Psychological Test and Assessment Modeling, Volume 60, 2018 (3), 347-368

Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC

Kentaro Yamamoto[1], Lale Khorramdel[2] & Hyo Jeong Shin[2]

Abstract

PIAAC is one of the first international large-scale assessments that implemented a multistage adaptive testing (MST) design. The design consists of multiple layers of adaptation to administer the most relevant and efficient set of questions based on the estimated proficiency of respondents. The benefits of the MST design were evaluated in terms of the comparability of item parameters across countries and the test efficiency. To assess the comparability across countries, item-by-country interactions were examined using item response theory (IRT) models. The efficiency of the MST design was calculated and compared to a nonadaptive design with a fixed item format. Moreover, possible effects of the position of item sets on item difficulty, which would present a problem for implementing MST, were examined. Results show a higher test efficiency in the MST design, only small item position effects and a high comparability of item parameters across different countries and languages.

Keywords: multistage adaptive testing, IRT, item position, test efficiency, measurement invariance

[1]*Correspondence concerning this article should be addressed to:* Kentaro Yamamoto, Educational Testing Service, Research and Development, Center for Global Assessment, 660 Rosedale Road, Princeton, NJ 08541, USA, e-mail: kyamamoto@ets.org

[2]Educational Testing Service

International large-scale assessments compare the skills, knowledge, and behaviors of various populations across countries and economies with a focus on group scores as opposed to large-scale testing programs that focus on individual test scores (Kirsch, Lennon, von Davier, Gonzalez, & Yamamoto, 2013). They aim at populations and subpopulations from a diverse group around the world (von Davier, Sinharay, Oranje, & Beaton, 2006) and need to account for heterogeneous performance within and across participating countries. An increasing number of participating countries necessitates measuring a broader range of proficiency levels not just within but also across countries. As a result, new methodologies are developed and applied to increase fairness, measurement reliability, and test efficiency. This has also led to large-scale assessments moving from paper-based assessment (PBA) to computer-based assessment (CBA), allowing the measurement of new constructs and the collection of additional information (such as timing and other process data) that can be used to improve proficiency estimation and reduce measurement error. Moreover, CBAs allow the implementation of adaptive test designs that aim to increase the efficiency, validity, and accuracy of the measured construct of interest by matching the administration of test items to the proficiency level of test takers.

Adaptive tests have been shown to obtain more consistently efficient and precise measurements of examinees across the entire proficiency distribution compared to traditional linear tests (Lord, 1980; Wainer, 1990), especially with regard to the ends of the proficiency scale (Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1974). Adaptive test designs may also reduce the linking error in large-scale assessments (Wu, 2010) and potentially increase engagement and test taking motivation (Arvey, Strickland, Drauden, & Martin, 1990; Asseburg & Frey, 2013) especially for low performing respondents (Betz & Weiss, 1976), hence reducing nonresponse and random responding, which both are an issue in large-scale assessments. But there are also studies that assume or show no increase in motivation under adaptive testing (Bergstrom, Lunz, & Gershon, 1992; Eggen, 2004; Ling, Attali, Finn, & Stone, 2017; Ponsoda, Olea, Rodriguez, & Revuelta, 1999; Wise, 2014), especially with regard to higher-performing test takers (Frey, Hartig, & Moosbrugger, 2009). However, most of these studies focused on item-level computerized adaptive testing (CAT) and individual test scores and might not directly apply to multistage adaptive testing (MST) or group score and large-scale assessments. A positive impact of MST on test-taking motivation has yet to be sufficiently examined[3].

The Programme for the International Assessment of Adult Competencies (PIAAC) was one of the first international large-scale assessments introducing a CBA and an adaptive test design in the form of MST. PIAAC is a cyclical internationally standardized survey that measures adults' proficiency in the key information-processing skills of Literacy, Numeracy, and Problem Solving in Technology-rich Environments (Organisation for Economic Co-operation and Development, 2016). The PIAAC target population is a

[3] We thank an anonymous reviewer for pointing out the importance of presenting the conflicting results of studies on the relation between motivation and adaptive testing. But we would like to stress that test taking motivation was not the main goal of introducing an adaptive design in PIAAC and that the current study was not designed to examine this issue.

household sample with respondents between the ages of 16 and 65. Twenty-four countries participated in PIAAC Round 1 and nine additional countries in Round 2 (more additional countries will be tested in 2018 in PIAAC Round 3). This paper illustrates and evaluates the PIAAC MST design and shows that such a design can be implemented in large-scale assessments despite their design constraints, leading to increased test efficiency and better meeting the needs of countries with various performance levels. The PIAAC adaptive test design can be used as an example to establish similar designs for other international large-scale assessments that focus on group-level scores.

The current paper aims to illustrate the challenges and possible solutions for introducing adaptive testing into international large-scale assessments based on the example of PIAAC. In the following sections, we will present advantages of MST compared to CAT in the context of large-scale assessments, particularly for PIAAC. Then, we will describe the MST design implemented in PIAAC in more detail, and present results of the evaluation of the design. In the discussion of the findings, we will compare the advantages of the PIAAC test design to its limitations and discuss the generalizability to other large-scale assessments.

Expected advantages of MST in PIAAC

Adaptive tests can be roughly distinguished as belonging to one of two groups: item-level adaptive tests and multistage adaptive tests (Lord, 1971; Zenisky, Hambleton, & Luecht, 2010). Item-level adaptive tests, or CATs, have been in use for some time. Their use has been described in psychological assessment (Kubinger, 2016) as well as educational assessment (Weiss & Kingsbury, 1984). Item-level adaptive testing is particularly attractive for the testing programs that focus on the assessment of individuals, where a relatively narrow construct is assessed based on a large item pool such as GRE® (Robin, Steffen, & Liang, 2014) or CPA examinations (Breithaupt, Zhang, & Hare, 2014). However, implementing CAT in this form in a large-scale assessment is rarely feasible due to certain goals and design constraints (Reese, Schnipke, & Luebke, 1999). Issues include the need of sufficient construct coverage across all proficiency levels, especially when assessing constructs that are based on a broader construct definition, comparability of the data and scale, and the need to balance the distribution of item contents and item types, and item positions to avoid biased item parameter and proficiency estimates, for example due to possible position effects.

