik, L. H. (Eds.) (2003). *Key competencies for a successful life and a well-functioning society*. Göttingen, Germany: Hogrefe & Huber.
- Schooler, C., Mulatu, M. S., & Oates, G. (1999). The continuing effects of substantively complex work on the intellectual functioning of older workers. *Psychology and Aging*, 14, 483–506.
- Schuller, T., & Desjardins, R. (2007). *Understanding the social outcomes of learning*. Paris: Organisation for Economic Co-operation and Development.
- Schultz, T. (1963). *The economic value of education*. New York: Columbia University Press.
- Spence, M. (1973). Job market signalling, *Quarterly Journal of Economics*, 87(1), 355-374.
- Thurow, L. C. (1975). *Generating inequality*. New York: Basic Books.
- Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45-65). Göttingen, Germany: Hogrefe & Huber.
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Chapter 4: Translation, Adaptation, and Verification of Test and Survey Materials

Andrea Ferrari and Laura Wayrynen, cApStAn; Dorothee Behr and Anouk Zabal, GESIS

4.1 Overview

The PIAAC assessment instruments (comprising cognitive instruments and the BQ) were originally developed in English, but administered to sampled adults in their own language. It follows that the successful *localization* of assessment instruments is an important component of the project. Some definitions, first, of terms which we shall use throughout this chapter:

- *Localization* can be defined, in general terms, as the process of adapting a product or service to a particular language and culture. A successfully localized product or service is one that appears to have been developed within the local culture. For international comparative assessment studies, like PIAAC, the challenge is to localize test and questionnaire items while maintaining the comparability of collected assessment results and contextual data across countries and languages.
- The *localization process* can be broken down into *translation/adaptation* and *validation*. The words *translation* and *adaptation* are used jointly because the term translation is deemed too restrictive to describe the process of culturally adjusting a test rather than literally translating it. An adaptation may consist in changing the picture of a stimulus, in changing the combination of July/summer to July/winter (or January/summer) for the Southern hemisphere, in changing a coeducational school context to a boys' or girls' school context for certain countries, etc. It may, e.g., involve a change of wording, register, context, currency, measurement unit, or form of address. *Validation* refers to quality control steps which will be defined later.

In PIAAC, as in many major international assessment studies, the localization process followed a mostly decentralized model:

- The participating countries (National Centers) were each responsible for localizing assessment materials for use in their respective countries.
- The PIAAC Consortium guided and assisted the countries throughout the process, in particular by developing and conducting linguistic quality assurance (LQA) and linguistic quality control (LQC) processes.

In PIAAC, the LQA processes implemented by cApStAn in cooperation with other Consortium players included:

- Early resolution of potential localization issues, via preliminary scrutiny of source assessment materials to anticipate adaptation issues, ambiguities, cultural issues, or item translatability problems.
- Definition of the localization design, based on the OECD PISA (Programme for International Student Assessment) design. The minimum standards to be followed by countries included a double translation and reconciliation design, making use of professional staff, and attending the training sessions organized by the Consortium. The key quality-control steps included in the design were the verification of National Centers' initial submissions by verifiers appointed, trained and monitored by cApStAn, a final check of instruments after post-verification revision by National Centers, and layout corrections by Consortium technical staff, and the documentation of all steps leading to the finalized localized instruments.
- Preparation of general translation and adaptation guidelines, separately for the BQ and the assessment materials. These key documents set out requirements and roles, translation traps, pointers on linguistic difficulty, psychometric traps, cultural adaptations, etc. They are further described in sections 4.3.1 and 4.3.2.
- Preparation of centralized tools for documenting and monitoring the successive translation, adaptation and verification activities: the VFFs (Verification Follow-up Forms) and BQAS (Background Questionnaire Adaptation Spreadsheets) used in the Field Test and later the MMFs (Main Study Translation-Adaptation-Verification Monitoring Forms). These tools included detailed item-specific translation and adaptation guidelines such as advice on adaptations that were mandatory, desirable or ruled out; advice on terminology problems and idiomatic expressions, literal or synonymous matches, that is, between stimuli and items to be echoed, patterns in response options to be echoed, formatting issues, and so on. Figure 4.1 shows an example of a VFF with item-specific guidelines

Figure 4.1: Example of a VFF

PIAAC FIELD TRIAL 2009 VERIFICATION FOLLOW-UP FORM COMPUTER-BASED													
Country: PT Target language: pt		UNIT: Election Results	PIAAC ID: C302BC02	ALL ID: COREQ2S1									
PLEASE INSERT NEW LINES, IF NEEDED, TO DOCUMENT ADDITIONAL ISSUES													
LOCATION	ENGLISH SOURCE	PROPOSED TARGET VERSION	CONSORTIUM RECOMMENDATION	NPM COMMENT	VI INTE								
stimulus	Nationwide Manufacturing Company Union Council ELECTION RESULTS		Note: 'Union' is to be understood as trade union, i.e. an organization representing workers										
stimulus	Posting Date: June 22, 2000		Eliminate ', 2000' versus ALL version										
stimulus	The election of a new member of the Union Council for election group 3, at the Carver plant took place on June 21, 2000.		The name 'Carver' may be changed. Note: 'plant' means here 'factory' Eliminate ', 2000' versus ALL version.										
stimulus	The results of the election were as follows:												
stimulus	<table border="1"> <thead> <tr> <th>Candidates</th> <th>Number of votes</th> </tr> </thead> <tbody> <tr> <td>A. Greer</td> <td>120 votes</td> </tr> <tr> <td>H.A. Holliday</td> <td>80 votes</td> </tr> <tr> <td>G. F. Reynolds</td> <td>29 votes</td> </tr> </tbody> </table>	Candidates	Number of votes	A. Greer	120 votes	H.A. Holliday	80 votes	G. F. Reynolds	29 votes		Names of people may be changed. Keep the three numbers aligned over each other.		
Candidates	Number of votes												
A. Greer	120 votes												
H.A. Holliday	80 votes												
G. F. Reynolds	29 votes												
stimulus	Consequently Mr. A. Greer was formally elected as a member of the Union Council for Nationwide Manufacturing Company.		If name 'A. Greer' is changed, change it here too										
stimulus	In accordance with article 16, paragraph 1 of the Union Council bylaws, any interested party may lodge a complaint with the council within one week after publication of these results.												

- Provision of training sessions for countries' translation teams or their trainers of translations. A general session was provided at a meeting in Lisbon, Portugal, in October 2008, and modular workshops (for the various types of materials) were provided at a Barcelona, Spain, meeting in March 2009.
- Provision of a translation training kit so that further training sessions could be held in countries. The kit included a customizable PowerPoint presentation, materials for hands-on exercises, confidentiality forms, and so on.
- Continued assistance to National Centers throughout the localization process (help desk via ticketing system, see Chapter 6).

In PIAAC, the implemented LQC processes included:

- Verification by the Consortium of target versions submitted by National Centers against the source versions, with reporting of residual errors and undocumented deviations, and expert advice where corrective action was needed:
 - For Field Test instruments: full verification of all national materials
 - For Main Study instruments: “focused” verification of changes made by countries to their finalized Field Test national materials (whether to echo changes made to the source version or at the initiative of the National Centers), extra checks for risky cases as needed, and full verification of newly translated materials

- A final check procedure after National Centers carried out their post-verification revision of instruments and Consortium technical staff made layout corrections, again with reporting and follow-up of residual errors and/or unresolved issues.
- The scope of verification included all translated instruments viewed by respondents (computer-administered test units, help and orientations, BQ, paper test booklets) as well as language-dependent automated scoring rules (for the “highlight in stimulus” response mode and numeric entry response mode), paper scoring guides, and the “CAPI workflow” file used by interviewers to conduct the questionnaire and assessment sessions.

