Implementation of CBT in the PIAAC Field Test and CAT in the PIAAC

(September 12, 2011)

Kentaro Yamamoto

[Ph.D., Deputy Director, Center for Global Assessment ETS, Princeton NJ]

Education

Ph. D., Educational Psychology, University of Illinois Urbana-Champaign (1987)
M.S., Statistics, University of Illinois Urbana-Champaign (1986)
M.S., Psychology, Portland State University (1983)

Experience

2004-present, Deputy Director, Center for Global Assessment ETS, Princeton, NJ.
1999-present, Principal Research Scientist, ETS, Princeton, NJ.
1995-1999, Senior Research Scientist, ETS, Princeton, NJ.
1987-1995, Research Scientist, Educational Testing Service, Princeton, NJ.
2006-date, Technical Advisor for the PISA, OECD.
2006-date, Designed the PIAAC, responsible for the psychometrics of the PIAAC, OECD.
1998-2005, Designed the ALL, Adult Literacy and Life Skills Survey, responsible for the psychometrics of the ALL.
1992-2000, Designed the IALS, International Adult Literacy Survey, responsible for the psychometrics of the IALS.
1990-1994, Directed the NALS data analyses, National Adult Literacy Survey, responsible for the psychometrics of the NALS.
1987-1991, Directed the NAEP data analyses, National Assessment of Educational Progress, responsible for the science and mathematics of the NAEP.

Selected Publications

Kirsch, I., Braun, H., Yamamoto, K., & Sum, A. (2007) America's Perfect Storm: Three Forces Changing Our Nation's Future, Policy Information Center, Center for Global Assessment, Educational Testing Service, Princeton, NJ.

von Davier, M.,& Yamamoto, K. (2006) Mixture-Distribution and HYBRID Rasch Models, in von Davier, M., & Cartensen, C. (Eds.) *Statistics for Social and Behavioral Sciences* (pp. 99-118). New York: Springer.

von Davier, M.,& Yamamotto, K. (2004) Partially Observed Mixtures of IRT Models: an Extension of the Generalized Partial-Credit Model, *Applied Psychological Measurement*, 28,6,389-406.

Sum, A., Kirsch, I., & Yamamoto, K. (2004) Literacy Proficiencies and Labor Market Success: Key Findings from National and International Surveys, Policy Information Center, Center for Global Assessment, Educational Testing Service, Princeton, NJ.

Yamamoto, K., & Everson, H. (1997) Modeling the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model. in Rost, J. & Langeheine R. (Eds.) Applications of Latent Trait and Latent Class Models in the Social Sciences, New York, U.S.A.; Waxmann Munster.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155-174.

Introduction

The number and frequency of International Large Scale Assessments of various populations in many cognitive skill domains have been increasing, and so has the intensity of the attention countries and the media place upon the findings. The major global and consistent efforts for in-school populations have been the Trends in International Mathematics and Science Study (TIMSS) since the 1960s and the Progress in International Reading Literacy Study (PIRLS) since the 1990s, both sponsored by the International Association for the Evaluation of Educational Achievement (IEA); as well as the Programme in International Student Assessment (PISA) since the 2000s and for adult populations the International Adult Literacy Survey since the 1990s, both sponsored by the Organization for Economic Co-operation and Development (OECD). Most recently the Programme for the International Assessment of Adult Competencies (PIAAC) has been developed for adult populations and data collection is under way.

The PIAAC is the most comprehensive international survey of adult skills ever undertaken. In response to the growing awareness of literacy as human capital, the PIAAC is designed to provide policy makers and other stakeholders with information about the critical skills individuals need to maintain and enhance their ability to meet changing work conditions and societal demands. To better reflect the changing nature of information, the PIAAC has been developed as a computer-based assessment, allowing the inclusion of assessment tasks that represent the real-world demands faced by adults today. The survey focuses on the key areas of numeracy and literacy, including measures of electronic reading and of reading component skills for those demonstrating more limited literacy proficiency. For the first time in an adult survey, the PIAAC also assesses problem-solving skills in technology rich environments. In addition, the survey includes an extensive questionnaire that collects a broad range of information ranging from demographic data to information about how skills are used in a variety of adult contexts such as at home and work, and also in the community.

All surveys have critical points in three areas to make sure validity and reliability as well as comparability of inferences can be assured, namely 1) survey instrumentation and design, 2) sampling and data collection operations, and 3) psychometric analyses. Due to limited space and time, it was decided to present a general overview and look at how they relate to each other instead of a detailed treatment of any of the areas. The guiding principle of any feature of surveys and decisions in regard to the psychometric analyses is to establish comparability of inferences, assessment constructs, instrumentation, survey design, and psychometric analyses.

