Chapter 9
Large-Scale Assessments of Adult Literacy

Irwin Kirsch, Mary Louise Lennon, Kentaro Yamamoto, and Matthias von Davier

Educational Testing Service's (ETS's) work in large-scale adult literacy assessments has been an ongoing and evolving effort, beginning in 1984 with the Young Adult Literacy Survey in the United States. This work has been designed to meet policy needs, both in the United States and internationally, based on the growing awareness of literacy as human capital. The impact of these assessments has grown as policy makers and other stakeholders have increasingly come to understand the critical role that foundational skills play in allowing individuals to maintain and enhance their ability to meet changing work conditions and societal demands. For example, findings from these surveys have provided a wealth of information about how the distribution of skills is related to social and economic outcomes. Of equal importance, the surveys and associated research activities have contributed to large-scale assessment methodology, the development of innovative item types and delivery systems, and methods for reporting survey data in ways that ensure its utility to a range of stakeholders and audiences.

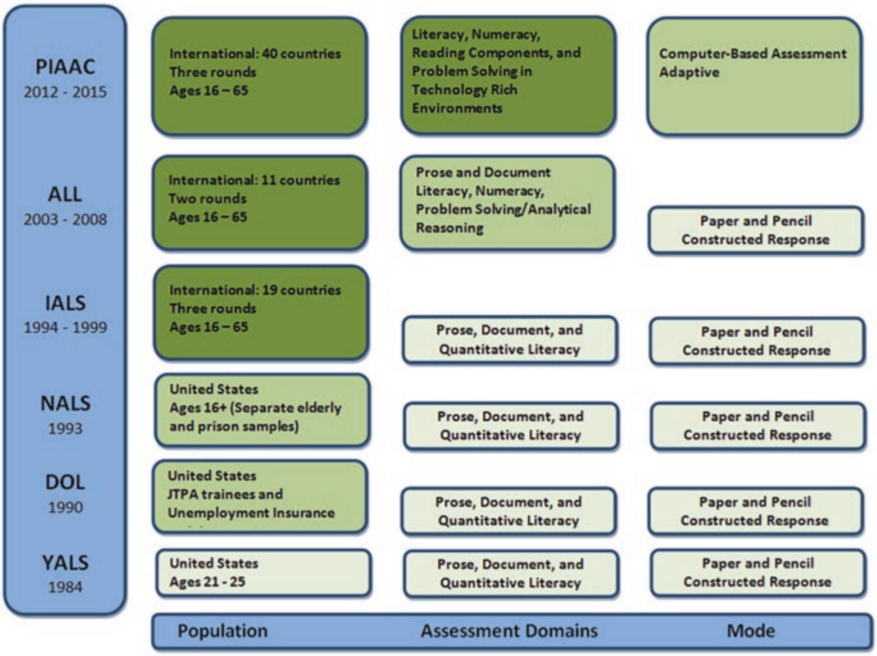
The chronology of ETS's large-scale literacy assessments, as shown in Fig. 9.1, spans more than 30 years. ETS served as the lead contractor in the development of these innovative assessments, while the prime clients and users of the assessment outcomes were representatives of either governmental organizations such as the National Center for Education Statistics (NCES) and Statistics Canada, or trans-governmental entities such as the Organisation for Economic Co-operation and Development (OECD). These instruments have evolved from a single-language, paper-based assessment focusing on a U.S. population of 16- to 25-year-olds to an adaptive, computer-based assessment administered in almost 40 countries and close to 50 languages to adults through the age of 65. By design, the assessments have been linked at the item level, with sets of questions from previous assessments

This work was conducted while M. von Davier was employed with Educational Testing Service.

I. Kirsch (✉) • M.L. Lennon • K. Yamamoto • M. von Davier
Educational Testing Service, Princeton, NJ, USA
e-mail: ikirsch@ets.org

© Educational Testing Service 2017

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_9



	Population	Assessment Domains	Mode
PIAAC 2012 - 2015	International: 40 countries Three rounds Ages 16 – 65	Literacy, Numeracy, Reading Components, and Problem Solving in Technology Rich Environments	Computer-Based Assessment Adaptive
ALL 2003 - 2008	International: 11 countries Two rounds Ages 16 – 65	Prose and Document Literacy, Numeracy, Problem Solving/Analytical Reasoning	Paper and Pencil Constructed Response
IALS 1994 - 1999	International: 19 countries Three rounds Ages 16 – 65	Prose, Document, and Quantitative Literacy	Paper and Pencil Constructed Response
NALS 1993	United States Ages 16+ (Separate elderly and prison samples)	Prose, Document, and Quantitative Literacy	Paper and Pencil Constructed Response
DOL 1990	United States JTPA trainees and Unemployment Insurance	Prose, Document, and Quantitative Literacy	Paper and Pencil Constructed Response
YALS 1984	United States Ages 21 - 25	Prose, Document, and Quantitative Literacy	Paper and Pencil Constructed Response

Fig. 9.1 ETS's large-scale literacy assessments. Note. *ALL* = Adult Literacy and Life Skills Survey (Statistics Canada, Organisation for Economic Co-operation and Development [OECD]), *DOL* = Department of Labor Survey, *JPTA* = Job Training Partnership Act, *IALS* = International Adult Literacy Survey (Statistics Canada, OECD), *PIAAC* = Programme for the International Assessment of Adult Competencies (OECD), *YALS* = Young Adult Literacy Survey (through the National Assessment of Educational Progress)

included in each new survey. This link has made it possible to look at changes in skill levels, as well as the distribution of those skills, over time. Each of the assessments has also expanded upon previous surveys. As Fig. 9.1 illustrates, the assessments have changed over the years in terms of who is assessed, what skills are assessed, and how those skills are assessed. The surveys have evolved to include larger and more diverse populations as well as new and expanded constructs. They have also evolved from a paper-and-pencil, open-ended response mode to an adaptive, computer-based assessment.

In many ways, as the latest survey in this 30-year history, the Programme for the International Assessment of Adult Competencies (PIAAC) represents the culmination of all that has been learned over several decades in terms of instrument design, translation and adaptation procedures, scoring, and the development of interpretive schemes. As the first computer-based assessment to be used in a large-scale household skills survey, the experience derived from developing and delivering PIAAC—including research focused on innovative item types, harvesting log files, and delivering an adaptive assessment—helped lay the foundation for new computer based large-scale assessments yet to come.

This paper describes the contributions of ETS to the evolution of large-scale adult literacy assessments in six key areas:

- Expanding the construct of literacy
- Developing a model for building construct-based assessments
- Expanding and implementing large-scale assessment methodology
- Linking real-life stimulus materials and innovative item types
- Developing extensive background questionnaires to link performance with experience and outcome variables
- Establishing innovative reporting procedures to better integrate research and survey data

9.1 Expanding the Construct of Literacy

Early work in the field of adult literacy defined literacy based on the attainment of certain grade level scores on standardized academic tests of reading achievement. Standards for proficiency increased over the decades with “functional literacy” being defined as performance at a fourth-grade reading level during World War II, eighth-grade level in the 1960s, and a 12th grade level by the early 1970s. This grade-level focus using instruments that consisted of school-based materials was followed by a competency-based approach that employed tests based on nonschool materials from adult contexts. Despite this improvement, these tests still viewed literacy along a single continuum, defining individuals as either *literate* or *functionally illiterate* based on where they performed along that continuum. The 1984 Young Adult Literacy Survey (YALS) was the first in a series of assessments that contributed to an increasingly broader understanding of what it means to be “literate” in complex modern societies. In YALS, the conceptualization of literacy was expanded to reflect the diversity of tasks that adults encounter at work, home, and school and in their communities. As has been the case for all of the large-scale literacy assessments, panels of experts were convened to help set the framework for this assessment. Their deliberations led to the adoption of the following definition of literacy: “using printed and written information to function in society, to achieve one’s goals, and to develop one’s knowledge and potential” (Kirsch and Jungeblut 1986, p. 3).