4.2 Participation in the development of the source version

Early resolution of potential localization issues via preliminary scrutiny of source assessment materials is an upstream LQA process which aims to reduce the difficulties and workload encountered later downstream. cApStAn reviewed the first drafts of new cognitive materials and of the BQ (as of version 3.4) with an eye to anticipating adaptation issues, ambiguities, cultural issues, or item translatability problems, with suggestions for either rewording or adding item-specific translation/adaptation guidelines.

cApStAn also provided English translations of item submissions from participating countries in Japanese, Italian, German and French; some of these were selected to be part of the PIAAC item pool.

Throughout the localization process, cApStAn took care of an errata management process, whereby errors in the source identified by National Centers or verifiers were tracked and, depending on the nature of the error and the time of discovery, listed for correction in source and/or national versions either at Field Test or Main Study phase.

4.3 Testing languages and translation/adaptation procedures, including double translation design

4.3.1 Testing languages and translation/adaptation procedures for the BQs

The major bulk of translation occurred in preparation for the Field Test. Therefore, the focus in the following will be on translation activities prior to the Field Test rather than in preparation of the Main Study.

The BQ was translated/adapted from (international) English into 39 national versions comprising 26 languages including English. Table 4.1 displays the languages of the BQ for each country.

Table 4.1: Languages of BQ for each country

Country	Languages
Australia	English
Austria	German, Turkish, Serbo-Croatian
Canada	English, French
Chile ¹	Spanish
Cyprus ²	Greek
Czech Republic	Czech
Denmark	Danish
England/N. Ireland (UK)	English
Estonia	Estonian, Russian
Finland	Finnish, Swedish
Flanders (Belgium)	Dutch
France	French
Germany	German
Hungary ³	Hungarian
Ireland	English
Italy	Italian
Japan	Japanese
Korea, Rep. of	Korean
Netherlands	Dutch
Norway	Norwegian (BM), English
Poland	Polish
Portugal ⁴	Portuguese
Russian Fed. ⁵	Russian
Slovak Rep.	Slovak, Hungarian
Spain	Spanish, Catalan, Galician, Valencian, Basque
Sweden	Swedish
United States	English, Spanish

¹ Chile later dropped out of this cycle of PIAAC and joined PIAAC Round 2.

² Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

³ Hungary later dropped out of PIAAC.

⁴ Portugal later dropped out of PIAAC.

⁵ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Austria, Norway and the United States translated the BQ into more languages than they did for the assessment instruments. This was to accommodate important non-English speaking populations.

Prior to the BQ translation, each country, in cooperation with the Consortium, adapted the international BQ version to its local context. Adaptations at this stage mainly pertained to questions that, although measuring the same underlying concept, were in themselves substantively different from country to country (e.g., education, occupation or income items). Furthermore, countries were offered the opportunity to add items of country-specific interest not yet included in the international BQ. All such adaptations and national extensions were subject to approval by the Consortium. Chapter 3 describes the process of adaptation and extension in detail.

Once the adaptations and national extensions received signoff from the Consortium, a country-specific BQ version was built for each country, consisting of the (adapted) common set of international BQ items and the country-specific items. This version served as the basis for translation. The translation environments for the BQ translation were the Item Management Portal as well as specific translation software. Chapter 6 describes the technical tools.

Because comparability of survey materials is essential to any meaningful use of cross-national survey data, countries received a general guideline document laying down a quality framework for translation. The guideline document for the BQ translation focused on the one hand on the general translation process and on the other hand on issues to consider in the actual translation.

The guidelines on the general translation process included the recommended translation approach of double translation by two independent translators, followed by reconciliation. Double translation allows the spotting of misinterpretations or ambiguities, idiosyncratic wording or simply translator oversights; moreover, it offers stylistic variants among which to choose in light of a fluent translation. It has established itself as a state-of-the art approach in questionnaire translation. For reconciliation, team reconciliation was proposed to countries as a very efficient reconciliation method. Team reconciliation brings together at one table a unique mix of competencies: translators and linguistic experts, experts in the various domains of the questionnaire (education, work, etc.) as well as experts in questionnaire design and survey methodology. This broad range of expertise (translation, domain, design) is regarded as essential for producing high-quality questionnaire translations (Harkness, 2003; Harkness, Villar, & Edwards 2010). Alternatively, as a minimum, a single reconciler was required, ideally with input from a panel of experts in survey methodology and the various domains covered by the BQ. Translators were to be skilled practitioners, translating into their mother tongue and experienced or trained in questionnaire translation. Reconcilers were to have strong language skills in both source and target languages and be knowledgeable about questionnaire translation, questionnaire design, and the content domains covered by the BQ.

The general BQ guidelines also specified an overall framework for the BQ translation. The fact that a number of items in the BQ had been taken (changed or unchanged) from other surveys was acknowledged. Countries were given freedom to consult already existing translations from these surveys. However, it was stressed that in the end, the adherence and comparability to the PIAAC BQ was the crucial factor and would be the basis for verification.

The guidelines on issues to consider during translation specified that countries were to produce a questionnaire translation that maintains the measurement properties and the meaning of the source questionnaire, while at the same time being as fluent and understandable as possible. The overall task was to strike the right balance between faithfulness and fluency. The general message to countries was to produce the best possible *translation*. Any adaptations – beyond those already been agreed on – that countries deemed necessary had to be documented by countries and submitted to the Consortium for approval. Adaptations in this case were understood as intended deviations from the source version going beyond the changes that typically occur through translation. While the adaptations occurring *prior* to the translation phase applied to all countries in the same manner (e.g., all countries had to implement their own education measures), adaptations *during* the translation phase, if occurring at all, affected individual countries only. Countries were provided with an Excel tool in which to document adaptation needs: They were asked to provide an explanatory back translation of their chosen translation into English and to justify their decision. Back-translation in PIAAC was thus seen as a tool enabling communication with the Consortium and allowing for a commonly understood documentation. It did not serve as an assessment tool in itself.

Furthermore, countries were given item-specific translation guidelines. These provided further clarifications (e.g., on the meaning of terms or phrases or on characteristics of response categories) for a certain number of questionnaire items. The need for these clarifications had been identified by expert reviews focusing on potential translation problems and results of the cognitive pre-test. Furthermore, a so-called advance translation had been conducted on a pre-final version of the BQ (cf. Dorer, 2012). The goal of this translation was to identify problems in the source questionnaire while it was still under development and to take appropriate action (e.g., adding item-specific guidelines, changing wording).

During a one-day workshop at the NPM meeting in Barcelona in March 2009, NPMs, national staff responsible for the translation process, or translators themselves were introduced to the specificities of the translation of the BQ. The workshop covered the technical environment of the questionnaire translation (Item Management Portal, translation software), the different types of BQ translation guidelines, as well as good translation practice and discussion. The national teams were encouraged to replicate (parts of) the workshop in their countries with their chosen personnel.

During the translation and reconciliation process itself, countries were given the opportunity to ask queries about any problems they encountered (regarding meaning, technical issues, etc.). These queries were submitted by countries within the Open Ticket Request System (OTRS); GESIS monitored and answered the BQ questions and liaised with other Consortium partners as needed – in particular ROA as item developer of the BQ.

After reconciliation, the BQ translations were submitted by the countries to the Consortium along with any documentation on special translation decisions and desired adaptations. The BQ translations (and adaptations) underwent the same verification procedures as the assessment materials. Subsequent parts of this chapter present the verification process.

After the Field Test, countries had the opportunity to correct translation errors that had come to their attention in the course of fieldwork or their own analyses. Furthermore, the Consortium provided each country with a PIAAC Field Test Report which included recommendations to

check certain specific items (where applicable). However, modifications to the questionnaire were required to be restricted to those that were absolutely necessary, i.e. to correct *errors*, but not make any changes such as stylistic improvements which could otherwise affect item functioning for items which had proved to work well in the Field Test.

4.3.2 Testing languages and translation/adaptation procedures for the cognitive instruments

The cognitive instruments were translated/adapted from the international English source version into 35 national versions comprising 24 languages, as shown in Table below, which includes information on participation in the two core components of literacy and numeracy as well as the two optional components of problem solving in technology-rich environments (PSTRE) and reading components. (Note that some countries translated the BQ into more languages than they did for the assessment instruments; this information is given in the previous section.)