An MST design appears better suited to deal with such issues as it allows for more control of the item exposure. MST is a natural generalization of CAT and is described as a balanced compromise between linear test forms and item-level CAT, combining the advantages of both (being adaptive but allowing experts to review test forms and allowing respondents to change responses). It is an extension that allows the choice of the next item set (comprising several items) as opposed to choosing single items. This approach allows to control the presentation of items across different test forms for a better construct coverage and the possibility to balance the item position to prevent bias on parameter estimation. Moreover, it accumulates more information after each adaptive step, which can lead to

greater accuracy in the decision of the next adaptive path (compared to approaches that use single items for each adaptive decision or path). This reduces the likely dependence of the adaptive selection on item-by-country interactions (found in international large-scale assessments) as compared to those expected with item-level adaptive tests (Kirsch & Thorn, 2013).

International large-scale assessments make use of item sets (units) that are developed around a stimulus (i.e. several items share the same stimulus) and allow more freedom in test assembly. These item sets have to stay intact (i.e., all items of an item set have to be presented to the respondent) to fully represent the framework of the measured construct. Furthermore, they cannot be split to fit an item-level CAT without increasing testing time and reading effort. In current international large-scale assessments, the questions in these item sets are often associated with a realistic and more complex context to better resemble the measured construct. Examples include scenarios in which respondents are provided an overarching purpose for reading a collection of thematically related texts in order to respond to some larger integrative question or to write a recommendation based on a set of texts (Organisation for Economic Co-operation and Development, 2016). Furthermore, the possibility of using intact item sets in MST makes it easier to incorporate open-ended response items that are not automatically scoreable by the computer. The adaptive decision for the next set of items can be solely based on the automatically scored responses. Item-level CAT is usually based on automatically scoreable items (e.g., multiple-choice or open-ended response items). Items that cannot be automatically scored cannot be incorporated into item-level adaptive algorithms without increasing the testing time. However, PIAAC only contains short constructed-response items that can be automatically scored.

One might argue that a testlet-based CAT (Wainer, Bradlow, & Wang, 2007), which is also based on the use of items sets, could be used instead of MST. Testlet-based CAT designs have been studied more extensively and have been developed for individual-score reporting. However, their applications were not studied for group-score reporting and this approach is not used to estimate item parameters but assumes that item parameters are already known. In PIAAC and other international large-scale assessments, the test design is usually based on preliminary item parameters from a field test while the final item parameters are estimated in the main survey, which is based on the final design. Furthermore, to enable the estimation and recovery of item parameters in the main survey, PIAAC uses a linking approach within each adaptive stage; in other words, within one stage, item sets share a certain number of items and are linked to each other. Because of the importance of estimating item parameters in international large-scale assessments that are comparable across a large number of countries and multiple assessment cycles over time, the application of testlet-based CATs may not be feasible at this point. Another approach similar to MST is presented by CAT shadow tests (van der Linden & Veldkamp, 2004) which were shown to be similar to multistage tests when applying certain constraints (Choi et al., 2016). But shadow tests need a very large item pool and are computing-intensive. Hence, they are not an option for international large-scale assessments at this point.

For all these reasons, adaptive tests on the level of item sets such as MST designs appear to be more suitable and easier to implement in the context of group score assessments than multiple isolated questions (Oranje, Mazzeo, Xu, & Kulick, 2014). Therefore, it was

decided to adapt an MST design for PIAAC. This approach best allows for matching item difficulty with the abilities of respondents while meeting other design requirements (item parameter estimation, broad construct coverage, balancing item content, item type and the position of items, linking) at the same time.

Research questions and aims of the current paper

This paper aims to illustrate the PIAAC test design as an example how to introduce adaptive testing into an international large-scale assessment while, at the same time, accounting for constraints typical for such assessments. In other words, we aim to illustrate one example of how to combine adaptive features with general requirements and restrictions of large-scale assessments. The described design is specific to PIAAC but the rational and thoughts behind can be applied for other international large-scale assessments as well.

The main goals of the MST design in PIAAC are a) to optimize the delivery of test items to provide more reliable information about skills without increasing testing time, b) to enable a broad construct coverage, c) account for heterogeneous performances across respondents within and across the participating countries, and d) reduce the impact of possible item-by-country interactions to achieve comparable item parameters and test scores. To examine whether the design is able to meet these goals, we examined the efficiency of the MST design compared to a nonadaptive design and the number of item-by-country interactions for the domains Literacy and Numeracy. Moreover, we examined the presence of item position effects as requirement for the implementation of an adaptive design. Comparable item parameters and test scores in MST can only be achieved if the impact of item position is minimal.

Method

This section describes the final MST design for PIAAC and the methods used for evaluating it.

The multistage adaptive testing design in PIAAC

General aspects of the Design. The test design for PIAAC was based on a variant of matrix sampling (using different sets of items, MST, and different assessment modes) where each respondent was administered a subset of items from the total item pool. That is, different groups of respondents answered different sets of items. The assessment consists of a background questionnaire (BQ) administered in the beginning (30–40 minutes) followed by a cognitive assessment (60 minutes) measuring the four domains Literacy, Numeracy, Reading Components (RC), and PSTRE. Moreover, a link to prior adult surveys (IALS and ALL) was established through 60 percent of common Literacy and Numeracy linking items. PIAAC consists of two consecutive assessments in each cycle and for each participating country, a field test and a main study.