This definition both rejected an arbitrary standard for literacy, such as performing at a particular grade level on a test of reading, and implied that literacy comprises a set of complex information-processing skills that goes beyond decoding and comprehending text-based materials.

To better reflect this multi-faceted set of skills and abilities, performance in YALS was reported across three domains, defined as follows (Kirsch and Jungeblut 1986, p. 4):

- Prose literacy: the knowledge and skills needed to understand and use information from texts including editorials, news stories, poems, and the like
- Document literacy: the knowledge and skills required to locate and use information contained in job applications or payroll forms, bus schedules, maps, indexes, and so forth

- Quantitative literacy: the knowledge and skills required to apply arithmetic operations, either alone or sequentially, that are embedded in printed materials, such as in balancing a checkbook, figuring out a tip, completing an order form, or determining the amount of interest on a loan from an advertisement

Rather than attempt to categorize individuals, or groups of individuals, as literate or illiterate, YALS reported results for each of these three domains by characterizing the underlying information-processing skills required to complete tasks at various points along a 0–500-point reporting scale, with a mean of 305 and a standard deviation of about 50. This proficiency-based approach to reporting was seen as a more faithful representation of both the complex nature of literacy demands in society and the various types and levels of literacy demonstrated by young adults.

Subsequent research at ETS led to the definition of five levels within the 500-point scale. Analyses of the interaction between assessment materials and the tasks based on those materials defined points along the scale at which information-processing demands shifted. The resulting levels more clearly delineated the progression of skills required to complete tasks at different points on the literacy scales and helped characterize the skills and strategies underlying the prose, document, and quantitative literacy constructs. These five levels have been used to report results for all subsequent literacy surveys, and the results from each of those assessments have made it possible to further refine our understanding of the information-processing demands at each level as well as the characteristics of individuals performing along each level of the scale.[1]

With the 2003 Adult Literacy and Life Skills Survey (ALL), the quantitative literacy domain was broadened to reflect the evolving perspective of experts in the field. The new numeracy domain was defined as the ability to interpret, apply, and communicate numerical information. While quantitative literacy focused on quantitative information embedded in text and primarily required respondents to demonstrate computational skills, numeracy included a broader range of skills typical of many everyday and work tasks including sorting, measuring, estimating, conjecturing, and using models. This expanded domain allowed ALL to collect more information about how adults apply mathematical knowledge and skills to real-life situations. In addition, the ALL assessment included a problem-solving component that focused on analytical reasoning. This component collected information about the ability of adults to solve problems by clarifying the nature of a problem and developing and applying appropriate solution strategies. The inclusion of problem solving was seen as a way to improve measurement at the upper levels of the scales and to reflect a skill set of growing interest for adult populations.

Most recently, the concept of literacy was expanded again with the Programme for the International Assessment of Adult Competencies (PIAAC). As the first computer-based, large-scale adult literacy assessment, PIAAC reflected the changing nature of information, its role in society, and its impact on people's lives.

[1] See the appendix for a description of the information-processing demands associated with each of the five levels across the literacy domains.

The scope of the prose, document, and numeracy domains was broadened in PIAAC and the assessment incorporated two new domains, as follows:

- For the first time, this adult assessment addressed literacy in digital environments. As a computer-based assessment, PIAAC included tasks that required respondents to use electronic texts including web pages, e-mails, and discussion boards. These stimulus materials included hypertext and multiple screens of information and simulated real-life literacy demands presented by digital media.
- In PIAAC, the definition of numeracy was broadened again to include 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 (i.e., beliefs and attitudes) are necessary to meet the demands of numeracy effectively in everyday life.
- PIAAC included the new domain of problem-solving in technology-rich environments (PS-TRE), the first attempt to assess this domain on a large scale and as a single dimension. PS-TRE was defined as:

  > using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks. 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. (OECD 2012, p. 47)

  PS-TRE presented computer-based tasks designed to measure the ability to analyze various requirements of a task, define goals and plans, and monitor progress until the task purposes were achieved. Simulated web, e-mail and spreadsheet environments were created and respondents were required to use multiple, complex sources of information, in some cases across more than one environment, to complete the presented tasks. The focus of these tasks 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 contained a reading components domain, which included measures of vocabulary knowledge, sentence processing, and passage comprehension. Adding this domain was an important evolution because it provided more information about the skills of individuals with low levels of literacy proficiency than had been available from previous international assessments. To have a full picture of literacy in any society, it is necessary to have more information about these individuals because they are at the greatest risk of negative social, economic, and labor market outcomes.

9.2 Developing a Model for Building Construct-Based Assessments

A key characteristic of the large-scale literacy assessments is that each was based on a framework that, following Messick's (1994) construct-centered approach, defined the construct to be measured, the performances or behaviors expected to reveal that construct, and the characteristics of assessment tasks to elicit those behaviors. In the course of developing these assessments, a model for the framework development process was created, tested, and refined. This six-part process, as shown in Fig. 9.2 and described in more detail below, provides a logical sequence of steps from clearly defining a particular skill area to developing specifications for item construction and providing a foundation for an empirically based interpretation of the assessment results. Through this process, the inferences and assumptions about what is to be measured and how the results will be interpreted and reported are explicitly described.

1. *Develop a general definition of the domain.* The first step in this model is to develop a working definition of the domain and the assumptions underlying it. It is this definition that sets the boundaries for what will and will not be measured in a given assessment.
2. *Organize the domain.* Once the definition is developed, it is important to think about the kinds of tasks that represent the skills and abilities included in that

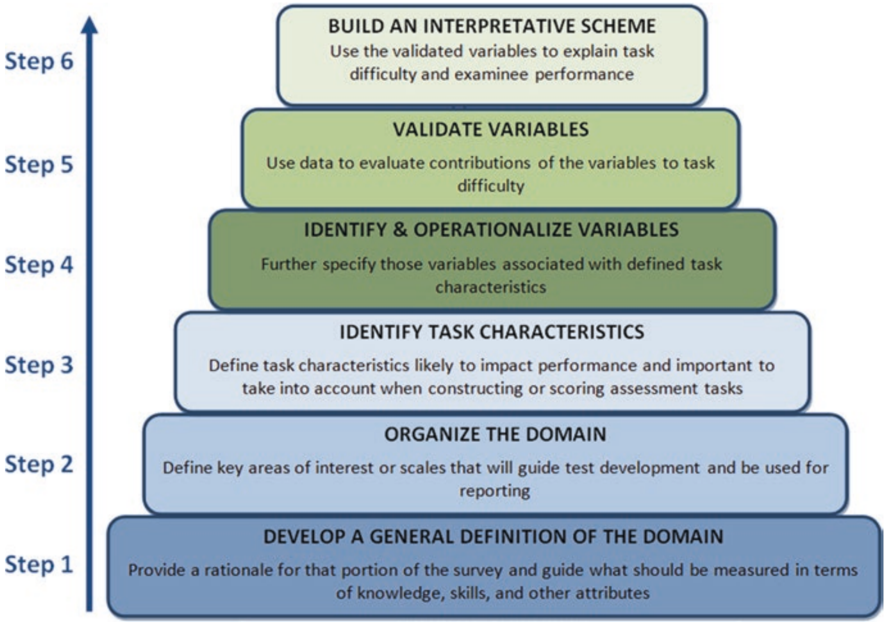

Step 6
BUILD AN INTERPRETATIVE SCHEME
Use the validated variables to explain task difficulty and examinee performance
Step 5
VALIDATE VARIABLES
Use data to evaluate contributions of the variables to task difficulty
Step 4
IDENTIFY & OPERATIONALIZE VARIABLES
Further specify those variables associated with defined task characteristics
Step 3
IDENTIFY TASK CHARACTERISTICS
Define task characteristics likely to impact performance and important to take into account when constructing or scoring assessment tasks
Step 2
ORGANIZE THE DOMAIN
Define key areas of interest or scales that will guide test development and be used for reporting
Step 1
DEVELOP A GENERAL DEFINITION OF THE DOMAIN
Provide a rationale for that portion of the survey and guide what should be measured in terms of knowledge, skills, and other attributes