Development of the Measurement Construct

The literacy constructs adapted in the PIAAC followed a series of assessments originating from the 1984 Young Adult Literacy Survey (YALS), National Adult Literacy Survey (NALS), International Adult Literacy Survey (IALS), and Adult Lifeskills and Literacy Survey (ALLS). Constructs were organized as prose, document and quantitative in the earlier assessments. For the recent ALLS, the new constructs of numeracy and problem

solving were added and quantitative literacy was dropped. A new problem solving construct for technology rich environments was developed for the PIAAC and so prose and document are now combined with literacy. A key characteristic of these surveys is that each was based on a framework following Messick's (1994) construct-centered approach, 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 that elicit those behaviors.

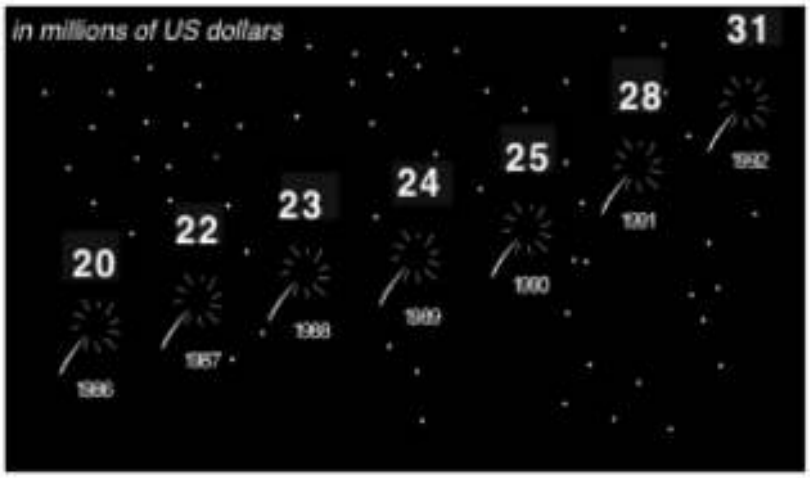
Having explicit literacy constructs allowed researchers to explore variables that explained differences in performance. This led to a departure from the prevailing difficulty models in the field focusing on the complexity of stimulus materials alone to a new understanding focusing on the relationship between the print materials that adults use in their everyday lives and the kinds of tasks they need to accomplish using such materials. Stimulus materials and questions were analyzed in terms of linguistic features and structures, as well as a range of processing variables related to task demands. The type of information based on the linguistic features and structures represents the complexity of information present in the stimuli. The typical stimuli that consist of series of sentences and paragraphs were identified as prose materials, and the printed information organized by a matrix format in various forms, including lists, were identified as document stimuli. Mosenthall and Kirsch (1991) further categorized document stimuli into six structures: simple, combined, intersecting and nested lists, and charts and graphs. They identified that the presence or absence of graphic organizers, including headings, bullets, and bold or italicized print, in the prose stimuli had an impact on the difficulty of items. Task demands can vary a great deal depending upon questions based on the same stimulus, that is to say both a simple question, such as locating synonymous information in the stimulus, and a very difficult question, such as making general inferences based on multiple pieces of information in the same stimulus. These task demands of – locate, cycle, integrate and generate – in combination with text features, helped explain what made some tasks more or less difficult than others (Kirsch, 2001).

The overall characteristics of stimuli and items in the NALS, IALS, ALL and now the PIAAC do not look like typical items used for any reading assessment. Items were situated in a more realistic context, all the information needed to answer is included in the stimulus, and items are centered on the notion of the stimulus materials. None of the items are multiple-choice items, and they all require open-ended responses. Instruction for the scoring of these answers was somewhat different by also being lenient in regard to the precision of responses but paying attention to their accuracy. This approach can result in ignoring minor deviations in spelling, the selection of significant digits by approximation can be allowed, and the beginning and the end of an underline can include the line above or below as long as no contradictory information is included.

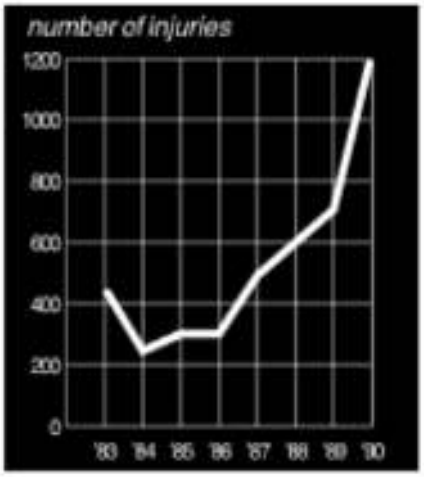
These features are brought about due to how adults tend to approach reading in their daily lives. Adults tend to approach reading texts and printed information with specific intention and purposes such as enjoyment, work, learning ideas presented

in the written text or tables, obtaining necessary transportation information, and so on, instead of the activity of reading itself. The figures in the following example were taken from an article that appeared in a newspaper. Respondents were asked to answer three questions based on their understanding of the figures and inferences one may be able to draw from them.