The MST design for the PIAAC main survey was prepared in multiple steps based on the analysis of the field test data. In a first step, the field test was used to examine the role of computer familiarity and to evaluate the equivalence of item parameters between the PBA and CBA. For this, respondents in the participating countries were randomly assigned to either the PBA or the CBA. In a second step, the field test was used to establish initial item parameters based on item response theory (IRT) models. These parameters were used to construct the adaptive testing algorithm for branching respondents in the main study MST design. More details about the PIAAC field test design and analysis in preparation of the final PIAAC MST design can be found in the PIAAC Technical Report (Organisation for Economic Co-operation and Development, 2013; Kirsch & Yamamoto, 2013). To enable adaptive testing in PIAAC, different item types (highlighting, clicking, single choice, multiple choice, and numeric entry) were scored automatically and instantaneously by the computer-based platform based on international and national scoring rules; see the PIAAC Technical Report for more information (Organisation for Economic Co-operation and Development, 2013). For describing the final PIAAC MST design, the terminologies described in Table 1 are used.

Table 1:
Terminologies used to describe the PIAAC Design

PBA (Nonadaptive)	CBA (Adaptive)
Item: refers to a task to which an examinee is directed to provide a response. The response is coded based on a coding guide; in PIAAC all items are machine coded	
Unit: refers to a short and mutually exclusive set of items in the PIAAC adaptive test design	
Cluster: refers to a mutually exclusive set of items in the PBA; one cluster takes 30 minutes testing time on average	Block: a set of units in the PIAAC adaptive test design; each respondent receives two blocks: one in adaptive stage 1 and one in adaptive stage 2
Booklet: each respondent in the nonadaptive PBA receives one booklet; a booklet consists of two 30-minute clusters (60 minutes on average)	Module: refers to a domain-specific set of two blocks across the adaptive stages in the PIAAC adaptive test design (one stage 1 block and one stage 2 block); one module takes 30 minutes testing time on average; each examinee receives two cognitive domains, that is, two modules (60 minutes on average)

The final MST design. The final PIAAC MST design as described in Figure 1 is not constrained to the cognitive assessment but also uses information from the background questionnaire (BQ). The first step in the adaptive design is to route respondents to either the PBA or the CBA based on their responses to questions from the BQ and a core set of questions focusing on information and communications technology (ICT) skills. Respondents who reported no familiarity with computers were routed to the PBA, as were respondents refusing to take the test on the computer. Respondents who reported familiarity with computers in the main study were routed to the CBA. The second level of adaptation was

within the CBA. PIAAC used a probability-based multistage adaptive algorithm where the cognitive items for Literacy and Numeracy were administered to respondents in an adaptive way (PSTRE was not administered adaptively). In other words, more able respondents received a more difficult set of items than less able respondents.

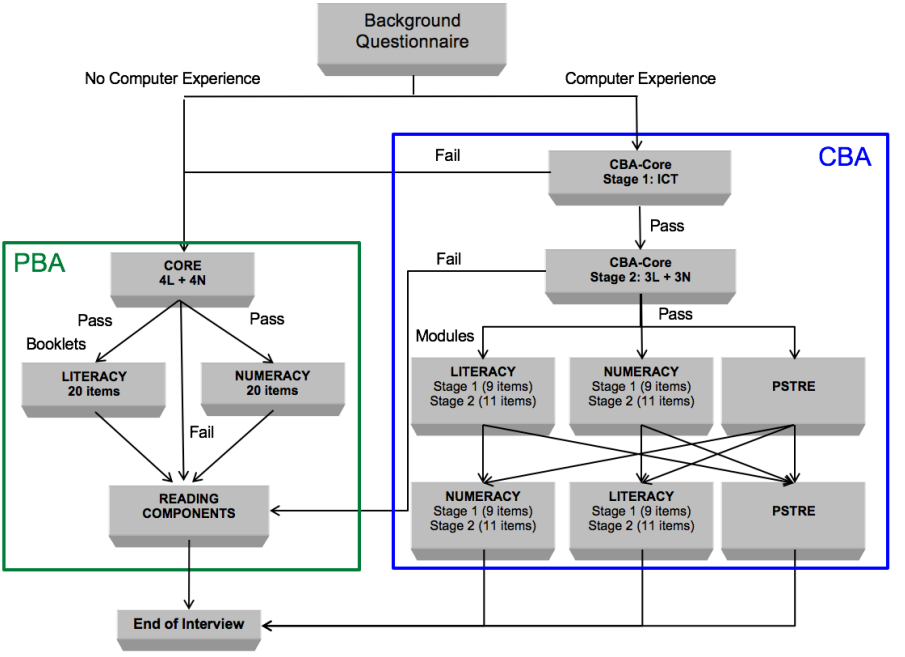

Figure 1:
PIAAC MST design for the Round 1 and Round 2 main study.

The paper-delivered branch included a 10-minute core assessment of Literacy and Numeracy skills. Respondents who performed at or above a minimum standard were randomly assigned to a 30-minute cluster of Literacy or Numeracy items, followed by a 20-minute assessment of Reading Component skills. The relatively small proportion of respondents who performed poorly on the paper-and-pencil core items skipped the Literacy and Numeracy items and were routed directly to the Reading Components.