Table 4.2: Translation by country for cognitive instruments

Country	Languages	Literacy/Numeracy	PSTRE	Reading Components
Australia	English	Yes	Yes	Yes
Austria	German	Yes	Yes	Yes
Canada	English, French	Yes	Yes	Yes
Chile ⁶	Spanish	Yes	Yes	Yes
Cyprus ⁷	Greek	Yes	NA	Yes
Czech Republic	Czech	Yes	Yes	Yes
Denmark	Danish	Yes	Yes	Yes
England/N. Ireland (UK)	English	Yes	Yes	Yes
Estonia	Estonian, Russian	Yes	Yes	Yes
Finland	Finnish, Swedish	Yes	Yes	NA
Flanders (Belgium)	Dutch	Yes	Yes	Yes
France	French	Yes	NA	NA
Germany	German	Yes	Yes	Yes
Hungary ⁸	Hungarian	Yes	Yes	Yes
Ireland ⁹	English	Yes	Yes	Yes
Italy	Italian	Yes	NA	Yes
Japan	Japanese	Yes	Yes	NA
Korea, Rep. of	Korean	Yes	Yes	Yes

⁶ Chile later dropped out of this cycle of PIAAC and joined PIAAC Round 2.

⁷ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

⁸ Hungary later dropped out of PIAAC.

⁹ Ireland joined late but was able to borrow and adapt the UK English version

Country	Languages	Literacy/Numeracy	PSTRE	Reading Components
Netherlands	Dutch	Yes	Yes	Yes
Norway	Norwegian (BM)	Yes	Yes	Yes
Poland	Polish	Yes	Yes	Yes
Portugal ¹⁰	Portuguese	Yes	Yes	Yes
Russian Fed. ¹¹	Russian	Yes	Yes	NA
Slovak Rep.	Slovak, Hungarian	Yes	Yes	Yes
Spain	Spanish, Catalan, Galician, Valencian, Basque	Yes	NA	Yes
Sweden	Swedish	Yes	Yes	Yes
United States	English	Yes	Yes	Yes

The translation environment for the cognitive instruments was the same as for the BQ translation: the OLT (Open Language Tool) translation software used for XLIFF files exchanged via the PIAAC Item Management Portal, described in detail in Chapter 6. XLIFF is the abbreviation of XML Localization Interchange File Format – a standard file format which permits making adaptable data editable and manageable within a localization process.

The National Centers were instructed on the principles and mechanics of translation/adaptation of PIAAC cognitive materials at the Lisbon NPM Meeting in October 2008, shortly before the release of the first cognitive materials (the literacy and numeracy link units). They received a general guidelines document prepared jointly by ETS and cApStAn, and attended an interactive training session on translation/adaptation procedures prepared jointly by DIPF and cApStAn. The training module included a detailed script, PowerPoint presentations, user manuals, various background and sample materials, and a hands-on session. It was shortly thereafter packaged and distributed as a “kit” so countries could replicate translation training locally.

Similarly to the general guidelines document for the BQ, its counterpart for cognitive materials stressed the need for very high quality translation in order to collect internationally comparable data – with the additional challenge, for cognitive materials, to “retain the *cognitive equivalence of tasks* as much as possible.”

The general guidelines included the recommended procedure of double translation by two independent translators, followed by reconciliation by a third person. The team reconciliation approach (more suitable for questionnaires) was not advocated, but a review of the reconciled version by national domain experts was recommended as an additional quality-enhancing procedure.

¹⁰ Portugal later dropped out of PIAAC.

¹¹ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

The general guidelines laid down requirements for translators and reconcilers, addressed security/confidentiality aspects, translation traps, the general principles for cultural adaptations (and detailed instructions for the adaptation of currency items), and explained the LQC processes that would follow the initial submission by National Centers of translated materials. It was explained that some PIAAC items have been taken from previous surveys (ALL and IALS) but with changes to accommodate the computer delivery mode. Countries were thus encouraged to use their existing national versions of these items as a basis for their translations, which would nevertheless be verified for equivalence to the PIAAC source version. It was also explained that countries would need to use specific software to enable the automated scoring of items using the “highlight in stimulus” response mode (Chapter 2).

Countries were also given item-specific translation guidelines (also referred to as “translation and adaptation rules” or “item-by-item notes”), conveniently echoed in the VFFs – the forms used to document the translation/adaptation and verification process. These guidelines were intended to draw the translators’ attention to possible terminology problems, translation traps, and issues for which adaptations were recommended, desirable, or ruled out.

At the NPM Meeting in Barcelona in March 2009, the National Centers were given workshops on the specificities of translating literacy units, numeracy units, problem-solving units, and reading components. The focus of these workshops 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 the construct plays 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.

Throughout the localization process (from initial double translations to reconciliation, then post-verification review, layout adaptation and final check), the National Centers were assisted via the OTRS ticketing system. Queries were routed to cApStAn, ETS or DIPF as appropriate.

As for the BQ, countries had the opportunity after the Field Test to correct translation errors that had come to their attention. Furthermore, the Consortium provided each country with feedback based on the Field Test data that included recommendations to check certain specific items (where applicable). As for the BQ, modifications to the cognitive items at the initiative of countries were required to be restricted to those that were absolutely necessary, that is, to correct *errors*, and to avoid “cosmetic” changes, carrying the risk of negative impact on item functioning for items which had proved to work well in the Field Test.

4.4 International verification of the national versions – Field Test

4.4.1 Assignment specification, verifier training

The following was the key “mission statement” for successful localization taken from the PIAAC Translation and Adaptation Guidelines:

In order to collect internationally comparable data in the study, the equivalence of all national versions is an essential requirement, which means that the translation of materials must be of extremely high quality in each of the national versions used by

participating countries. Within the assessment context, an additional goal is to retain the *cognitive equivalence* of tasks as much as possible, so that each item examines the same skills and invokes the same cognitive processes as the original version, while being culturally appropriate within the target country.

Essentially, verification is the LQC process put in place to check to what extent National Centers were successful in accomplishing the above objective, and correcting course as needed. Thus the verifiers' mission statement was to:

- Ensure linguistic correctness and cross-country equivalence of the different language versions of the PIAAC instruments
- Achieve the best possible balance between faithfulness to source and fluency in target
- Document interventions for both National Centers and the Consortium

The verifiers were selected from cApStAn's experienced team: They are native speakers of each of the target languages, highly proficient in English as source language and as 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 achievement tests. 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 attended a two-day training seminar in Krakow, Poland, in April 2009, organized by cApStAn with the participation of DIPF staff (or the follow-up session organized in Brussels, Belgium, in May 2009). They were instructed about the PIAAC Item Management Portal, the OLT software, the particularities of the different instruments to be verified (BQ, literacy units, numeracy units, problem-solving units, reading components), the subtleties of verifying scoring definitions for the highlight in stimulus response mode and the numeric entry response mode. The training seminar included presentations and hands-on exercises.

4.4.2 Overview of verification procedures

The National Centers submitted reconciled XLIFF files (or Word files in the case of paper-based instruments) for verification via the Item Management Portal, together with the appropriate filled-in monitoring instruments (VFF for cognitive units, BQAS and Dynamic Text Rules Spreadsheet for the BQ, or DTRS).

Verifiers were instructed to compare each sentence of the target version of the instruments with the corresponding sentence in the English source version, and:

- a) Examine whether the content of the items was equivalent across the two languages, with only appropriate and needed adaptations (for cognitive materials, this involved checking compliance with each item-specific guideline listed in the VFF).

- b) Examine whether the target language was linguistically correct and struck the right balance between faithfulness to source and fluency in the target language.
- c) When necessary, propose corrective action in the target language and document these interventions, in English, in the monitoring instrument. Documentation involved selecting an intervention category to identify the type of issue, selecting a severity code, and writing an explanatory comment (see below for details).
- d) Verifiers also checked and intervened as needed on scoring definitions proposed by countries for the highlight and numeric entry response modes (in cognitive units) and on dynamic text issues (in the BQ).
- e) Verifiers also checked national versions against the latest PIAAC errata list, maintained and regularly updated by cApStAn.