Fireworks in the Netherlands



Victims of fireworks



Questions 12 - 14. Use the charts on the opposite page to answer questions 12 through 14.

12. In what year were the fewest number of people in the Netherlands injured by fireworks?

_____

13. According to the charts, what was the value, in US dollars, of fireworks sold in the Netherlands in 1991?

_____

14. Describe the relationship between sales of fireworks and injuries due to fireworks.

_____

_____

_____

Establishing Comparability

Adaptation and Translation

If valid comparisons of assessment results are to be made across countries and from one cycle to the next, the equivalence of different language versions of the assessment instruments is essential. Equivalence refers to semantic equivalence (content), as well as equivalence in terms of register, style readability and other characteristics likely to affect psychometric properties. The guideline of adaptation can be unique to an item reflecting the measurement construct of the item. Adaptation often deals with units of numerical representation, unfamiliar to familiar names and expressions in order to improve cross-cultural functioning. These adaptations are understood as non-central to the item features that directly impact the type and complexity of questions, and thus impact upon the difficulty.

Scoring

Accurate and reliable scoring is a key component of establishing the comparability of inferences. Scoring is required to determine whether respondents have correctly answered the questions. Re-scoring is required as a quality assurance measure to determine whether the scoring rubrics have been applied consistently by every scorer within the country and without bias across the country.

Within-country reliability will require a second scorer to re-score a pre-defined number of cognitive instruments. The purposes for rescoring are to: 1) document the degree to which the same scores are given for the same responses, regardless of the scorer; and 2) identify items and scorers that have low inter-rater agreement (i.e., low consistency). Items with low inter-rater reliability will be further examined for possible ways to improve scoring accuracy through improved translation, instruction and/or training.

Across-country scoring requires scorers from different countries to score an identical set of responses. This is not possible to accomplish since the instruments are in multiple languages and adaptation has introduced differences even among English-language countries. For this reason, the sample responses in English were assembled to form anchor booklets. At least two bilingual scorers (fluent in the country's national language and English) will score the English language international anchor booklets to ensure the equivalence of scoring across countries. The scores of these two bilingual scorers will be compared and evaluated against the master scores for accuracy. Inaccurate scores should be investigated as any systematic deviations may require that country scores be corrected.

Unique Features of the Population Survey Design

In large-scale assessments, designers are challenged to find ways to minimize the amount of time any one student spends responding to a set of cognitive items while also maximizing the measurement construct described by the framework. This always requires a large number of tasks be created such that the number of items would be too large to be taken by any one person. Most large-scale surveys such as the PISA, TIMSS, IALS and ALL, accommodate the need for a minimum testing time over a large pool of items by using some form of incomplete block design.

In any incomplete design, each student is directed to respond to a subset of items from the selected subject domains following a particular pattern, while a methodological approach can be applied to describe how various populations and subgroups perform on the full set of items. Through such a design, a sufficient number of items representing the full range of a particular construct can be used to describe the skill distribution of the total population while somewhat minimizing the burden on each respondent. The typical incomplete design prescribes that all blocks in a domain replicate the measurement characteristics for the domain it contains.

The table below shows an example of the combination of two blocks of cognitive items used for the ALLS. The literacy items of the prose and document scales were pooled together to create four literacy blocks (L1-L4) to be well-matched and parallel to each other. Numeracy and problem solving items were created new for the ALL and divided into two blocks each, (N1 and N2) and (PS1 and PS2). The row labels represent the first block and the column labels represent the second block.

Eight blocks were paired according to the design below to produce 28 booklets, ensuring every block is balanced in terms of position in the booklet as well as a pairwise combination with literacy. For example, booklet 3 has two blocks of literacy, L2 at first then L3. The block L2 appears 4 times each at two positions, and is paired with both numeracy and problem solving blocks. This design together with the Item Response Theory (IRT) model makes it possible to place items in multiple blocks and skill proficiencies of respondents onto a single unified scale. Note that the covariance information between numeracy and problem solving scales cannot be assessed directly due to the omission of booklets including both numeracy and problem solving blocks in order to reduce the sample size. Many designs of international, large-scale surveys have similar characteristics.