The computer-delivered branch of the assessment first directed respondents to the CBA core section, which was composed of two stages taking approximately five minutes each. Poor performance on either stage of the CBA Core section resulted in switching over to the appropriate sections of the paper-and-pencil instruments. Respondents who failed CBA Core Stage 1 (which contained ICT-related items) were directed to begin the paper-based core section and proceed with the process outlined in the above bullet. Respondents who passed CBA Core Stage 1 but failed CBA Core Stage 2 (which contained six cognitive items) were then administered only the Reading Component items. Respondents who

performed well on both CBA core sections were routed to one of three possible outcomes (each taking approximately 50 minutes): respondents received a combination of Literacy and Numeracy modules, or a PSTRE module combined with either a Literacy or a Numeracy module, or only PSTRE modules. The Literacy and Numeracy modules each consisted of two adaptive stages. Each stage contained a number of blocks varying in difficulty, and each block consisted of several item units (a unit is a mutually exclusive set of items). In each stage, only one block was delivered to a respondent. The blocks within one stage were linked through a common item unit (see Table 2). This was necessary to provide stable item parameter estimates in the main study. Within each of these modules, a respondent took 20 items (nine items in Stage 1; 11 in Stage 2). Thus, respondents taking Literacy in Module 1 and Numeracy in Module 2 (or vice versa) answered 40 items. Each module was designed to take an average of 30 minutes. PSTRE is unique because of the nature of the domain. It was organized as two fixed sets of items: seven in Module 1 and seven in Module 2. These were also designed to take an average of 30 minutes. In contrast to Literacy and Numeracy, the assessment of PSTRE was not adaptive. Table 2 provides an overview of the design of the MST stages 1 and 2.

Table 2:
Design of the Main Study CBA Instruments for Literacy and Numeracy in the Integrated Design

STAGE 1 (18 unique items – 9 items per block. Each respondent takes 1 block)							
	Unit A1 4 items	*Unit B1* 5 items	*Unit C1* 4 items	*Unit D1* 5 items			
Block 1-1	X	X					
Block 1-2		X	X				
Block 1-3			X	X			
STAGE 2 (31 unique items – 11 items per block. Each respondent takes 1 block)							
	Unit A2 6 items	*Unit B2* 5 items	*Block C2* 3 items	*Unit D2* 3 items	*Unit E2* 3 items	*Unit F2* 5 items	*Unit G2* 6 items
Block 2-1	X	X					
Block 2-2		X	X	X			
Block 2-3				X	X	X	
Block 2-4						X	X

Note. One block consists of two or three item units, one module within a stage consists of two blocks.

Due to the diversity of the participants' country, language, and educational backgrounds, a deterministic assignment of stages would likely have resulted in certain subpopulations being exposed to only a small percentage of items. To help mitigate the potential impact of such a situation, item exposure rates for specified subpopulations were controlled through a set of conditional probability tables (Chen, Yamamoto, & von Davier, 2014). This was important for achieving comparable data and test scores.

Module selection

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

Choice of the first block for Literacy and Numeracy in Module 1 (Stage 1). The Literacy and Numeracy blocks in Stage 1 varied in difficulty. There were three levels of blocks: easy (Block 1), medium (Block 2), and difficult (Block 3). Three variables determined which block was chosen for a respondent:

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

These three variables were organized in a matrix that resulted in two thresholds that presented probabilities of being assigned to a certain block. The following matrix (see Table 3) provides an example, using the Stage 1 selection.

Table 3:
Example of the Probability Matrix Design for the Stage 1 Selection of Literacy and Numeracy Blocks

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

As shown in the Table 3 matrix, if a respondent had a high education level, was a native speaker, and scored high on CBA-Core Stage 2 (for a total score of 6), he or she would be assigned 0.280 and 0.580 as thresholds. Then a random number between 0 and 1 was generated. This respondent received the easier block if the random number was less than 0.280; the medium test if equal to or greater than 0.280 and less than 0.580; and the difficult test if equal to or greater than 0.580. This process ensured that respondents who were

native speakers, highly educated, and performed well on the core were most likely to receive the most difficult block at the first stage compared to other blocks. However, there was some probability they would receive one of the easier blocks.

Choice of the second block for Literacy and Numeracy in Module 1 (Stage 2). The four Literacy and Numeracy blocks in Stage 2 also varied in difficulty, with Block 1 being the easiest and Block 4 the most difficult. For this scenario, three thresholds were defined because there was one more category than in Stage 1. Thus, the test assignment for Stage 2 depended on the following three variables as shown in Table 4.

Table 4:
Example of the Probability Matrix Design for the Stage 2 Selection of Literacy and Numeracy Blocks

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

- Education level (EdLevel3) from the BQ: Low, medium, or high
- Native versus nonnative speaker: The respondent was considered a native speaker if his or her first language was one of the assessment languages
- CBA-Core Stage 2 score plus Stage 1 score: CBA-Core Stage 2 passing scores were between 3 and 6 while the results of Stage 1 were between 0 and 9

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

According to Table 4, if a respondent had a high education level, was a native speaker, and scored high on the CBA-Core Stage 2 (for example a total score of 6) and had the highest score in Stage 1 (9, for a total of 15 for both stages), he or she would be assigned thresholds of 0.140, 0.360 and 0.615. Then a random number between 0 and 1 was

generated. Thus, this respondent would have received Block 1 (easiest) if the random number was less than 0.140, Block 2 if equal to or greater than 0.140 and less than 0.360, Block 3 if equal to or greater than 0.360 and less than 0.615, or Block 4 (most difficult) if equal to or greater than 0.615.

Choice of the second module. After completing Module 1 (either the two blocks for Literacy or Numeracy or the Problem-Solving module), the respondent proceeded to Module 2. The selection between Module 1 and Module 2 was also based on random numbers (between 0 and 1).

- If the respondent completed Literacy as Module 1, he or she was assigned Numeracy as Module 2 (starting with numeracy orientation) if the random number was less than 0.75. Otherwise he or she continued with PSTRE as Module 2 (starting with PSTRE orientation).
- If the respondent completed Numeracy as Module 1, he or she was assigned Literacy as Module 2 (starting with literacy orientation) if the random number was less than 0.75. Otherwise he or she continued with PSTRE as Module 2 (starting with PSTRE orientation).