During the verification process, the need became apparent to refine the policy regarding the range of acceptable responses in numeracy items (for both the “exact match” and “number match” methods, see Chapter 5) In collaboration among the Numeracy Expert Group, DIPF and cApStAn, tables were prepared per country (or per group of countries sharing similar characteristics as regards, e.g., currency) in which the acceptable correct responses were listed for each item.

Likewise, during the verification process, a workflow was set up for error and exception management: corrupt file management, special requests by countries concerning units under verification or after final check, late submissions, upload of erroneous or incomplete files by countries, and so on. In hindsight, many of the problems were traced to the highlight response mode – a novelty in PIAAC. The presence of numerous and complex scoring definition “tags” in the XLIFFs made the files with highlight items more difficult to verify and subject to corruption. Furthermore, the workflow for scoring definition was not optimal, requiring too many steps: a) initial definition of scoring-related textblocks by country, b) then verification by cApStAn, c) then re-definition and re-verification in case of post-verification changes made by country and/or changes made at layout adaptation phase or at final check phase.

Verifiers were monitored and assisted by cApStAn staff, who also reviewed verified materials, liaising as needed with ETS/DIPF/ROA/CRP Henri Tudor on content and/or technical issues, before materials were “delivered” to countries.

Delivery took place via the Item Management Portal; countries were advised through OTRS when a batch of materials was verified, receiving precise instructions on how to further process the materials as well as a handy overview monitoring file.

These instructions are a convenient way to present the verification process in detail and are reused (in abridged form) in the two subsections that follow.

4.4.3 Detailed verification process – cognitive materials

Introduction – Process

The post-verification phase of the translation/adaptation/verification process for PIAAC assessment began after the verifier reviewed one or more batches of materials; the materials with suggested corrections and accompanying VFFs were made available on the IMP; and a Verification-Monitoring spreadsheet was provided, giving an overview of the verification outcomes for the verified batches. At this stage, it was the National Center’s responsibility to process the verification feedback and prepare for final check.

Background – Verification outcomes and how they were documented

PIAAC assessment materials were verified sentence by sentence, taking into account both general and item-specific translation and adaptation guidelines, with the aim to ensure the best possible balance between faithfulness to source version and fluency in the target version.

Verifiers’ suggested corrections were documented in VFFs, using a framework of intervention categories and severity codes (defining the nature and seriousness of identified issues). Figure 4.2 shows an example of a VFF showing a verifier’s intervention.

Figure 4.2: Example of VFF showing a verifier’s intervention

NPM COMMENT	VERIFIER INTERVENTION	SEVERITY CODE	VERIFIER COMMENT	DI
	Missing Info	2	'and places' missing from translation.	
	OK			

The severity codes have the following meaning:

- **Code 1** - serious error (likely to affect item functioning – must be addressed – will be rechecked)
- **Code 2** - minor error (better to correct, but not crucial, so will not be rechecked).
- **Code 3** - suggestion for improvement (implementation left to the discretion of the National Reviewer).

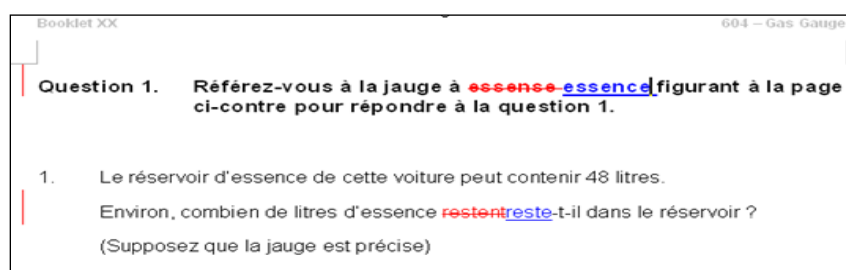
Only Code 1 errors gave rise to follow-up at final check, as explained later.

The verifiers’ suggested corrections were mostly implemented in the materials. (Exceptions: in some cases verifiers reported layout issues that they could not correct, or made suggestions that were better not implemented but left to countries’ initiative. Such exceptions were always

explicitly stated in the VFF: by default, verifiers' entries in the VFFs described problems that they went on to correct).

- Word units (paper-based) were corrected in “track changes” mode and needed to be processed by the National Reviewer (changes accepted or rejected or further modified). Figure 4.3 shows an example of a MS-Word file corrected in “track changes” mode.

Figure 4.3: Example of Word file corrected in ‘track changes’





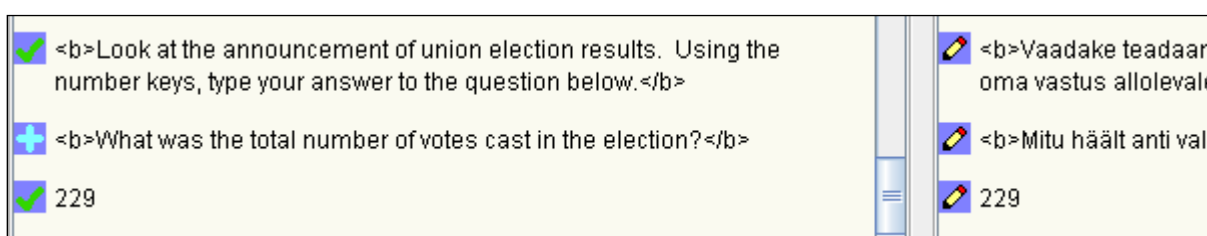
- XLIFF files (computer-based) were verified using OLT, which does not offer the “track changes” mode. Instead, to show where verifiers intervened, text segments were marked (on the left side or “source” side) either “approved”  (no changes made) or “translated”  (some edits made).
- The National Reviewer did not need to take any action inside these files except if he or she disagreed with an edit. Figure 4.4 shows an example of a verified XLIFF file viewed in the OLT interface.

Figure 4.4: Example of verified XLIFF file viewed in OLT



Processing verification feedback – Step 1: Getting an overview

National Reviewers were advised to first consult the Verification-Monitoring spreadsheet, which provided a handy overview of verification outcomes for the verified batches. Figure 4.5 shows an example of a Verification-Monitoring spreadsheet, showing different verification outcomes for the units within a batch.

Figure 4.5: Example of Verification-Monitoring spreadsheet showing verification outcomes

PIAAC2009FT VERIFICATION LITERACY		COUNTRY: ZEDLAND	LANGUAGE: ZEDISH	CODE: zd-ZD
Unit	Name	XLIFF VERIFICATION Computer-based	LAYOUT ISSUES? Computer-based	NAT. Comp FURTHER
LINKING ITEMS (LITERACY)				
300	Employment Ad	DONE: MINOR EDITS ONLY		
301	SIGH	DONE: NO EDITS AT ALL		
302	Election Results	DONE: NO EDITS AT ALL		
303	Preschool Rules	DONE: NO EDITS AT ALL	YES	
304	Contact Employer	DONE: MAJOR EDIT(S)		
305	TMN AntiTheft	DONE: MAJOR EDIT(S)	YES	
306	CANCO	DONE: MINOR EDITS ONLY		
307	MEDCO Aspirin	DONE: MINOR EDITS ONLY		

- A column indicates for each unit whether it was verified with no edits at all, or with minor edits only, or with (also) major edits. This information was designed to save work: a unit verified with no edits at all did not even need to be opened – it was unchanged versus the version submitted for verification; a unit verified with minor edits only (severity code 2 and/or 3) was not further checked at final check.
- For all computer-based units, a column indicated the presence of residual layout issues (text that did not fit or display correctly, etc.). These were either noticed during translation and confirmed by the verifier, or noticed during verification. A “YES” in this column alerted Consortium technical staff that action was needed to fix a layout problem – no action by countries was required.
- For numeracy units and (a few) literacy units that include items with the “numeric response mode,” an additional worksheet (Numeric-Entry-Scoring) indicated in column F the range of acceptable responses adapted to the country’s situation (currency values, metric/imperial). This followed a Consortium decision to uniformly extend the range of acceptable “exact match” responses, taking into account variability in the way respondents “spell” numbers (dot or comma as decimal separator, comma or dot or space as thousands separator, dash to indicate “no cents”). Column G indicated whether this implied the need to implement additional correct responses; this was handled centrally by the Consortium – no action by countries was required.