Fig. 9.2 Model for construct-based assessment

definition. Those tasks must then be categorized in relation to the construct definition to inform test design and result in meaningful score reporting. This step makes it possible to move beyond a laundry list of tasks or skills to a coherent representation of the domain that will permit policy makers and others to summarize and report information in more useful ways.

3. *Identify task characteristics*. Step 3 involves identifying a set of key characteristics, or task models, which will be used in constructing tasks for the assessment. These models may define characteristics of the stimulus materials to be used as well as characteristics of the tasks presented to examinees. Examples of key task characteristics that have been employed throughout the adult literacy assessments include contexts, material types, and information-processing demands.
4. *Identify and operationalize 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 need to be defined. These definitions are based on the existing literature and on experience with building and conducting other large-scale assessments. Defining the variables allows item developers to categorize the materials with which they are working, as well as the questions and directives they construct, so that these categories can be used in the reporting of the results. In the literacy assessments, for example, *context* has been defined to include home and family, health and safety, community and citizenship, consumer economics, work, leisure, and recreation; *materials* have been divided into continuous and noncontinuous texts with each of those categories being further specified; and *processes* have been identified in terms of type of match (focusing on the match between a question and text and including locating, integrating and generating strategies), type of information requested (ranging from concrete to abstract), and plausibility of distractors.[2]
5. *Validate variables*. In Step 5, research is conducted to validate the variables used to develop the assessment tasks. Statistical analyses determine which of the variables account for large percentages of the variance in the difficulty distribution of tasks and thereby contribute most towards understanding task difficulty and predicting performance. In the literacy assessments, this step provides empirical evidence that a set of underlying process variables represents the skills and strategies involved in accomplishing various kinds of literacy tasks.
6. *Build an interpretative scheme.* Finally in Step 6, an interpretative scheme is built that uses the validated variables to explain task difficulty and examinee performance. The definition of proficiency levels to explain performance along the literacy scales is an example of such an interpretative scheme. As previously explained, each scale in the literacy assessments has been divided into five progressive levels characterized by tasks of increasing complexity, as defined by the underlying information processing demands of the tasks. This scheme has been used to define what scores along a particular scale mean and to describe the survey results. Thus, it contributes to the construct validity of inferences based

[2] See Kirsch (2001) and Murray et al. (1997) for a more detailed description of the variables used in the IALS and subsequent assessments.

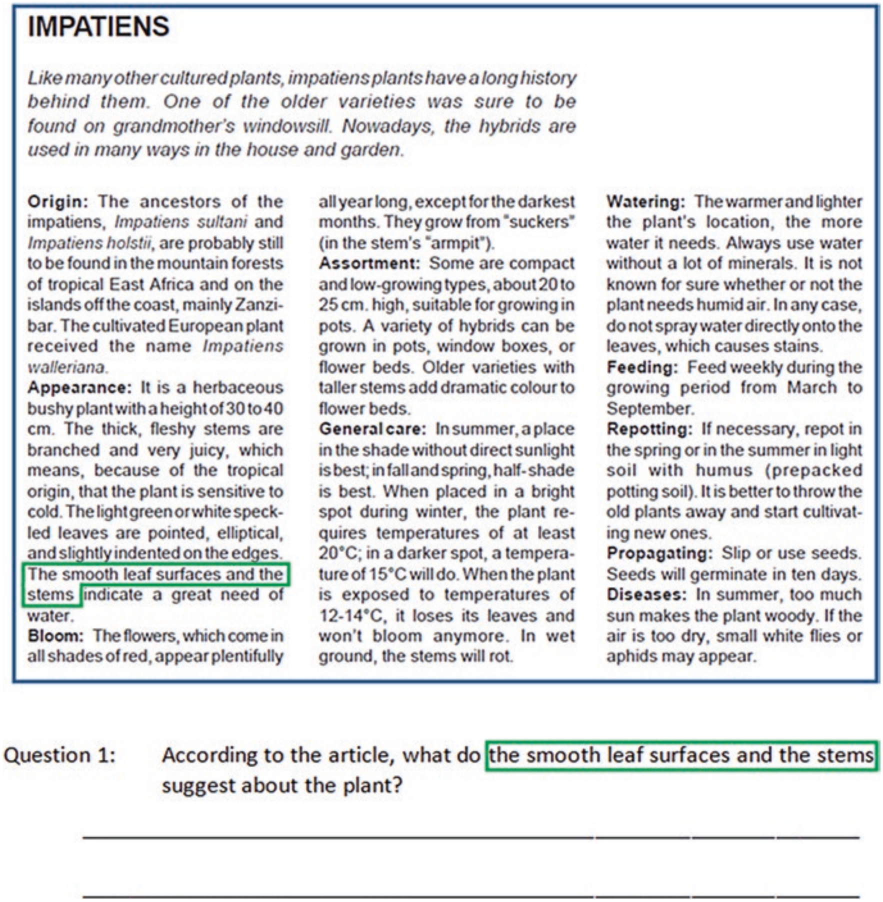
on scores from the measure (Messick 1989). Data from the surveys' background questionnaires have demonstrated consistent correlations between the literacy levels and social and economic outcomes, providing additional evidence for the validity of this particular scheme.

Advancing Messick's approach to construct-based assessment through the application of this framework development model has been one important contribution of the large-scale literacy surveys. This approach not only was used for each of these literacy assessments, but also has become an accepted practice in other assessment programs including the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Achievement (PISA) and the United Nations Educational, Scientific, and Cultural Organization's (UNESCO's) Literacy Assessment and Monitoring Programme (LAMP).

Employing this model across the literacy assessments both informed the test development process and allowed ETS researchers to explore variables that explained differences in performance. Research based on data from the early adult literacy assessments led to an understanding of the relationship between the print materials that adults use in their everyday lives and the kinds of tasks they need to accomplish using such materials. Prior difficulty models for both assessments and learning materials tended to focus on the complexity of stimulus materials alone. ETS's research focused on both the linguistic features and the structures of prose and document materials, as well as a range of variables related to task demands.

Analyses of the linguistic features of stimulus materials first identified the important distinction between continuous and noncontinuous texts. Continuous texts (the prose materials used in the assessments) are composed of sentences that are typically organized into paragraphs. Noncontinuous texts (document materials) are more frequently organized in a matrix format, based on combinations of lists. Work by Mosenthal and Kirsch (1991) further identified a taxonomy of document structures that organized the vast range of matrix materials found in everyday life—television schedules, checkbook registers, restaurant menus, tables of interest rates, and so forth—into six structures: simple, combined, intersecting, and nested lists; and charts and graphs. In prose materials, analyses of the literacy data identified the impact of features such as the presence or absence of graphic organizers including headings, bullets, and bold or italicized print.