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

The PIAAC instruments and sample

Analyses were based on 76 Literacy and 76 Numeracy items that were scored dichotomously, and the 14 problem-solving items were scored dichotomously or polytomously. Table 5 provides an overview of the number of items per assessment mode (PBA and CBA).

Data from 165,599 respondents from 24 countries in the Round 1 main study, and data from 43,221 respondents from 8 additional countries (out of 9; the other was excluded because it used PBA only) in the Round 2 main study were available for statistical analyses. For details about the country-specific sample sizes, please see the PIAAC Technical Report on the OECD website (Organisation for Economic Co-operation and Development, 2013).

Table 5:
Number of Cognitive Items per Assessment Mode and Domain in PIAAC

Domain	Assessment Mode	Number of Items
Literacy	CBA	52
	PBA	24
Numeracy	CBA	52
	PBA	24
PSTRE	CBA	14
Reading Components	PBA	100

Note. 18 Literacy and 17 Numeracy items were linking items between the PBA and CBA assessment mode, meaning these items were identical. Thus, PIAAC contained a total of 131 unique items.

Examining item position effects in the CBA. To implement the MST design in PIAAC without introducing any bias in the parameter estimation, one aim was to minimize any possible effect of item position in the CBA. An item position effect is present when a different position of items impacts the proportion of correct item responses, that is, the item difficulty or some other characteristic of the item. As a precaution, the PIAAC design in the CBA was set up in a way to counterbalance the potential effects of item position. Each respondent received two cognitive modules in the CBA, where each module comprised either Literacy, Numeracy, or PSTRE items. Each module of Literacy and Numeracy items appeared in two different positions within the assessment (see Figure 1). While an IRT based method to examine and account for item position effects in CAT calibrations was proposed by Frey, Bernhardt and Born (2017), this approach is not directly applicable to PIAAC, because the PIAAC design is complex as the item parameter estimation is influenced by different variables (e.g. linking across different countries, languages, assessment modes and assessment cycles, country-by-language interactions, etc.) and the scaling model in PIAAC cannot be changed without harming the trend measure. Finding and implementing a different scaling model to account for item position effects is not an option at this point because PIAAC is a large-scale survey that requires comparable and consistent skill inferences strictly attached to proficiency values based on the conditional probabilities specified by invariant item parameters across cycles regardless of item position.

The item position effect in PIAAC for the scales Literacy and Numeracy was examined using the average weighted proportion of correct responses (it was not possible to examine position effects on the PSTRE domain as the different PSTRE modules comprised different items, in contrast to Literacy and Numeracy).

The weighted proportion correct for an item was calculated as follows:

$$P_i = \frac{\sum_k WP_k \sum_j W_j \left(x_{ji} = 1 | k \right)}{\sum_k WP_k \left(\sum_j W_j \left(x_{ji} = 1 | k \right) + \sum_j W_j \left(x_{ji} = 0 | k \right) + \sum_j W_j \left(x_{ji} = 2 | k \right) \right)} \quad (1)$$

where the proportion correct on item ***i*** was calculated by using standardized weights[4] of path ***k*** (WP_k), final PIAAC sampling weights for the respondent ***j*** (W_j), and score responses correct "1", incorrect "0", and omit "2".

Examining the Efficiency of the MST Design. For evaluating the test efficiency of the MST design compared to a nonadaptive design, the field test data from Round 1 were analyzed. The relative efficiency of the PIAAC MST for Literacy and Numeracy is shown over an average of linear tests of equal length based on the same identical item sets defined as the ratio of two square root of test information curves: the value proportional to measurement errors. Identical item parameters were used for both the hypothetical MST condition and the nonadaptive condition (linear tests). For the nonadaptive condition, the test information curve was calculated as the average of the item information of the same number of items as in the adaptive paths.

Examining Item-By-Country and Item-By-Language Interactions. In the PIAAC Round 1 main study, international or common item parameters were estimated in a multiple-group IRT model (Bock & Zimowski, 1997; Yamamoto & Mazzeo, 1992) with countries divided by languages as separate groups based on the assumption of measurement invariance of item parameters across all groups. A unique or country-specific item parameter was estimated in case an item showed misfit to the common parameter, meaning item-by-country/language interactions could be identified. The more common item parameters that could be retained, the higher the comparability of data and test scores across countries and language. The estimation of common and unique parameters followed the procedures outlined in Glas and Jehangir (2013), Glas and Verhelst (1995), Oliveri and von Davier (2011, 2014), Yamamoto (1997) and, Yamamoto and Mazzeo (1992). For more details on the estimation of common and unique parameters, refer also to the PIAAC technical report (Organisation for Economic Co-operation and Development, 2013). All analyses were performed separately for each cognitive domain.

All cognitive items were calibrated based on the two-parameter logistic model (2PLM; Birnbaum, 1968) for dichotomously scored items and the generalized partial credit model (GPCM; Muraki, 1992) for polytomously scored items. The 2PLM and GPCM assume that a single latent trait is sufficient to represent the data (unidimensional models). Typically, their use is motivated by the need to summarize overall performance parsimoniously within a single domain. All models[5] were estimated using the software *mdltm* (von Davier, 2005). The software provides marginal maximum likelihood estimates obtained using customary expectation-maximization methods, with optional acceleration. Not-reached items in the PBA were scored as missing and omitted responses (any missing response followed by a valid response) were scored as incorrect. In the CBA, where it was possible to assess response times per item, nonresponses due to rapid omission were differentiated from

[4]Unique path weights were calculated for every path for each country in order to make statistics of item proportions correct comparable across countries. If they are applied to the respondents who went through a particular path, the proportion of each path would be identical for all countries. The target proportion for each path was set as average proportion across all countries.