Based on examination of the Verification-Monitoring spreadsheet, the National Reviewer could decide for which units to consult the VFF, which provided details on the corrections made. In turn, based on consultation of the VFF, the National Reviewer could decide which XLIFF files or Word files to open for processing.

Processing verification feedback – Step 2: Consulting/annotating the VFF

No post-verification entry by the National Reviewer was required in the VFF except in the case of disagreement with a Code 1 correction (Code 2 and Code 3 corrections could be freely accepted or disregarded/undone). In case of disagreement with a Code 1 correction, the National Reviewer was to enter in the “Discussion” column of the VFF a justification for not following the verifier’s advice or correcting differently.

Processing verification feedback – Step 3a: Finalizing Word files

All Word files (paper-based units) with edits (including only minor edits) had to be opened so the corrections in track changes could be processed: accepted or rejected or (hopefully rarely) further modified. For reminder (see step 2), in case of rejection or further modification with regard to a Code 1 correction, the National Reviewer was to enter an explanation in the VFF.

Word files that were verified with no edits at all (as indicated in the monitoring spreadsheet) did not need to be opened – they were identical to the versions submitted for verification.

Finalized Word files needed to be uploaded to the IMP. This included Word files in which all corrections were just “accepted.” There was no need to upload Word files that had been verified with no edits at all and which had not been further changed by the National Reviewer.

Processing verification feedback – Step 3b: Finalizing XLIFF files

The text in XLIFF files (computer-based units) was corrected during verification and the corrections/suggestions were not “provisional” (not in “track changes” mode). Verified XLIFF files only needed to be opened if the National Reviewer wished to undo or (hopefully rarely) further modify a correction – or to implement a suggestion listed in the VFF but not actually implemented by the verifier. As a reminder (see step 2), in case of undoing or further modification with regard to a Code 1 correction, the National Reviewer had to enter an explanation in the VFF.

Note that if the National Reviewer made post-verification changes to the stimulus text of a literacy unit that includes “highlight” items, this could affect the definition of textblocks for scoring. He or she had to send an OTRS ticket in that case.

Finalized XLIFF files were uploaded to the IMP.

Processing verification feedback – Step 4: Returning the annotated Verification-Monitoring spreadsheet

After the above steps are completed, the National Reviewer was to return the Verification-Monitoring spreadsheet with the “Further Edits” columns filled in. Figure 4.6 shows an example of a Verification-Monitoring spreadsheet returned by a National Centre, showing where the National Reviewer has made post-verification changes.

Figure 4.6: Example of Verification-Monitoring spreadsheet filled in by Nat. Reviewer, to show where post-verification changes were made in computer-based units

PIAAC2009FT VERIFICATION					
LINK LITERACY		COUNTRY: AUSTRALIA	LANGUAGE: ENGLISH	CODE: en-AU	
Unit	Name	XLIFF VERIFICATION Computer-based (Outcome)	SCORING VERIFICATION Computer-based (Outcome)	LAYOUT ISSUES? Computer-based (DIPF/ETS will fix)	NAT. REVIEWER Computer-based FURTHER EDITS?
300	Employment Ad	DONE: NO EDITS AT ALL	DONE: NO EDITS AT ALL		
301	SIGH	DONE: NO EDITS AT ALL	NOT APPLICABLE		
302	Election Results	DONE: NO EDITS AT ALL	DONE: NO EDITS AT ALL		YES
303	Preschool Rules	DONE: NO EDITS AT ALL	DONE: MINOR EDITS ONLY		
304	Contact Employer	DONE: MINOR EDITS ONLY	DONE: NO EDITS AT ALL		
305	TMN AntiTheft	DONE: MINOR EDITS ONLY	DONE: NO EDITS AT ALL		
306	CANCO	DONE: NO EDITS AT ALL	DONE: NO EDITS AT ALL		

Final check

In the course of the final check procedure, units were reviewed in the following cases:

- Units were checked and corrected for residual layout issues and extension of acceptable “exact match” responses (technical final check).
- All units with major corrections (Code 1 corrections) were double-checked for correct implementation of such corrections (linguistic final check).
- In the case of computer-based units, this check was only needed for those Code 1 corrections for which the National Reviewer signaled disagreement in the VFF.
- In the case of paper-based units, this check was carried out on assembled booklets (PDF files), which were produced centrally by the Consortium.
- The workflow did not foresee another loop with units being returned to countries following the final check. Only the VFF was returned to countries upon completion of the final check, indicating for each Code 1 correction either “OK” or a comment suggesting that the issue was not satisfactorily solved. In the latter case, the National Reviewer still had the chance to address this, by making the recommended changes and re-uploading the affected units to the IMP.

4.4.4 Detailed verification process – BQ

Introduction – Process

The post-verification phase of the translation/adaptation verification process for the PIAAC BQ began after the verifier reviewed the nine XLIFF files of the BQ and annotated the BQAS; the materials with suggested corrections and accompanying BQAS were made available on the IMP

and a Verification-Monitoring spreadsheet, was provided, which gave an overview of the verification outcomes.

At this stage, it was the reviewer's responsibility to process the verification feedback. There was no final check phase for the BQ.

BQ – Verification outcomes and how they are documented



The verifier compared each segment of the national target version (right-hand side of the XLIFF files) with the national source version (NSV, left-hand side of the XLIFF files). Both general and item-by-item guidelines were taken into account.

The verifiers' suggested corrections were documented in columns 16a and 16b of the BQAS, using the same framework of intervention categories as for the Direct Assessment, but without severity codes. Figure 4.7 shows an example of a BQAS with a verifier's intervention.

Figure 4.7: Example of BQAS showing a verifier's intervention

1	2	16a	16b
Int. Question No	International English Version	Verifier Intervention	Verifier Comment
B_Q15d	Compared to your employer at the time, how useful do you think this training would be if you were working for a different employer? Would you say it was ...	Missing info	Missing "Would you say it was ...". Added by verifier.
	INTERVIEWER: Read categories to respondent.	OK	
	Much less useful	OK	
	Somewhat less useful	OK	
	Equally useful	OK	
	Somewhat more useful	OK	
	Much more useful	OK	

The verifiers' suggested corrections were mostly implemented in the materials. (Exceptions: in some cases verifiers made suggestions that were better not implemented but left to the country's initiative. Such exceptions were explicitly stated in the BQAS: by default, verifiers' entries in the BQAS described problems that they had corrected.)

- As for the Direct Assessment, XLIFF files were verified using OLT, which does not offer the "track changes" mode. Instead, to show where verifiers intervened, text segments were marked (on the left side or "source" side) either "approved"  (no changes made) or "translated"  (some edits made). The National Reviewer did not need to take any action inside these files except if he or she disagreed with an edit.

Processing verification feedback – Step 1: Getting an overview

National Reviewers were advised to first consult the BQ worksheet of the Verification-Monitoring spreadsheet, which provides a handy overview of verification outcomes. Figure 4.8 shows an example of a Verification-Monitoring spreadsheet with different verification outcomes for each of the BQ sections.