On the task side of the difficulty equation, these analyses also identified strategies required to match information in a question or directive with corresponding information in prose and document materials. These strategies—locate, cycle, integrate, and generate—in combination with text features, helped explain what made some tasks more or less difficult than others (Kirsch 2001). For example, locate tasks were defined as those that required respondents to match one or more features of information stated in the question to either identical or synonymous information in the stimulus. A locate task could be fairly simple if there was an exact match between the requested information in the question or directive and the wording in the stimulus and if the stimulus was relatively short, making the match easy to find.



IMPATIENS

Like many other cultured plants, impatiens plants have a long history behind them. One of the older varieties was sure to be found on grandmother's windowsill. Nowadays, the hybrids are used in many ways in the house and garden.

Origin: The ancestors of the impatiens, *Impatiens sultani* and *Impatiens holstii*, are probably still to be found in the mountain forests of tropical East Africa and on the islands off the coast, mainly Zanzibar. The cultivated European plant received the name *Impatiens walleriana*.
Appearance: It is a herbaceous bushy plant with a height of 30 to 40 cm. The thick, fleshy stems are branched and very juicy, which means, because of the tropical origin, that the plant is sensitive to cold. The light green or white speckled leaves are pointed, elliptical, and slightly indented on the edges. The smooth leaf surfaces and the stems indicate a great need of water.
Bloom: The flowers, which come in all shades of red, appear plentifully all year long, except for the darkest months. They grow from "suckers" (in the stem's "armpit").
Assortment: Some are compact and low-growing types, about 20 to 25 cm. high, suitable for growing in pots. A variety of hybrids can be grown in pots, window boxes, or flower beds. Older varieties with taller stems add dramatic colour to flower beds.
General care: In summer, a place in the shade without direct sunlight is best; in fall and spring, half-shade is best. When placed in a bright spot during winter, the plant requires temperatures of at least 20°C; in a darker spot, a temperature of 15°C will do. When the plant is exposed to temperatures of 12-14°C, it loses its leaves and won't bloom anymore. In wet ground, the stems will rot.
Watering: The warmer and lighter the plant's location, the more water it needs. Always use water without a lot of minerals. It is not known for sure whether or not the plant needs humid air. In any case, do not spray water directly onto the leaves, which causes stains.
Feeding: Feed weekly during the growing period from March to September.
Repotting: If necessary, repot in the spring or in the summer in light soil with humus (prepacked potting soil). It is better to throw the old plants away and start cultivating new ones.
Propagating: Slip or use seeds. Seeds will germinate in ten days.
Diseases: In summer, too much sun makes the plant woody. If the air is too dry, small white flies or aphids may appear.

Question 1: According to the article, what do the smooth leaf surfaces and the stems suggest about the plant?

____

____

Fig. 9.3 Sample prose task

As an example, see Fig. 9.3. Here there is an exact match between "the smooth leaf surfaces and the stems" in the question and in the last sentence in the second paragraph of the text.

Analyses showed that the difficulty of locate tasks increased when stimuli were longer and more complex, making the requested information more difficult to locate; or when there were distractors, or a number of plausible correct answers, within the text. Difficulty also increased when requested information did not exactly match the text in the stimulus, requiring respondents to locate synonymous information. By studying and defining the interaction between the task demands for locate, cycle, integrate, and generate tasks and features of various stimuli, the underlying information-processing skills could be more clearly understood. This research allowed for improved assessment design, increased interpretability of results, and development of derivative materials, including individual assessments[3] and instructional materials.[4]

In 1994, the literacy assessments moved from a national to an international focus. The primary goal of the international literacy assessments—International Adult Literacy Survey (IALS), ALL, and PIAAC—was to collect comparable international data that would provide a broader understanding of literacy across industrialized nations.

One challenge in meeting the goal of ensuring comparability across different national versions of the assessment was managing the translation process. Based on the construct knowledge gained from earlier assessments, it was clear that translators had to understand critical features of both the stimulus materials and the questions. Training materials and procedures were developed to help translators and project managers from participating countries reach this understanding. For example, the translation guidelines for the content shown in Fig. 9.3 specified the following:

- Translation must maintain literal match between the key phrase “the smooth leaf surfaces and the stems” in the question and in the last sentence in the second paragraph of the text.
- Translation must maintain a synonymous match between *suggest* in question and *indicate* in text.

Understanding task characteristics and the interaction between questions and stimulus materials allowed test developers to create precise translation guidelines to ensure that participating countries developed comparable versions of the assessment instruments. The success of these large-scale international efforts was in large part possible because of the construct knowledge gained from ETS research based on the results of earlier national assessments.

9.3 Expanding and Implementing Large-Scale Assessment Methodology

The primary purpose of the adult literacy large-scale assessments has been to describe the distribution of literacy skills in populations, as well as in subgroups within and across populations. The assessments have not targeted the production of

[3] These individual assessments include the Test of Applied Literacy Skills (TALS), a paper-and-pencil assessment with multiple forms; the *PDQ Profile*™ Series, an adaptive computer-based assessment of literacy proficiency; and the Health Activities Literacy Test, an adaptive computer-based assessment of literacy tasks focusing on health issues.

[4] Using information from this research, ETS developed P.D.Q. Building Skills for Using Print in the early 1990s. This multi-media, group-based system includes more than 100 h of instruction focusing on prose, document, and quantitative literacy, as well as workbooks and instructional support materials.

scores for individual test takers, but rather employed a set of specialized design principles and statistical tools that allow a reliable and valid description of skill distributions for policy makers and other stakeholders. To describe skills in a comparable manner in international contexts, the methodologies utilized needed to ensure that distributions were reported in terms of quantities that describe differences on scales across subgroups in meaningful ways for all participating entities.

The requirement to provide comparable estimates of skill distributions has been met by using the following methodological tools:

- Models that allow the derivation of comparable measures across populations and comparisons across literacy assessments
- Survey methodologies that provide representative samples of respondents
- Procedures to ensure scoring accuracy and to handle missing data
- Forward-looking designs that take advantage of context information in computer-based assessments

Taken together, these methodological tools facilitate the measurement goal of providing reliable, valid, and comparable estimates of skill distributions based on large-scale literacy assessments.

9.3.1 Models Allowing the Derivation of Comparable Measures and Comparisons Across Literacy Assessments

The goal of the literacy assessments discussed here has been to provide a description of skills across a broad range of ability, particularly given that the assessments target adults who have very different educational backgrounds and a wider range of life experiences than school-based populations. Thus the assessments have needed to include tasks that range from very easy to very challenging. To enable comparisons across a broad range of skill levels and tasks, the designs for all of the adult literacy assessments have used "incomplete block designs". In such designs, each sampled individual takes a subset of the complete assessment. The method of choice for the derivation of comparable measures in incomplete block designs is based on measurement models that were developed for providing such measures in the analyses of test data (Lord 1980; Rasch 1960). These measurement models are now typically referred to as item response theory (IRT) models (Lord and Novick 1968).