[5]For further information regarding the models discussed, see Fischer and Molenaar (1995) and van der Linden and Hambleton (1997, 2016), or von Davier and Sinharay (2014) for the use of these models in the context of international comparative assessments.

nonresponses after interaction with the stimuli (based on literature on response latencies; cf. Setzer & Allspach, 2007; Wise & DeMars, 2005; Wise & Kong, 2005). Thus, omitted responses were only treated as wrong if a respondent spent more than five seconds on an item. If a respondent spent less than five seconds, the nonresponse was considered not attempted and treated as a missing value.

The PIAAC Round 2 main study used the common item parameters that were estimated in Round 1 to examine item-by-country/language interactions through a fixed item parameter linking (i.e., item parameters were fixed to those obtained in Round 1, and the fit of those parameters in the Round 2 data was evaluated for each country and language).

To identify item-by-country/language interactions, fit statistics were calculated using the root mean square deviation (RMSD). The RMSD is a standardized index of the discrepancy between the observed item characteristic curve (ICC) and the model-based ICC, both in terms of slope and location (intercept) of the item response function. The RMSD is always between 0 and 1 and computed by:

$$RMSD = \sqrt{\int (P_o(\theta) - P_e(\theta))^2 f(\theta) d\theta} \qquad (2)$$

with P_o for the observed percent of correct responses, P_e for the expected percent of correct responses, and $d\theta$ for the distribution of θ. As in the PIAAC operational scaling (Yamamoto, Khorramdel, & von Davier, 2013), poorly fitting ICCs were revealed using a RMSD > 0.15 criterion (a value of 0 indicates no discrepancy; in other words, a perfect fit of the model). There is no general rule about using the RMSD criterion. In numerous studies and other international large-scale assessments, a criterion of RMSD > 0.2 is used. In PIAAC, a stricter criterion was defined with the goal to exclude any bias and increase measurement precision.

A low number of item-by-country/language interactions would indicate that the data provided by the MST design in PIAAC show a high comparability and measurement invariance across groups leading to comparable test scores and increasing the validity of the PIAAC assessments.

Results

This section presents the results for the module position effect analysis and the IRT scaling for examining item-by-country and item-by-language interactions in the PIAAC Round 1 and Round 2 main studies, and the MST design efficiency in the PIAAC Round 1 Main Study.

Item position effects in the CBA

Table 6 shows the average proportion correct for items in a given module. The average proportion is calculated from the weighted and standardized data for all participating countries. The average proportions correct across all countries are virtually identical between the first and second module position (within 1 percentage point) regardless of paired domains. Only a slight item position effect was found: 2.9 % for Literacy items and 1.2 % for Numeracy items. Hence, there seems to be no large effect of item position enabling comparable item parameters. This finding is supported by the results described in the section about item-by-country interactions below where only a few interactions could be found.

Table 6:
Average Proportion Correct by Content-Related Module Order

Country	Average of Literacy Items 1st Module		Average of Numeracy Items 1st Module		Average of Literacy Items 2nd Module		Average of Numeracy Items 2nd Module	
	LIT-NUM	LIT-PS2	NUM-LIT	NUM-PS2	NUM-LIT	PS1-LIT	LIT-NUM	PS1-NUM
Round 1								
Australia	61.5 %	61.0 %	64.5 %	65.1 %	58.9 %	58.8 %	63.4 %	63.1 %
Austria	56.7 %	58.8 %	67.8 %	67.2 %	53.0 %	55.9 %	67.2 %	67.1 %
Canada	58.7 %	58.4 %	63.8 %	62.5 %	54.6 %	55.6 %	61.9 %	62.4 %
Cyprus	49.4 %		60.7 %		45.8 %		60.8 %	
Czech Rep.	53.5 %	54.4 %	68.6 %	65.4 %	53.9 %	51.6 %	64.7 %	66.5 %
Denmark	58.7 %	57.2 %	68.9 %	68.2 %	55.0 %	55.2 %	67.0 %	68.1 %
England/N. Ireland (UK)	58.0 %	57.6 %	60.5 %	60.8 %	52.2 %	51.8 %	59.9 %	60.4 %
Estonia	57.0 %	57.1 %	65.7 %	65.1 %	54.2 %	54.7 %	65.4 %	66.9 %
Finland	65.5 %	65.2 %	72.5 %	74.0 %	63.3 %	62.6 %	70.2 %	67.9 %
Flanders (Belgium)	60.0 %	57.9 %	67.2 %	69.7 %	57.1 %	58.5 %	67.3 %	65.5 %
France	52.1 %		60.2 %		48.4 %		58.8 %	
Germany	57.1 %	56.6 %	66.3 %	67.5 %	53.0 %	51.9 %	65.9 %	65.3 %
Ireland	56.3 %	56.4 %	60.7 %	60.9 %	52.1 %	50.7 %	58.9 %	56.5 %
Italy	47.5 %		56.9 %		44.2 %		55.6 %	
Japan	67.0 %	68.9 %	75.7 %	76.1 %	64.3 %	64.1 %	73.9 %	74.1 %
Korea	57.2 %	57.1 %	62.9 %	63.4 %	56.9 %	57.8 %	62.9 %	60.6 %
Netherlands	62.8 %	62.3 %	68.5 %	69.3 %	59.6 %	61.1 %	69.0 %	66.8 %
Norway	60.3 %	61.0 %	69.2 %	68.2 %	59.1 %	57.2 %	66.2 %	68.9 %
Poland	56.6 %	55.9 %	61.5 %	60.8 %	51.3 %	54.2 %	62.1 %	60.2 %
Russian Fed.	53.7 %	52.9 %	56.5 %	58.4 %	52.5 %	50.4 %	57.5 %	56.0 %
Slovak Rep.	54.5 %	55.4 %	67.2 %	66.9 %	53.8 %	53.9 %	67.0 %	66.7 %
Spain	48.4 %		55.7 %		44.8 %		55.4 %	
Sweden	62.4 %	64.7 %	69.7 %	70.6 %	58.5 %	61.9 %	67.0 %	68.9 %
United States	57.8 %	56.7 %	56.9 %	58.8 %	52.1 %	54.9 %	56.8 %	55.0 %
Round 2								
Chile	34.4 %	34.5 %	42.8 %	37.0 %	29.8 %	28.3 %	40.0 %	38.0 %
Greece	44.6 %	42.0 %	55.7 %	57.5 %	38.7 %	41.1 %	52.5 %	53.6 %
Israel	51.1 %	51.4 %	60.4 %	59.5 %	46.9 %	47.8 %	59.1 %	60.9 %
Lithuania	47.6 %	48.9 %	64.3 %	65.4 %	47.9 %	48.3 %	62.2 %	61.3 %