Use of Background Variables

Background variables can be viewed as being one of three types according to the functions they fulfill. The first type of variables is often demographic variables used to describe the distribution of skills among the population such as gender, age,

Booklet Design of the ALL

		2nd Block							
		L1	L2	L3	L4	N1	N2	PS1	PS2
1st Block	L1		1	2		9			21
	L2			3	4		11	19	
	L3	5			6	10			22
	L4	7	8				12	20	
	N1		13		14		17		
	N2	15		16		18			
	PS1	23		24					27
	PS2		25		26			28	

locality of residence, and so on. The second type of variables is often variables used to describe the learning experiences and conditions impacting on the attainment of skills, such as education, training, education of the parents, and so on. The last type of variables is thought to be describing the direct or indirect social outcomes due to differences in skills, such as civic participation, voting, income, work history, health, and so on. All these variables can be used to aggregate respondents for reporting the proficiency distribution.

Population Modeling

The booklet design presented above solicits relatively few cognitive responses from each sampled respondent per subscale but maintains a wide range of construct representation when responses are summed up for all respondents. The reduced number of cognitive responses per respondent leads to proportionately less information, thus greater uncertainty in regard to the respondent's skill and thus it is not suitable for estimating the skills of an individual respondent. The size of the uncertainty, i.e., measurement error, based solely on the responses of an individual respondent is too large to be ignored and conventional statistical analyses would produce bias. Population modeling uses two steps, first to reduce measurement error through the empirical Bayes method and then transfer the variability of a posterior likelihood function through imputed values. Bayesian population modeling together with multiple imputations enables us to estimate population characteristics more efficiently and capture the remaining uncertainty of proficiency estimates. In order to correctly capture relationships between background variables and proficiency skills unique to each country, separate population modeling is required.

In addition to cognitive skills, background questions were also computerized to be able to ask more appropriate questions for the unique circumstances of the respondent, expediency and accuracy of data collection, and construction of a database. With paper-based background questionnaires, this is very difficult and results in complicated procedures to adapt a series of questions for each respondent, such as questions depending on the previous questions, and other word path dependent questions. Complicated data collection procedures administered by interviewer often invite non-random errors and it takes longer to collect and record background information.
Computerization enables us to collect additional information during the background data collection, such as the time of the data collection, duration of each question, correction path of erroneously entered answers, interviewer identity, and possibly the location of the data collection. Some of the data cleaning functions can be carried out during the data collection, such as the verification of the range of responses, relevance of the information, and categorization of values. These data can be recorded to a database following the prescribed record layout, ready to be summarized and analyzed.

The first cycle of the PIAAC is the collaboration among the 25 participating countries and an international consortium of organizations. Well over 125,000 adults aged 16-65 will take part in this assessment that is conducted in participants' homes by trained interviewers. The PIAAC is administered on laptop computers, making it the first

computer-based assessment to be used in a large-scale household survey.

IRT Linkage to the Previous Surveys

The PIAAC is linked to the International Adult Literacy Survey (IALS) and Adult Literacy and Life Skills Survey (ALL) for literacy and numeracy scales. The unique survey design and expanded role of the field study has made it possible to maintain this linkage that has made it possible to look at changes in skill levels, as well as the distribution of those skills, over time. For the field study, altogether there were 82 literacy items and 77 numeracy items combining both paper-based and computer-based items. About half of the items were newly developed for the PIAAC and the other half were common to previous surveys. About two thirds of the common items were common across two modes of presentations. The detailed design of the linking items is presented below and the design enables us to examine the comparability of the IRT item parameters between paper-based and computer-based administrations on the common items. The design also allows us to compare the equivalence of the PIAAC IRT parameters on paper administration against previous IALS/ALL survey results.

Establishing the equivalence of scales and inferences between the paper-based IASL/ALL and the computer-based PIAAC takes two steps. The first is to establish the equivalence between the IASL/ALL and PIAAC scales on paper-based items. Sampled respondents were randomly assigned to one of the paper-based booklets or a set of computer-based items to ensure aggregate skill distributions are equal across all groups with different sets of items. Every item by country deviations from the common item parameters were examined using the IRT-based method. The field study data supported the IRT parameter invariance between the IALS/ALL to the PIAAC for the paper-based items for the aggregate of all countries. As expected, a few items showed item characteristic deviations for some countries. The second step is to establish the equivalence at the item parameter levels between paper-based items and computer-based items. The majority of the 25 common linking items showed quite a good fit to the common item parameters derived from the IASL/ALL data. The paper-based items adapted for the computer-based assessment showed a fair amount of deviation for the literacy scale compared to items for the numeracy