Figure 4.8: Example of Verification-Monitoring spreadsheet showing verification outcomes

PIAAC2009FT VERIFICATION BACKGROUND-QUESTIONNAIRE		COUNTRY: ZEDLAND	LANGUAGE: ZEDISH	CODE: zd-ZD
Unit	Name	XLIFF VERIFICATION Computer-based (Outcome)	ADAPTATION ISSUES (CONTENT) / ROA ADVICE NEEDED (ROA advice needed)	ADAPTATION ISSUES (DYNAMIC TEXT) / CRP ADVICE NEEDED (CRP advice needed)
BQ	BQ Section AB (bq)	DONE, WITH EDITS	NO	NO
BQ	BQ Section C (bq)	DONE: NO EDITS AT ALL	NO	NO
BQ	BQ Section D (bq)	DONE, WITH EDITS	NO	NO
BQ	BQ Section E (bq)	DONE, WITH EDITS	YES	NO
BQ	BQ Section F (bq)	DONE, WITH EDITS	NO	NO
BQ	BQ Section G (bq)	DONE, WITH EDITS	NO	YES
BQ	BQ Section H (bq)	DONE, WITH EDITS	NO	NO
BQ	BQ Section I (bq)	DONE, WITH EDITS	NO	NO
BQ	BQ Section J (bq)	DONE: NO EDITS AT ALL	NO	NO

- A column indicated for each section of the BQ whether it was verified with or without edits. A unit verified with no edits did not need to be opened – it is unchanged versus the version submitted for verification (e.g. Sections C and J in Figure 4.8 above).
- Another column indicated the possible occurrence of residual adaptation issues that the verifier was unable to resolve and that (may have) required consultation with the BQ group (e.g. Section E in Figure 4.8 above). Usually an OTRS ticket was sent by cApStAn to the BQ group concerning such issues, and the issue was resolved or needed to be resolved between the National Reviewer and the BQ group.
- A last column indicated the possible occurrence of dynamic text issues, e.g., when the country commented in the DTRS that a given question did not require gender-related duplication or that a given past tense/present tense question required the introduction of an additional segment (e.g., Section G in Figure 4.8 above). Such issues were transmitted to CRP, and CRP contacted the reviewer concerning the best way to handle such issues.

Processing verification feedback – Step 2: Consulting/annotating the BQAS

The National Reviewer entered post-verification comments in the BQAS, for example, in the case of disagreement with a correction. If there was an additional iteration with the BQ group

(see bullet points 2 and 3 of Step 1), there might be a Consortium comment in the BQAS. This would need to be taken into account when finalizing the BQ. In case of disagreement with a proposed correction, the reviewer sent the BQAS to ROA.

BQAS with post-verification comments were uploaded to the IMP.

Processing verification feedback – Step 3: Finalizing XLIFF files

The text in XLIFF files was corrected during verification and the corrections/suggestions were not “provisional” (not in “track changes” mode). Verified XLIFF files only needed to be opened if the National Reviewer wished to undo or (hopefully rarely) further modify a correction – or to implement a suggestion listed in the BQAS but not actually implemented by the verifier.

Finalized XLIFF files were uploaded to the IMP.

Processing verification feedback – Step 4: returning the annotated Verification-Monitoring spreadsheet

After the above steps were completed, the National Reviewer returned the Verification-Monitoring spreadsheet with the “Further Edits” column filled in. Figure 4.9 shows an example of a Verification-Monitoring spreadsheet returned by a National Centre, showing where the National Reviewer has made post-verification changes.

Figure 4.9: Example of Verification-Monitoring spreadsheet filled in by Nat. Reviewer, to show where post-verification changes have been made in the XLIFF files of BQ sections

PIAAC2009FT VERIFICATION BACKGROUND QUESTIONNAIRE		COUNTRY: ZEDLAND	LANGUAGE: ZEDISH	CODE: zd-ZD	
Unit	Name	XLIFF VERIFICATION Computer-based	ADAPTATION ISSUES (CONTENT) / ROA ADVICE NEEDED	ADAPTATION ISSUES (DYNAMIC TEXT) / CRP ADVICE NEEDED	NAT. REVIEWER Computer-based
		(Outcome)	(ROA advice needed)	(CRP advice needed)	FURTHER EDITS?
BQ	BQ Section AB (bq)	DONE, WITH EDITS	NO	NO	NO
BQ	BQ Section C (bq)	DONE: NO EDITS AT ALL	NO	NO	NO
BQ	BQ Section D (bq)	DONE, WITH EDITS	NO	NO	YES
BQ	BQ Section E (bq)	DONE, WITH EDITS	YES	NO	NO
BQ	BQ Section F (bq)	DONE, WITH EDITS	NO	NO	NO
BQ	BQ Section G (bq)	DONE, WITH EDITS	NO	YES	YES
BQ	BQ Section H (bq)	DONE, WITH EDITS	NO	NO	NO
BQ	BQ Section I (bq)	DONE, WITH EDITS	NO	NO	NO
BQ	BQ Section J (bq)	DONE: NO EDITS AT ALL	NO	NO	NO

Final check

Different from Direct Assessment units, there was no final check procedure for the BQ. The Verification-Monitoring spreadsheet and the BQAS with the National Reviewer’s annotations were archived to keep a trace of the “history” of each national version of the BQ instrument, and reused when preparing the Main Study instrument.

4.5 International verification of the national versions – Main Study

The guiding principle of PIAAC Main Study translation/adaptation and verification activities was to control and limit the changes made by National Centers to their finalized Field Test national versions of assessment instruments, and carry out a verification focused on just these changes, with exceptions as needed and more extensive checks in identified “risky” cases, as well as a full verification of newly translated materials (the Main Study CAPI Workflow, Help Screens and Orientations).

The above scheme applied to “Phase I” of a two-phase process for revising and adapting the materials for the Main Study, devised in order to accommodate the tight timeline between the Field Test and Main Study. Phase I took place from May to November 2010, prior to analysis of the Field Test data, and focused on correcting issues associated with wording, scoring and layout that were identified by countries and the Consortium.

Phase II followed immediately after analysis of the Field Test data, from December 2010 to January 2011, and focused on identifying and correcting errors to the PIAAC instrumentation based on analysis of the Field Test data. During Phase II, a number of cognitive and BQ items that did not work well in the Field Test for a majority of countries were dropped. In addition, country review was allowed for a very limited set of cognitive items that functioned well for most, but not all, countries. Countries were asked to document the possible source of error and proposed solutions (beyond fixes which might already have been made during Phase I). A very limited number of last-minute changes were thus made at Phase II. These were discussed, approved and tracked, but not formally “verified” owing to the time pressure.

The rest of this section will describe the verification processes implemented during Phase I, where the great majority of Field Test to Main Study revisions were made. Countries were instructed on procedures at the NPM Meeting in Frankfurt in June 2010 as well as in the preparatory run-up to that meeting.

4.5.1 Main Study verification of literacy, numeracy and problem-solving units

The starting point of the process was the MMF (Main Study Translation-Adaptation-Verification Monitoring Form). One MMF for each of five batches (Link Literacy, Link Numeracy, New Literacy, New Numeracy, Problem Solving) was prepared and initially sent to National Centers with instructions to take note of the Main Study revisions and checks requested by the Consortium and add their requests for “national” changes (with a strong recommendation to limit these to corrections of errors, avoiding cosmetic or stylistic changes).

Note: specially customized MMFs were prepared for Hungary, which had dropped out of the Field Test process and rejoined for the Main Study (MMFs based on Slovak-Hungarian materials) and for the minority-language versions for Spain (MMFs with additional checks for Catalan, Galician, Basque, and Valencian materials which had not been final-checked at Field Test).

Countries’ requests for changes were evaluated by the item development teams (sometimes with the assistance of verifiers who provided linguistic advice in a “pre-verification phase”) and a Consortium recommendation for each request (approval, approval with caution, or rejection) was documented in the MMF.

On the basis of this preliminary work, the MMF was further filled out with cells to document and follow up on all “agreed revisions” (in both computer-based materials, in a first stage, and paper-based materials, later), using a color scheme to facilitate differential processing:

- Blue cells: text changes to be implemented by country and subject to verification and, if applicable, to final check.
- Yellow cells: layout changes or numeric scoring changes to be implemented by the Consortium’s technical teams (DIPF or ETS) and subject to country’s sign-off.
- Mauve cells (in literacy units only): revision of text blocks for the scoring of highlight items, to be implemented by country and subject to a verification procedure (see later).

Figure 4.10 shows an example of a MMF documenting the Main Study verification process of cognitive units.