New Zealand	57.4 %	56.9 %	64.0 %	62.7 %	54.1 %	55.4 %	61.4 %	61.8 %
Singapore	54.7 %	53.5 %	66.7 %	64.7 %	51.8 %	51.6 %	65.6 %	67.8 %
Slovenia	48.2 %	48.9 %	60.9 %	63.3 %	45.0 %	46.0 %	61.0 %	60.8 %
Turkey	33.9 %	34.7 %	47.1 %	46.8 %	30.5 %	33.6 %	47.2 %	47.4 %
Average across Round 1 and Round 2 countries								
$Average_1$	55.3 %	55.2 %	63.4 %	63.4 %	52.2 %	52.7 %	62.3 %	62.1 %
$Average_2$	49.4 %		58.4 %		45.8 %		57.7 %	

Note. $Average_1$ is based on the countries that participated in the PSTRE domain. $Average_2$ is based on the countries that did not participated in the PSTRE domain. Jakarta (Indonesia) received only PBA forms and is, therefore, not included in this table.

Efficiency of the MST design

Figure 2 shows the relative efficiency of the PIAAC MST for Literacy and Numeracy over an average of linear tests of equal length. The ratio efficiency gain is shown on the vertical axis, whereas the Literacy and the Numeracy scales are shown on the horizontal axis. The MST is shown to be 10–30 % more efficient for Literacy and 4–31 % more efficient for Numeracy compared to the nonadaptive linear tests. This means that we can obtain the same amount of test information as we might expect from a test that is 10–30 % longer with regard to Literacy, and 4–31 % percent longer with regard to Numeracy.

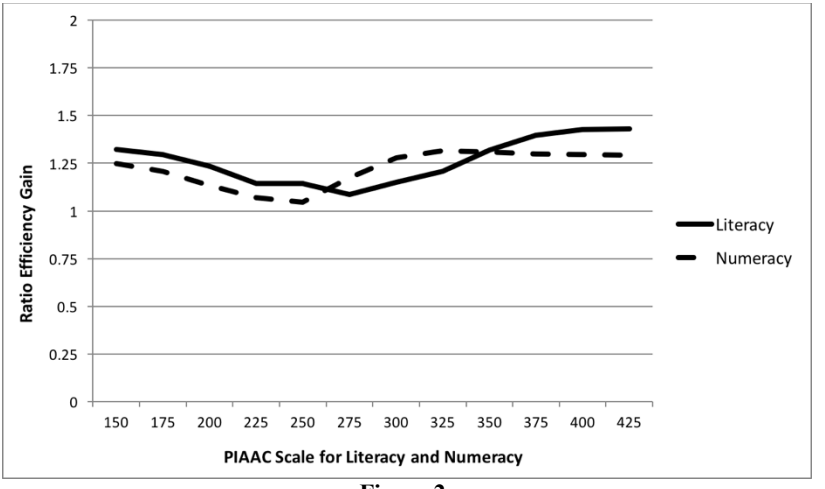

Figure 2:
Efficiency of the multistage adaptive testing model of the Literacy and Numeracy scale used in PIAAC.

There is no proficiency range where adaptive testing is less informative, with more gains for extreme scale scores. This improvement in measurement precision was one of the major goals of the MST design in PIAAC.

Item-By-Country and Item-By-Language interactions

For Literacy, unique parameters were estimated for 8 % of item-by-country interactions in Round 1 and 6 % in Round 2. For Numeracy, unique item parameters were estimated for 7 % of item-by-country interactions in Round 1 and 3 % in Round 2. For PSTRE, unique item parameters were estimated for 3 % of item-by-country interactions in Round 1 and 3.6 % in Round 2. (Although PSTRE was embedded in the MST design but not administered adaptively, its results are still reported here.) Overall, a high comparability of the item parameters across countries and languages could be achieved: 92 % and 94 % for Literacy and 93 % and 97 % for Numeracy.

Discussion

Based on the example of the PIAAC test design, this paper aims to illustrate the challenges and possible solutions for introducing adaptive testing into international large-scale assessments. Such assessments need to meet certain goals and standards and have various constraints. Usually, a sufficient and broad construct coverage needs to be ensured in all proficiency levels for a high comparability across groups within countries as well as across different countries. At the same time, comparability of the assessed constructs needs to be established for each country across assessment cycles over time to ensure a stable trend measure. The influence of context effects such as item position and mode effects as well as item-by-country/language interactions need to be largely mitigated or prevented. The PIAAC test design aims to account for all these constraints and to combine them with an adaptive test administration to increase test efficiency and accuracy, and to better meet various performance levels within and across countries.

The PIAAC MST design uses information from the background questionnaire (BQ) and the cognitive assessment and was based on two levels of adaptation: 1) Based on respondent's computer skills and experience assessed through a series of background questions as well as responses to core items, respondents were routed to either a paper-based (PBA) or a computer-based assessment (CBA); 2) within the CBA, respondents' proficiency levels with regard to responses to prior cognitive items as well as information about their educational level and native language was used to assign the different adaptive stages. In addition, a probability-based multistage adaptive algorithm was used to control the item exposure rate to enable a broad construct coverage and to minimize item-by-country interactions.