IRT models are generally considered superior to simpler approaches based on sum scores, particularly in the way omitted responses and incomplete designs can be handled. Because IRT uses the full information contained in the set of responses, these models are particularly useful for assessment designs that utilize a variety of item types arranged in blocks that cannot be set up to be parallel forms of a test. Incomplete block designs do not allow the comparison of sum scores of aggregated responses because different blocks of items may vary in difficulty and even in the number of items. IRT models establish a comparable scale on which items from different blocks, and from respondents taking different sets of items, can be located, even in sparse incomplete designs. These models are powerful tools to evaluate whether the information provided for each individual item is comparable across populations of interest (see, for example, Yamamoto and Mazzeo 1992). In particular, the linking procedures typically used in IRT have been adapted, refined, and generalized for use in international assessments of adult literacy. More specifically, recent developments in IRT linking methods allow a more flexible approach to the alignment of scales that takes into account local deviations (Glas and Verhelst 1995; Yamamoto 1998; von Davier and von Davier 2007; Oliveri and von Davier 2011; Mazzeo and von Davier 2014; Glas and Jehangir 2014). The approach applied in IALS, ALL and PIAAC enables international assessments to be linked across a large number of common items while allowing for a small subset of items in each country to function somewhat differently to eliminate bias due to occasional item-by-country interactions. IRT has been the measurement method of choice not only for ETS's adult literacy assessments, but also for national and international assessments of school-age students such as the National Assessment of Educational Progress (NAEP), PISA, and Trends in International Mathematics and Science Study (TIMSS).

The integration of background information is a second important characteristic of the analytical methodologies used in the adult literacy assessments. Background data are used for at least two purposes in this context. First and foremost, they provide information about the relationship between demographic variables and skills. This makes it possible to investigate how the distribution of skills is associated with variables including educational attainment, gender, occupation, and immigration status of groups. These are among the variables needed to answer questions that are of interest to policy makers and other stakeholders, such as, "How are skills distributed in immigrant vs. nonimmigrant populations?" and "What is the relationship between literacy skills and measures of civic engagement such as voting?" In addition, background data provide auxiliary information that can be used to improve the precision of the skills measurement. This use of background data is particularly important because the available background data can help alleviate the effects of limited testing time for respondents by using the systematic differences between groups of respondents to strengthen the estimation of skills.[5]

While one of the main aims of ETS's large-scale literacy assessments has been to provide data on human capital at any given point in time, the extent to which skills change over time is also of fundamental interest. IRT models provide a powerful tool to link assessments over cycles conducted in different years. In much the same way that IRT allows linking of scales and provides comparable measures across blocks of different items within an assessment, and across countries, IRT can also be used to link different assessments over time. This link is only possible because significant efforts have been made across the literacy assessments to collect data in a manner that supports reusing sets of items over time while regularly renew-

[5] The interested reader is referred to Mislevy et al. (1992) for a description of this approach and to von Davier et al. (2006) for an overview and a description of recent improvements and extensions of the approach.

ing the item pool. The particular design principles applied ensure that new and previously used blocks of items are combined into test booklets in such a way that each assessment is also connected to multiple assessments over time. Because IRT estimation methods have been developed and extended to facilitate analyses of incomplete designs, these methods are particularly well suited to analyze multiple links across assessments. Statistical tools can be used to evaluate whether the items used repeatedly in multiple assessments are indeed comparable across assessments from different years and provide guidance as to which items to retain and which parts of the assessment have to be renewed by adding new task material.

9.3.2 Survey Methodologies That Provide Representative Samples of Respondents

The description of populations with respect to policy-relevant variables requires that members of the population of interest are observed with some positive probability. While it is not a requirement (or possibility) to assess every individual, a representative sample has to be drawn in order to provide descriptions of populations without bias. The adult literacy assessments have typically used methods common to household surveys, in which either a central registry of inhabitants or a list of addresses of dwellings/households of a country is used to randomly draw a representative random sample of respondents. This list is then used to select an individual at random, get in contact with those selected and ask the selected individual to participate in the survey. To account for unequal chances of being selected, the use of sampling weights is necessary. The importance of sampling and weighting for an accurate estimate of skill distributions is discussed in more detail in contributions summarizing analytic strategies involving sampling and weights for large-scale assessments by Rust (2014) and Rutkowski et al. (2010).

One particular use of these survey methodologies in large-scale assessments, and a contribution of ETS's adult assessments, is the projection of skill distributions based on expected changes in the population. The report, *America's Perfect Storm: Three Forces Changing Our Nation's Future* (Kirsch et al. 2007) shows how evidence regarding skill distributions in populations of interest can be projected to reflect changes in those populations, allowing a prediction of the increase or decline of human capital over time.

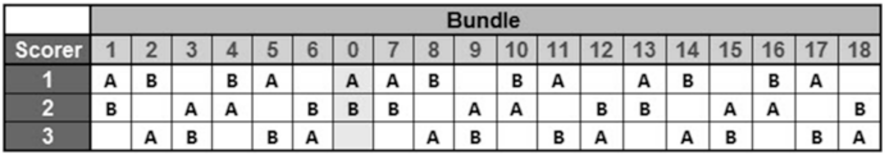
9.3.3 Procedures to Ensure Scoring Accuracy

One measurement issue that has been addressed in large-scale literacy assessments is the need to ensure that paper-and-pencil (as well as human-scored computer-based) tasks are scored accurately and reliably, both within and across countries participating in the international surveys. Many of the assessment tasks require respondents to provide short, written responses that typically range in length from single-word responses to short phrases or sentences. Some tasks ask for responses to be marked on the stimulus. On paper, respondents may be asked to circle or underline the correct answer whereas on the computer, respondents may be required to mark or highlight the response using the mouse or another input device. So while responses are typically quite short, scorers in all participating countries must follow a well-developed set of scoring rules to ensure consistent scoring. All of the adult literacy surveys prior to PIAAC were conducted as paper-and-pencil assessments, scored by national teams of trained scorers. While PIAAC is largely a computer-based assessment using automated scoring, a paper-and-pencil component has been retained, both to strengthen the link between modes and to provide an option for respondents without the requisite technical skills to complete the assessment on the computer. To ensure reliable and comparable data in all of the adult literacy surveys, it was critical that processes were developed to monitor the accuracy of human scoring for the short constructed responses in that mode within a country, across countries, and across assessments over time.

Without accurate, consistent and internationally comparable scoring of paper-and-pencil items, all subsequent psychometric analyses of these items would be severely jeopardized. For all of the large-scale adult literacy assessments, the essential activities associated with maintaining scoring consistency have been basically the same. Having items scored independently by two different scorers and then comparing the resulting scores has been the key required procedure for all participating countries. However, because the number of countries and number of languages has increased with each international assessment, the process has been refined over time. In IALS, the procedure used to ensure standardized scoring involved an exchange of booklets across countries with the same or similar languages. Country A and Country B thus would score their own booklets; then Country A would second score Country B’s booklets and vice versa. In cases where a country could not be paired with another testing in the same language, the scorers within one country would be split into two independent groups, and booklets would be exchanged across groups for rescoring.

Beginning with ALL, the use of anchor booklets was introduced. This common set of booklets was prepared by test developers and distributed to all countries. Item responses in these booklets were based on actual responses collected in the field as well as responses that reflected key points on which scorers were trained. Because responses were provided in English, scoring teams in each country designated two bilingual scorers responsible for the double-scoring process. Anchor booklets were used in PIAAC as well. The new aspect introduced in PIAAC was the requirement that countries follow a specified design to ensure that each booklet was scored twice and that scorers functioned both as first and second scorer across all of the booklets. Figure 9.4 shows the PIAAC design for countries that employed three scorers. The completed booklets were divided up into 18 bundles of equal size. Bundle 0 was the set of anchor booklets to be scored by bilingual Scorers 1 and 2.