PIAAC Field Test Items

	Literacy		Numeracy		PS
	Paper	CBA	Paper	CBA	CBA
Link both modes	25	25	25	25	
Link Paper only	0	0	2	0	
Link CBA only	0	17	0	13	
New both modes	0	0	5	5	
New Paper	10	0	3	0	
New CBA		30	0	29	23
Total	35	72	35	72	23
Grand Total	82		77		23

scale. However, increased deviations across countries were observed. The main study instruments were developed using the items with greater commonalities among countries and stronger linkage to the IALS/ALL scales. The two sequential links of the scales described above are necessary and the linkage between the PIAAC and IALS/ALL can't be established without both being successful.

In addition to maintaining important connections to past assessments, the PIAAC is also an innovative survey in a number of important ways. Two types of innovation will be discussed. In some areas, the innovations in the PIAAC are incremental. That is, they represent a shift in thinking or practice – reflecting the origins of the word "innovation" which means "to renew or change." In this sense, the PIAAC builds upon the foundations laid by earlier adult literacy assessments and other large-scale assessments. Examples of this type of incremental innovation include extending knowledge from previous assessments about scoring paper and pencil items to the computer-based environment and building upon previous psychometric models that allow us to interpret different classes of missing data due to non-response. In other instances, the innovations in PIAAC truly lay new groundwork. To develop national assessment materials, new tools were needed to be built that would allow countries to translate the computer-based tasks and perform layout and online scoring checks. In order to deliver this complex adaptive assessment, a platform had to be developed that was manageable for countries and that would work across a variety of case management systems. And because the computer-based tasks yield a broader range and amount of data, models and procedures had to be developed to facilitate the collection and interpretation of that data.

PIAAC Main Survey Design

More recent developments in the measurement of individual skills suggest adaptive testing as a means toward achieving test optimization in terms of the total testing time or amount of information per unit testing time over non-adaptive tests. While the advantages of adaptive testing are clear over traditional tests, this poses difficulties due to the inflated dependency of early administration, sensitivity to inaccurate item parameters, and over exposure of some subsets of items while there is under-exposure of others. The PIAAC implemented multistage adaptive testing design, i.e., a cluster of items is administered depending upon the skills of the respondent instead of item level adaptive testing. By grouping several items together, it reduces the over reaction to the responses to the items administered in the beginning of the test and the uncertainty of the item parameters of a particular item becomes less critical due to the fact the adaptive decision is based on the responses to a set of items.

Implementation of the Multi Stage Adaptive Design in the PIAAC

The following PIAAC assessment design is implemented and there are two versions of the PIAAC assessments: computer-based and paper-and-pencil. ALL participants are first given survey questions on their background information, including, but not limited to the following: education, whether they are native to the language used in the survey, and any computer experience. Respondents answering no to computer experience will be given the paper-and-pencil assessment. Only those

reporting some computer experience will be given a basic computer skill test of "CBA-Core Stage 1", like how to use a mouse, how to open a file, how to enter text, and so forth. Then each person will take a six-question "CBA-Core Stage 2" test. Only respondents having passed the Core test will receive the computer version of the assessment.

The multi-stage adaptation takes place twice within each scale of literacy and numeracy. The problem solving scale does not use any adaptation design.

In the CBA assessment, a respondent has an equal probability to get a test on literacy, numeracy, or problem solving in a technology-rich environment (i.e., how to use a computer in daily life). All sessions have several versions of the test booklets. The test booklets in literacy or numeracy are ordered by difficulty and are chosen to best match the respondents' abilities. Literacy and numeracy both have two stages of test booklets. Stage 1 has three test booklets; each with nine questions, and Stage 2 has four test booklets, each with 11 questions.

In the Stage 1 adaptation, a test booklet is selected based on the respondent's background information, such as their highest level of education, and language skills and responses on the Core test. In the Stage 2 adaptation, a test booklet is chosen based on the responses to the items in the Core Stage 1, in addition to all the information used in the Stage 1 adaptation.

The data collection is under way in 25 countries. The final data fit for analyses is expected to be ready by late 2012 and numerous reports, public database and supportive documents will be published in late 2013. Readers can monitor the developments at http://www.oecd.org/piaac.

The PIAAC has broken new ground in terms of the domains being assessed, the technology used to build and deliver the large scale population assessment, design of the survey itself and the analyses methodology in terms of the data being captured and how that data is analyzed. Many features found in the current PIAAC foreshadow the future of the ILSA and new features are being contemplated.

Figure 1. PIAAC routing diagram