Figure 4.10: Example of MMF with blue and yellow cells showing the verification process

PIAAC MMF (MAIN STUDY TRANSLATION ADAPTATION VERIFICATION MONITORING FORM) NEW NUMERACY Language_Country: et-EE							
Unit N°	Unit Name	Location	MS revisions and checks	Additional MS changes	Feedback on MS changes	AGREED CBA REVISION ACTION (WHO) • FOLLOW-UP	AGREED PBA REVISION ACTION (WHO) • FOLLOW-UP
634	Peanuts	Stimulus	NONE			NONE	N.A.
		Question 1 (C634P001)	NONE	Change “Mitu grammi (g) süsivesikuid sisaldab pakitaia pähkleid?” Reason: better understandable	FEEDBACK: You will be able to make this revisions to CBA using the tool to be introduced at the June meeting in Frankfurt. Those revisions will then be verified by cApStAn.	TEXT CHANGE - IMPLEMENTED BY COUNTRY, SUBJECT TO VERIFICATION Country comment: DONE Verifier comment: OK, no further changes needed Country post-verif comment/signoff: N/A Verifier final check: N/A	N.A.
		Question 1 - Scoring (Left Panel Numeric Entry Number Match)	NONE			NONE	N.A.
		Question 2 (C634P002)	NONE			NONE	N.A.
		Question 2 - Scoring (Left Panel Numeric Entry Number Match)	NONE			NONE	N.A.
635	Parking Map	Stimulus	REVISION CBA: The “3” representing the 3 km/mi mark should be aligned to the ruler.			LAYOUT CHANGE Implemented by ETS ETS comment: DONE Country comment/sign-off: It's ok	N.A.

Countries were provided with a manual explaining how to process the MMF and with the technical instructions for accessing materials on the IMP (or on the online “Copernicus” in the case of problem-solving units) and checking changes made by the Consortium (yellow cells), making changes under their responsibility (blue cells), and, for the two literacy batches, checking/correcting the “text blocks” used for highlight scoring and running a full testing protocol on highlight items.

After this pass by countries, the materials moved to verification phase.

Verifiers had read-only access to the units (via the preview facility on the IMP or on Copernicus) and were instructed to check all blue cells in the MMF, making sure that agreed changes were implemented correctly (and not-approved changes were not implemented). Verifiers were further advised that the changes were approved or rejected by the Consortium based on information

given by the country (not always very detailed or informative) and mostly with little or no knowledge of the language; therefore they were allowed to contradict or question the decision in cases where an agreed change could make the item easier or more difficult, linguistically poor, or causing an additional problem that was not taken into account by the country.

If the verifier detected no issue, the blue cell would be completed with markings that no further processing was needed (no need for final check). Otherwise, the verifier would describe the issue and suggest corrective action, and the blue cell would be marked for final check after post-verification review by the country.

Verifiers were also instructed to make use of the “Diff report” facility on the portal to detect and process any undocumented changes made by the country (this was a key feature to enable a “safe” focused verification procedure).

To verify the correct scoring of literacy items with the highlight response mode, a more efficient and focused procedure was put in place for the Main Study.

After countries revised their units, which included checking/correcting the text blocks and testing the highlight items, DIPF classified the national versions as low, medium or high risk, based on a review of the problems found at Field Test and of the quality and thoroughness of the Main Study scoring testing.

It was agreed that cApStAn would carry out a sample-based check of each country’s testing by performing a certain number of testing steps and checking that one received the same expected results. The list of testing steps to be performed was variable depending on the country’s classification. At minimum (low risk category): cApStAn tested all items for which the scoring rules were changed between Field Test and Main Study, any residual issues from Field Test testing (on a case by case basis), and three to five test cases chosen at random in other units than those already tested. For the medium risk category, cApStAn added six to nine test cases chosen at random in other units than those already tested. For the high risk category, cApStAn ran one or two additional test cases in each and every item.

Results were reported in the mauve cells of the Literacy MMFs with details in the separate scoring sheet where countries had documented their testing. For national versions that “passed” the validation procedure, countries were advised to nevertheless retest their scoring in case of text changes suggested by verifier that could affect the definition of text blocks (after implementing these at post-verification review stage). Such cases were clearly identified in the MMF. Countries that failed the validation procedure were asked to recheck and retest all highlight items and given further assistance.

4.5.2 Main Study verification of the BQ

As for literacy, numeracy and problem-solving units, in the Main Study, the principle was to verify only changes made to the BQ since the Field Test. The environment and process, however, were quite different. BQ sections were verified by reviewing and editing XLIFF files using OLT. The XLIFF files submitted by each country were specially prepared “partial” ones containing only the segments that countries needed or wished to change (not the entire BQ text), following a process of approval of national changes carried out with ROA.

When viewed in the OLT interface, verifiers see a “customized” (and approved) source version on the left and the country’s target version on the right. They were instructed to verify that the target texts are linguistically correct and match the customized source texts, make corrections as needed, and document these corrections.

The documentation and follow-up of verification corrections occurred in a new monitoring form created for the Main Study, replacing the unwieldy BQAS used in the Field Test. The “Main Study BQ Verification Report” form was designed to allow National Centers to easily identify where edits were made by verifiers and revert to their original translations. Follow-up columns were included for possible comments on verification issues by ROA (content issues) and/or CRP (technical issues, e.g., missing segments for dynamic text variants), who were invited to add comments after the verifier’s pass and verification review by cApStAn, and indicate or confirm any issues considered as crucial and thus subject to final check.

The edited XLIFF files for BQ sections and the BQ Verification Report form were then sent to country reviewers for post-verification processing. Countries were instructed that they could make post-verification edits (e.g., to undo or further modify a correction made by the verifier), that they were free to comment on their choices or not for non-crucial issues, e.g., minor linguistic defects, but were required to comment on issues marked as crucial and subject to final check. Countries were further advised that the most useful way to reply was, for example, “OK, we agree with verifier/ROA advice so no further change was made to the already corrected segment” or “We have changed to xxx because of reason yyy.”

After the country’s post-verification review, the files came back to cApStAn for final check. If an issue marked for final check was found not to be satisfactorily resolved, there could be one more iteration with the country before final signoff. Figure 4.11 shows an example of a BQ Verification Report with the documentation of a particular issue through all successive steps.

Figure 4.11: Example of BQ Verification Report with an issue documented through all successive steps

PIAAC MAIN STUDY VERIFICATION REPORT BQ								
Language:			Country: Portugal		Code: pt_PT			
XLIFF Segment No	Identifier (XLIFF comment text)	Source version	Translated version	Verifier’s suggested version	Verifier comment	CRP - ROA comment	Country post-verif comment /Signoff	Verifier final check
SECTION E 8	Code: E_Q02a / Type: Question	In what kind of business, industry or service did you work? Please give a full description.	Em que tipo de empresa ou organismo trabalhava? Por favor, responda de forma detalhada. Por exemplo: ensino pré-escolar, tribunal, centro de saúde, câmara municipal, fiação de fibras de algodão, fabricação de tecidos de malha, preparação e conservação de peixe, fabricação de pão, comércio a retalho de vestuário, construção de estradas, etc.	Em que tipo de negócio, indústria ou serviço trabalhava? Por favor, responda de forma detalhada. Por exemplo: ensino pré-escolar, tribunal, centro de saúde, câmara municipal, fiação de fibras de algodão, fabricação de tecidos de malha, preparação e conservação de peixe, fabricação de pão, comércio a retalho de vestuário, construção de estradas, etc.	Consistency (according to the instructions that follow the question). Translation is much more specified than source version, includes examples of professions. Verifier did NOT change	ROA: Consistency: agree in principle, as long as this improves the clarity of the translated version Translation more specified than source: see above.	X We would like to keep the previous version to be consistent with other national surveys, like Labor Force Survey. This translation/extension was approved to the FT and would like to keep it.	OK, but then better be consistent using “empresa ou organismo” to translate “business, industry or service” in all occurrences → We have thus changed to “empresa ou organismo” also in Section E segment 9 (see below) and in Section D segments 8 and 9 (see above) PT: Ok, thank you!

Note: for three national versions (Japan, Korea and the Russian Federation¹²), the countries requested and obtained approval to revise the entire BQ. The verification of these three versions was hence full rather than partial (changes only), but followed the same procedures described in this section.