In an ideal world, the results of these double-scoring procedures would confirm that scoring accuracy was 100% and that scorers were perfectly consistent with each



	Bundle																		
Scorer	1	2	3	4	5	6	0	7	8	9	10	11	12	13	14	15	16	17	18
1	A	B		B	A		A	A	B		B	A		A	B		B	A	
2	B		A	A		B	B	B		A	A		B	B		A	A		B
3		A	B		B	A			A	B		B	A		A	B		B	A

Fig. 9.4 Double-scoring design for PIAAC. Cells marked with "A" represent the first scorer for each bundle

other. Although this level of consistency is never obtained due to random deviations, scoring accuracy in the adult literacy surveys tends to be around 96%.

When scoring discrepancies occur, experience has shown that they fall into two distinct classes. The first type of discrepancy reveals a consistent bias on the part of one scorer, for example when one scorer is consistently more lenient than others. Because countries are required to send rescoring data for analysis at set points during the scoring process, when this situation is found, problematic scorers must be retrained or, in some cases, dismissed.

The second type of discrepancy that can be revealed through analysis of the rescoring data is more challenging to address. This occurs when the scoring results reveal general inconsistencies between the scorers, with no pattern that can be attributed to one scorer or the other. This issue has been relatively rare in the adult literacy assessments. When it has occurred, it is generally the result of a problem with an item or an error in the scoring guides. One procedure for addressing this situation includes conducting a review of all inconsistently scored responses to determine if there is a systematic pattern and, if one is found, having those items rescored. Additionally, the scoring guides for such items can be revised to clarify any issue identified as causing inconsistent scoring. When a specific problem cannot be identified and resolved, model based adjustments such as assigning unique item parameters to account for this type of country-by-item deviation may be required for one or more countries to reflect this ambiguity in scoring.

9.3.4 Statistical Procedures for Handling Missing Data

A second key methodological issue developed through experience with the large-scale literacy assessments involves the treatment of missing data due to nonresponse. Missing responses reduce the amount of information available in the cognitive assessment and thus can limit the kinds of inferences that can be made about the distribution of skills in the population based on a given set of respondents. More specifically, the relationship between skills and key background characteristics is not measured well for respondents with a high proportion of item nonresponse. This issue has been addressed in the large-scale literacy assessments by estimating conditioning coefficients based on the performance of respondents with sufficient cognitive information and applying the parameters to those respondents for whom there is insufficient performance data. This solution allows stable estimation of the model and ensures that regression of performance data on background variables is based on cases that provide sufficiently accurate information.

The two most common but least desirable ways to treat missing cases are a) to ignore them and b) to assume all missing responses can be equated to incorrect responses. Ignoring missing responses is acceptable if one can assume that missing cases occur at random and that the remaining observed cases are representative of the target population. In this case, the result would be slightly larger standard errors due to reduced sample size, and the other estimates would remain unbiased. Randomly missing data rarely occur in real data collections, however, especially in surveys of performance. If the incidence of nonresponse varies for major subgroups of interest, or if the missing responses are related to the measurement objective— in this case, the measurement of literacy skills—then inferring the missing data from observed patterns results in biased estimates. If one can be sure that all missingness is due to a lack of skill, the treatment as incorrect is justified. This treatment may be appropriate in high-stakes assessments that are consequential for respondents. In surveys, however, the respondent will not be subjected to any consequences, so other reasons for missingness, such as a lack of motivation, may be present.

To address these issues, different approaches have been developed. In order to infer reasons for nonresponse, participants are classified into two groups based on standardized coding schemes used by interviewers to record reasons for nonparticipation: those who stop the assessment for literacy-related issues (e.g., reading difficulty, native language other than language of the assessment, learning disability) and those who stop for reasons unrelated to literacy (e.g., physical disability, refusal for unspecified reason). Special procedures are used to impute the proficiencies of individuals who complete fewer than the minimum number of tasks needed to estimate their proficiencies directly.

When individuals cite a literacy-related reason for not completing the cognitive items, this implies that they were unable to respond to the items. On the other hand, citing a reason unrelated to literacy implies nothing about a person's literacy proficiency. When an individual responds to fewer than five items per scale— the minimum number needed to directly estimate proficiencies—cases are treated as follows:

- If the individual cited a literacy-related reason for not completing the assessment, then all consecutively missing responses at the end of a block of items are scored as wrong.
- If the individual cited a reason unrelated to literacy, then all consecutively missing responses at the end of block are treated as not reached.

A respondent's proficiency is calculated from a posterior distribution that is the product of two functions: a conditional distribution of proficiency, given responses to the background questionnaire; and a likelihood function of proficiency, given responses to the cognitive items (see Murray et al. 1997, for more detail). By scoring missing responses as incorrect for individuals citing literacy-related reasons for stopping the assessment, the likelihood function is very peaked at the lower end of the scale—a result that is believed to accurately represent their proficiency.

Because PIAAC was a computer-based assessment, information was available to further refine the scoring rules for non-response. The treatment of item level missing data in paper-and-pencil assessments largely has to rely on the position of items. In order to define the reason for not responding as either volitional or being based on having never been exposed to (not reached) the items, the location of the 'last' item for which a response was observed is crucial. In computer-based assessments, non-response can be treated in a more sophisticated way by taking timing data and process information into account. While the problem of rapid guessing has been described in high-stakes assessment (Wise and DeMars 2005), the nature of literacy surveys does not compel respondents to guess, but rather to skip an item rapidly for some reasons that may be unrelated to skills, for example perceived time pressure or a lack of engagement. If an item was skipped in this way – a rapid move to the next item characterized by a very short overall time spent on the item (e.g., less than 5 s) and the minimal number of actions sufficient to 'skip' the item, PIAAC applied a coding of 'not reached/not administered' (OECD 2013; Weeks et al. 2014). If, however a respondent spent time on an item, or showed more than the minimum number of actions, a missing response would be assumed to be a volitional choice and counted as not correct.

9.3.5 Forward-Looking Design for Using Context Information in Computer-Based Assessments

The methodologies used in large-scale assessments are well developed, and variants of essentially these same methodologies are used in all major large-scale literacy assessments. While this repeated use implies that the current methodology is well suited for the analyses of assessments at hand, new challenges have arisen with the advent of PIAAC.

As a computer-based assessment, PIAAC presents two important advantages—and challenges—when compared to earlier paper-and-pencil assessments. First is the wealth of data that a computer can provide in terms of process information. Even seemingly simple information such as knowing precisely how much time a respondent spent on a particular item can reveal important data that were never available in the paper-and-pencil assessments. The use of such data to refine the treatment of non-response data, as described above, is one example of how this information can improve measurement. Second is the opportunity to design adaptive assessments that change the selection of items depending on a respondent's performance on previous sets of items. These differences result in both new sources of information about the performance of respondents and a change in the structure of the cognitive response data given that not all test takers respond to the same set of items.

Modern psychometric methodologies are available that can improve estimation in the face of such challenges. Such methods can draw upon process and navigation data to classify respondents (Lazarsfeld and Henry 1968) with respect to the typical paths they take through scenario-based tasks, such as the ones in PIAAC's problem-solving domain. Extensions of IRT models can reveal whether this or other types of classifications exist besides the skills that respondents apply (Mislevy and Verhelst 1990; Rost 1990; von Davier and Carstensen 2007; von Davier and Rost 1995; von Davier and Yamamoto 2004; Yamamoto 1989). Additional information such as response latency can be used to generate support variables that can be used for an in-depth analysis of the validity of responses. Rapid responders (DeMars and Wise 2010) who may not provide reliable response data can potentially be identified using this data. Nonresponse models (Glas and Pimentel 2008; Moustaki and Knott 2000; Rose et al. 2010) can be used to gain a deeper understanding of situations in which certain types of respondents tend not to provide any data on at least some of the items. Elaborate response-time models that integrate latency and accuracy (Klein Entink et al. 2009; Lee 2008) can be integrated with current large-scale assessment methodologies.

9.4 Linking Real-Life Stimulus Materials and Innovative Item Types

From the first adult literacy assessment onward, items have been based on everyday materials taken from various adult situations and contexts including the workplace, community, and home. In the 1993 National Adult Literacy Survey (NALS), for example, sets of open-ended questions required respondents to use a six-page newspaper that had been created from articles, editorials, and advertisements taken from real newspapers. In PIAAC, simulation tasks were based on content from real websites, advertisements, and e-mails. For each of the large-scale literacy assessments, original materials were used in their entirety, maintaining the range of wording, formatting, and presentation found in the source. The inclusion of real-life materials both increased the content validity of the assessments and improved respondent motivation, with participants commenting that the materials were both interesting and appropriate for adults.

Each of the large-scale literacy assessments also used open-ended items. Because they are not constrained by an artificial set of response options, these open-ended tasks allowed respondents to engage in activities that are similar to those they might perform if they encountered the materials in real life. In the paper-and-pencil literacy assessments, a number of different open-ended response types were employed. These included asking respondents to underline or circle information in the stimulus, copy or paraphrase information in the stimulus, generate a response, and complete a form.

With the move to computer-based tasks in PIAAC, new ways to collect responses were required. The design for PIAAC called for the continued use of open-ended response items, both to maintain the real-life simulation focus of the assessment and to maintain the psychometric link between PIAAC and prior surveys. While the paper-and-pencil surveys allowed respondents to compose answers ranging from a word or two to several sentences, the use of automated scoring for such responses was not possible, given that PIAAC was delivered in 33 languages. Instead, the response modes used for computer-based items in this assessment included highlighting, clicking, and typing numeric responses—all of which could be scored automatically. Throughout previous paper-and-pencil assessments, there had always been some subset of respondents who marked their responses on the stimulus rather than writing answers on the provided response lines. These had been considered valid answers, and scoring rubrics had been developed to train scorers on how such responses should be scored. Thus electronic marking of text by highlighting a phrase or sentence or clicking on a cell in a table fit within existing scoring schemes. Additionally, previous work on a derivative computer-based test for individuals, the PDQ Profile Series, had shown that item parameters for paper-and-pencil items adapted from IALS and ALL were not impacted when those items were presented on the computer and respondents were asked to highlight, click, or type a numeric response. PIAAC thus became the first test to employ these response modes on a large scale and in an international context.

Taking advantage of the computer-based context, PIAAC also introduced new types of simulation items. In reading literacy, items were included that required respondents to use scrolling and hyperlinks to locate text on a website or provide responses to an Internet poll. In the new problem-solving domain, tasks were situated in simulated web, e-mail, and spreadsheet environments that contained common functionality for these environments. Examples of these new simulation tasks included items that required respondents to access information in a series of e-mails and use that information to schedule meeting rooms via an online reservation system or to locate requested information in a complex spreadsheet where the spreadsheet environment included “find” and “sort” options that would facilitate the task.

In sum, by using real-life materials and open-ended simulation tasks, ETS’s large-scale literacy assessments have sought to reflect and measure the range of literacy demands faced by adults in order to provide the most useful information to policy makers, researchers, and the public. Over time, the nature of the assessment materials and tasks has been expanded to reflect the changing nature of literacy as the role of technology has become increasingly prevalent and important in everyday life.

9.5 Developing Extensive Background Questionnaires to Link Performance With Experience and Outcome Variables

One important goal of the large-scale literacy assessments has been to relate skills to a variety of demographic characteristics and explanatory variables. Doing so has allowed ETS to investigate how performance is related to social and educational outcomes and thereby interpret the importance of skills in today’s society. It has also enhanced our understanding of factors related to the observed distribution of literacy skills across populations and enabled comparisons with previous surveys.

For each of the literacy assessments, respondents completed a background questionnaire in addition to the survey's cognitive measures. The background questions were a significant component of each survey, taking up to one-third of the total survey time. In each survey, the questionnaire addressed the following broad issues:

- General language background
- Educational background and experience
- Labor force participation
- Literacy activities (types of materials read and frequency of use for various purposes)
- Political and social participation
- Demographic information

As explained earlier, information collected in the background questionnaires is used in the psychometric modeling to improve the precision of the skills measurement. Equally importantly, the background questionnaires provide an extensive database that has allowed ETS to explore questions such as the following: What is the relationship between literacy skills and the ability to benefit from employer-supported training and lifelong learning? How are educational attainment and literacy skills related? How do literacy skills contribute to health and well being? What factors may contribute to the acquisition and decline of skills across age cohorts? How are literacy skills related to voting and other indices of social participation? How do reading practices affect literacy skills?

The information collected via the background questionnaires has allowed researchers and other stakeholders to look beyond simple demographic information and examine connections between the skills being measured in the assessments and important personal and social outcomes. It has also led to a better understanding of factors that mediate the acquisition or decline of skills. At ETS, this work has provided the foundation for reports that foster policy debate on critical literacy issues. Relevant reports include Kirsch et al. (2007), Rudd et al. (2004) and Sum et al. (2002, 2004).

9.6 Establishing Innovative Reporting Procedures to Better Integrate Research and Survey Data

Reports for each of the large-scale surveys have gone beyond simply reporting distributions of scores on the assessment for each participating country. As noted above, using information from the background questionnaire has made it possible to link performance to a wide range of demographic variables. Other reporting innovations have been implemented to make the survey data more useful and understandable for policy makers, researchers, practitioners, and other stakeholders.

The PIAAC data, conjointly with IALS and ALL trend data, are available in the Data Explorer (http://piaacdataexplorer.oecd.org/ide/idepiaac/), an ETS-developed web-based analysis and reporting tool that allows users to query the PIAAC database and produce tabular and graphical summaries of the data. This tool has been designed for a wide range of potential users, including those with little or no statistical background. By selecting and organizing relevant information, stakeholders can use the large-scale data to address questions of importance to them.

In addition to linking performance and background variables, survey reports have also looked at the distribution of literacy skills and how performance is related to underlying information-processing skills. Reports have included item maps that present sample items in each domain, showing where these items performed on the literacy scale and discussing features of the stimuli and questions that impact difficulty. Such analyses have allowed stakeholders to understand how items represent the construct and thereby allow them to generalize beyond the pool of items in any one assessment. These reports were also designed to provide readers with a better understanding of the information-processing skills underlying performance. Such an understanding has important implications for intervention efforts.

9.7 Conclusion

During the 30 years over which the six large-scale adult literacy assessments have been conducted, literacy demands have increased in terms of the types and amounts of information adults need to manage their daily lives. The goal of the assessments has been to provide relevant information to the variety of stakeholders interested in the skills and knowledge adults have and the impact of those skills on both individuals and society in general. Meeting such goals in this ever-changing environment has required that ETS take a leading role in the following:

- Expanding the construct of literacy
- Developing a model for building construct-based assessments
- Expanding and implementing large-scale assessment methodology to ensure reliable, valid, and comparable measurement across countries and over time
- Taking an approach to test development that focuses on the use of real-life materials and response modes that better measure the kinds of tasks adults encounter in everyday life[6]
- Developing extensive background questionnaires that make it possible to link performance with experience and outcome variables, thereby allowing the survey data to address important policy questions
- Developing reporting procedures that better integrate survey data with research

These efforts have not just expanded knowledge of what adults know and can do; they have also made important contributions to understanding how to design, conduct, and report the results of large-scale international assessments.