¹² Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

4.5.3 Main Study verification of the CAPI workflow, Help screens, and orientations

For the Field Test, these “ancillary materials” were translated centrally (by the Consortium) to mitigate the heavy translation workload for National Centers. For the Main Study, with the lesser translation workload, these translations followed the “decentralized” model and were thus subject to full verification.

The verification process was similar to the one implemented for the BQ and described in the section above: translation/adaptation by country in XLIFF files using OLT, verification via edits to the XLIFF files documented in a Verification Report form, post-verification review by country documented in the same form, but no final check in the case of these materials.

An important difference was an extra column in the Verification report form labeled “Special instructions, checks, errata.” These were instructions for verifiers, specially prepared after thorough analysis of the files, with a view in particular to ensure key matches between elements appearing in these files with the translations used in test units (e.g., names of units, correct responses for “core” items scored by the interviewer, names of the problem-solving “environments” and tooltips, etc.), and in the CAPI interface. Figure 4.12 shows an excerpt from a Workflow-Help-Orientation verification report with some checks to be performed by the verifier.

Figure 4.12: Excerpt of WF-HELPS-OR Verification Report showing some checks to be performed by verifiers

PIAAC MAIN STUDY VERIFICATION REPORT ORIENTATIONS		
XLIFF Segment No	Source version	Special instructions, checks, errata
ORIENTATION - LITERACY		
22	Forward	To be understood as per context "Go forward", not as "Forward an email". The translation must match the term used in the country's PS Web environment
29, 30, 31, 32	Comma Period Slash for fractions Dash for negative numbers	These terms should be translated consistently across the Helpscreen-Literacy, Helpscreen-Numeracy, Orientation-Literacy and Orientation-Numeracy files
36	Thank you. Go to the next question OR Click here to go back and change your answer	The translation must match the translation used in e.g. 318-Civil engineering (New-Lit), Q1, when you click on any link.
40, 41, 42, 43, 57, 58	Email Web	"Email" and "Web" should match the terms used on the environment tabs in the country's New Literacy unit 327 Summer Streets . NOTE: you need to click on the link on the opening page to arrive in a screen in which both buttons are shown.
45	... To review how to answer a question, click on Help	Though formatting of references to buttons is not consistent in the source, make sure that it is in the target: either capitalization or other consistent formatting of NEXT, SPACE, BACKSPACE, HELP, BACK throughout the Helpscreen + Orientations files Note: source should be To review how to answer a question, click on HELP.

4.5.4 Main Study verification of paper-based materials

For the Main Study, paper-based materials were prepared and verified on an “easier” timeline than computer-based materials, from March to June 2011. Procedures differed for the various materials, which comprised:

- Paper-based test units, assembled in three booklets: Core, Literacy and Numeracy.
- Reading components exercises, assembled in one RC booklet
- Scoring Guides for the Core, Literacy and Numeracy booklets

For the paper-based test units, the Literacy and Numeracy MMFs used to document and approve Field Test to Main Study changes during Phase I were “exhumed” (as a reminder, these included changes to paper-based as well as computer-based units, which needed to be considered together) and complemented with new changes or checks resulting from the Phase II revision process. Countries were instructed to make changes under their responsibility (corresponding to blue cells in the MMFs, see earlier description of the process for computer-based units) to the final Main Study-Word files from the Field Test, in track changes mode, and send these to the Consortium’s pre-press specialist, Danielle Baum. Danielle constructed initial PDF booklets based on the Main Study master versions, implementing the formatting/layout changes under Consortium responsibility (corresponding to yellow cells in the MMFs).

Verification of the correct implementation of agreed changes was carried out on these PDF booklets, with reporting of issues (and suggested corrections) in the MMFs. Verifiers also had access to the Main Study-Word files showing the Field Test>Main Study changes in track changes mode, which was handy if they needed to see the previous wording. In addition, verifiers could preview the computer-based version (where applicable) to ensure alignment of changes in paper-based with changes in computer-based. (Note: after a number of discrepancies were found in Spain-Galician and Spain-Basque materials between the PBA and the CBA, a full PBA to CBA identicalness check was carried out for these two versions.)

The MMFs were then used for post-verification review by countries, implementation of corrections by the Consortium's DTP specialist, final check by cApStAn on revised PDF booklets, and signoff by countries.

Note that the booklets also included a cover page and an introduction. These elements were also verified (classically, for equivalence to source and linguistic correctness), but the verification was documented and followed-up in a different MMF, together with the scoring guides and reading components, the "Guides and Booklets MMF."

The three scoring guides were verified in Main Study-Word, with suggested corrections implemented in track changes mode and followed-up (post-verification review by country and final check of crucial issues) via the "Guides and Booklets MMF." The "Guidelines for Scorers" section was verified classically, for equivalence to source and linguistic correctness. For the scoring sections, verifiers were instructed also to check that unit names, question stems and other elements matched the actual units, and that scoring instructions were properly adapted according to precise instructions inserted in the MMF. Figure 4.13 shows an example of a MMF with verifier's interventions in the scoring sections of a scoring guide.

Figure 4.13: Example of MMF showing verification interventions in scoring sections

PIAAC MAIN STUDY 2011 Country: Portugal Target language: Portuguese					
PLEASE INSERT NEW LINES, IF NEEDED, TO DOCUMENT ADDITIONAL ISSUES					
LOCATION	ENGLISH SOURCE	CONSORTIUM RECOMMENDATION	VERIFIER INTERVENTION	SEVERITY CODE	VERIFIER COMMENT
	Question 2: List two ways in which CIEM helps people who will lose their jobs because of a departmental reorganization.		Consistency	2	Name (in the singular form) not consistent with the booklet. Changed by verifier.
	1 Mentions BOTH of the following: • They act as a mediator for employees OR mediation • They assist with finding new positions [Note: Do not accept "Job Data Bank"; "Guidance"; "Courses"; or "Career Change Projects" These responses should receive a score of 7]	Attention: "Job Data Bank", "Guidance", "Courses" and "Career Change Projects" (literal matches with stimulus - see the dot points in left column).	OK		
313 - International calls	Question 3: Identify the two situations in which you might have to dial 098.		OK		
	1 Mentions BOTH of the following: • For help connecting a call • (To make calls in countries where the list says) the service is manual AND/OR via operator 7 Any other response 0 Stimulus and response page(s) left completely blank Correct answer: 1	Attention: Maintain literal/quasi-literal matches with stimulus	Consistency	2	Translation of "call" not consistent with the booklet. Changed by verifier.

The reading components were treated as a special case, given that a subset of materials used in the Field Test was selected for the Main Study, with no Field Test-to-Main Study changes. The “verification” of these materials consisted of a careful check that national materials were correctly assembled (using the correct selected materials as in the Main Study international master version), from the verified and finalized Field Test versions. To ensure the latter, verifiers were instructed to randomly check one or two verifier interventions from the Field Test in each section (vocabulary, sentence processing and passage comprehension).

References

- Dorer, B. 2011. Advance translation in the 5th round of the European Social Survey (ESS). *FORS working paper series*, (Paper No. 2011-4). Lausanne: FORS.
- Harkness, J. A. 2003. Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). Hoboken, NJ: John Wiley & Sons.
- Harkness, J. A., Villar, A. & Edwards, B. 2010. Translation, Adaptation, and Design. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. P. Mohler, B-E. Pennell, T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 117-140). Hoboken, NJ: John Wiley & Sons.
- Dept S., Ferrari, A. and Wäyrynen, L., 2010. Developments in Translation Verification Procedures in Three Multilingual Assessments: A Plea for an Integrated Translation and Adaptation Monitoring Tool. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. P. Mohler, B-E. Pennell, T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 157-173). Hoboken, NJ: John Wiley & Sons.
- Upsing B., Gissler G., Goldhammer F., Rölke H., Ferrari A., 2011. Localisation in International Large-Scale Assessments of Competencies: Challenges and Solutions. In R. Schäler (Ed.), *Localisation Focus, Vol.10- Issue 1*. Limerick, Ireland.