[6] Sample PIAAC items are available at http://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm.

Appendix: Description of the Five Levels for Prose, Document, and Numeracy Domains

	Prose	Document	Numeracy
Level 1 (0–225)	Most of the tasks in this level require the respondent to read a relatively short text to locate a single piece of information that is identical to or synonymous with the information given in the question or directive. If plausible but incorrect information is present in the text, it tends not to be located near the correct information.	Tasks in this level tend to require the respondent either to locate a piece of information based on a literal match or to enter information from personal knowledge onto a document. Little, if any, distracting information is present.	Tasks in this level require the respondent to show an understanding of basic numerical ideas by completing simple tasks in concrete, familiar contexts where the mathematical content is explicit with little text. Tasks consist of simple, one-step operations such as counting, sorting dates, performing simple arithmetic operations, or understanding common and simple percentages such as 50%.
Level 2 (226–275)	Some tasks in this level require respondents to locate a single piece of information in the text; however, several distractors or plausible but incorrect pieces of information may be present, or low-level inferences may be required. Other tasks require the respondent to integrate two or more pieces of information or to compare and contrast easily identifiable information based on a criterion provided in the question or directive.	Tasks in this level are more varied than those in level 1. Some require the respondents to match a single piece of information; however, several distractors may be present, or the match may require low-level inferences. Tasks in this level may also ask the respondent to cycle through information in a document or to integrate information from various parts of a document.	Tasks in this level are fairly simple and relate to identifying and understanding basic mathematical concepts embedded in a range of familiar contexts where the mathematical content is quite explicit and visual with few distractors. Tasks tend to include one-step or two-step processes and estimations involving whole numbers, interpreting benchmark percentages and fractions, interpreting simple graphical or spatial representations, and performing simple measurements.

(continued)

	Prose	Document	Numeracy
Level 3 (276–325)	Tasks in this level tend to require respondents to make literal or synonymous matches between the text and information given in the task, or to make matches that require low-level inferences. Other tasks ask respondents to integrate information from dense or lengthy text that contains no organizational aids such as headings. Respondents may also be asked to generate a response based on information that can be easily identified in the text. Distracting information is present but is not located near the correct information.	Some tasks in this level require the respondent to integrate multiple pieces of information from one or more documents. Others ask respondents to cycle through rather complex tables or graphs that contain information that is irrelevant or inappropriate to the task.	Tasks in this level require the respondent to demonstrate understanding of mathematical information represented in a range of different forms, such as in numbers, symbols, maps, graphs, texts, and drawings. Skills required involve number and spatial sense; knowledge of mathematical patterns and relationships; and the ability to interpret proportions, data, and statistics embedded in relatively simple texts where there may be distractors. Tasks commonly involve undertaking a number of processes to solve problems.
Level 4 (326–375)	These tasks require respondents to perform multiple-feature matches and to integrate or synthesize information from complex or lengthy passages. More complex inferences are needed to perform successfully. Conditional information is frequently present in tasks at this level and must be taken into consideration by the respondent.	Tasks in this level, like those at the previous levels, ask respondents to perform multiple-feature matches, cycle through documents, and integrate information; however, they require a greater degree of inference. Many of these tasks require respondents to provide numerous responses but do not designate how many responses are needed. Conditional information is also present in the document tasks at this level and must be taken into account by the respondent.	Tasks at this level require respondents to understand a broad range of mathematical information of a more abstract nature represented in diverse ways, including in texts of increasing complexity or in unfamiliar contexts. These tasks involve undertaking multiple steps to find solutions to problems and require more complex reasoning and interpretation skills, including comprehending and working with proportions and formulas or offering explanations for answers.

(continued)

	Prose	Document	Numeracy
Level 5 (376–500)	Some tasks in this level require the respondent to search for information in dense text that contains a number of plausible distractors. Others ask respondents to make high-level inferences or use specialized background knowledge. Some tasks ask respondents to contrast complex information.	Tasks in this level require the respondent to search through complex displays that contain multiple distractors, to make high-level text-based inferences, and to use specialized knowledge.	Tasks in this level require respondents to understand complex representations and abstract and formal mathematical and statistical ideas, possibly embedded in complex texts. Respondents may have to integrate multiple types of mathematical information, draw inferences, or generate mathematical justification for answers.

References

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10*(3), 207–229. https://doi.org/10.1080/15305058.2010.496347

Glas, C. A. W., & Jehangir, K. (2014). Modeling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 97–116). New York: Chapman & Hall

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907–922. https://doi.org/10.1177/0013164408315262

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–96). New York: Springer.

Kirsch, I. S. (2001). *The international adult literacy survey (IALS): Understanding what was measured* (Research Report No. RR-01-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2001.tb01867.x

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults* (NAEP Report No. 16-PL-01). Princeton: Educational Testing Service.

Kirsch, I. S., Braun, H., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future* (Policy Information Report). Princeton: Educational Testing Service.

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*, 621–640. https://doi.org/10.1348/000711008X374126

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*, 1–15. https://doi.org/10.3758/PBR.15.1.1

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 229–258). New York: Chapman & Hall.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillian.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(1), 13–23. https://doi.org/10.3102/0013189X023002013

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215. https://doi.org/10.1007/BF02295283

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document process. *Discourse Processes, 14*, 147–180. https://doi.org/10.1080/01638539109544780

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A, 163*, 445–459. https://doi.org/10.1111/1467-985X.00177

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (Eds.). (1997). *Adult literacy in OECD countries: Technical report on the first international adult literacy survey*. Washington, DC: National Center for Education Statistics.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264128859-en

OECD. (2013). Technical report of the Survey of Adult Skills (PIAAC). Retrieved from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315–333.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling non-ignorable missing data with IRT* (Research Report No. RR-10-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 3*, 271–282. https://doi.org/10.1177/014662169001400305

Rudd, R., Kirsch, I., & Yamamoto, K. (2004). *Literacy and health in America* (Policy Information Report). Princeton: Educational Testing Service.

Rust, K. (2014). Sampling, weighting, and variance estimation in international large scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 117–154). New York: Chapman & Hall.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*, 142–151. https://doi.org/10.3102/0013189X10363170

Sum, A., Kirsch, I. S., & Taggart, R. (2002). *The twin challenges of mediocrity and inequality: Literacy in the U.S. from an international perspective*. Princeton: Educational Testing Service.

Sum, A., Kirsch, I. S., & Yamamoto, K. (2004). *A human capital concern: The literacy proficiency of U.S. immigrants*. Princeton: Educational Testing Service.

von Davier, M., & Carstensen, C. (Eds.). (2007). *Multivariate and mixture distribution Rasch models*. New York: Springer. https://doi.org/10.1007/978-0-387-49839-3

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–382). New York: Springer. https://doi.org/10.1007/978-1-4612-4230-7_20

von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformation. *Methodology, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis*. Invited lecture at the ETS Spearman invitational conference, Philadelphia, PA.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol.* 26. *Psychometrics* (pp. 1039–1056). Amsterdam: Elsevier.

Weeks, J., von Davier, M., & Yamamoto, K. (2014). Design considerations for the Programme for International Student Assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 259–276). New York: Chapman & Hall.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class model* (Research Report No. RR-89-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K. (1998). Scaling and scale linking. In T. S. Murray, I. S. Kirsch, & L. B. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the First International Adult Literacy Survey* (pp. 161–178). Washington, DC: National Center for Education Statistics.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*, 155–173. https://doi.org/10.2307/1165167

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 2.5 International License (http://creativecommons.org/licenses/by-nc/2.5/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

cc BY NC