To ensure the success of the MST design in PIAAC and the comparability of item parameters, one aim was to minimize any possible effect of item position by balancing the item position through the order of modules (a module consists of two blocks; one block consists of several units; one unit consists of several items). Each module of Literacy and Numeracy items appeared in two different positions within the assessment. Results based on the main study data show a slight cluster position effect for Literacy modules (2.9 %) and Numeracy modules (1.2 %) on the percent of correct responses. However, the IRT scaling based on the 2PLM and GPCM provided comparable item parameters achieving high

comparability and measurement invariance (92 % and 94 % for Literacy and 93 % and 97 % for Numeracy in the PIAAC Round 1 and Round 2 assessments, respectively).

Finally, the MST design achieved a higher test efficiency compared to a nonadaptive design. It was shown to be 10–30 % more efficient for Literacy and 4–31 % more efficient for Numeracy, with higher test efficiency for higher and lower performing levels. An additional advantage of the MST design was the possibility to include different item types to ensure a broad construct coverage. Moreover, it is assumed that the use of item sets instead of individual items for adaptive decisions reduces the likely impact of item-by-country interactions (e.g., due to differential item functioning) on the adaptive path selection compared to item-level adaptive tests.

By implementing a MST design, PIAAC is able to provide more efficient and more accurate measures, especially for higher and lower performing respondents and countries. Thus, PIAAC provides policy makers and researchers not only with a rich but also more accurate source of information to understand the distributions of human capital in their country and the connections between these skills and important social, educational, and labor market outcomes. At the same time, the test design still meets the general goals and constraints of international large-scale assessments. It allows to establish a stable link across assessment modes, different countries and languages, and over time providing a stable trend measure. Overall, the MST design in PIAAC showed to be successful and can now serve as an example for how to prepare and implement adaptive testing in international-large scale assessments. The illustrated design was, of course, uniquely designed for PIAAC, but the rational and reasons behind this design (combining adaptive features with general requirements and restrictions of large-scale assessments) can be applied for other international large-scale assessments as well. The procedures and findings can also be used to establish MST for assessments focusing on individual test scores if a sufficient construct coverage at the individual level has to be achieved (i.e., with regard to assessing different constructs or different subscales of one construct) and if the impact of differential item functioning on the adaptive path decision is a possibility and should be reduced.

A limitation of MST in general is that it is not purely adaptive, in other words, not adaptive for every item, which limits the gain in measurement precision. However, the PIAAC test design combines a number of design constraints typical for international large-scale assessments with the advantages of adaptive testing. This led to certain limitations with regard to the adaptiveness of the design while, on the other hand, improving the stability of parameter estimation and comparability of test scores. The next PIAAC cycle will include a larger item pool and further refinements of the adaptive procedures for less proficient respondents with the aim to achieve an increase in measurement precision for respondents with a wider range of proficiencies.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695–716. doi: 10.1111/j.1744-6570.1990.tb00679.x

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*, 92–104.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137–149. doi: 10.1207/s15324818ame0502_4

Betz, N. E., & Weiss, D. J. (1976). *Psychological effects of immediate knowledge of results and adaptive ability testing* (Research Report No. 76-4). Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

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

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 433–448). New York, NY: Springer-Verlag. doi: 10.1007/978-1-4757-2691-6_25

Breithaupt, K. J., Zhang, O. Y., & Hare, D. R. (2014). The multistage testing approach to the AICPA uniform certified public accounting examinations. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 343–354). Boca Raton, FL: Chapman and Hall/CRC.

Chen, H., Yamamoto, K., & von Davier, M. (2014). Controlling MST exposure rates in international large-scale assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 391–409). Boca Raton, FL: Chapman and Hall/CRC.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Enschede, Netherlands: Print Partners Ipskamp.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York, NY: Springer.

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Handling of item position effects in the development of computerized adaptive tests]. *Diagnostica*, *63*, 167–178. doi: 10.1026/0012-1924/a000173

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests [Effects of adaptive testing on test-taking motivation in the example of the Frankfurt adaptive test for measuring attention]. *Diagnostica, 55*, 20–28. doi: 10.1026/0012-1924.55.1.20

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.* Boca Raton, FL: CRC Press.

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. 69–95). New York, NY: Springer. doi: 10.1007/978-1-4612-4230-7_5

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff, Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto, *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11), Dordrecht, Netherlands: Springer. doi: 10.1007/978-94-007-4629-9_1

Kirsch, I., & Thorn, W. (2013). The Programme for International Assessment of Adult Competencies – an Overview. In Organisation for Economic Co-operation and Development (2013), *Technical Report of the Survey of Adult Skills (PIAAC)*, foreword (pp. 5–24), PIAAC, OECD Publishing. Retrieved from http://www.oecd.org/site/piaac/All%20PIACC%20Technical%20Report%20final.pdf

Kirsch, I., & Yamamoto, K. (2013). PIAAC assessment design. In Organisation for Economic Co-operation and Development (2013), *Technical Report of the Survey of Adult Skills (PIAAC)*, chapter 1 (pp. 27–43), PIAAC, OECD Publishing. Retrieved from http://www.oecd.org/site/piaac/All%20PIACC%20Technical%20Report%20final.pdf

Kubinger, K. D. (2016). Adaptive testing. In K. Schweizer & C. DiStefano (Eds.), *Principles and Methods of Test Construction. Standards and Recent Advances*. *Psychological Assessment – Science and Practice, Vol. 3* (pp. 104–119). Göttingen, Germany: Hogrefe.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement, 41*, 495–511. doi: 10.1177/0146621617707556

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227–242. doi: 10.1007/BF02297844

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. doi: 10.1002/j.2333-8504.1992.tb01436.x

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53(*3), 315–333. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. doi: 10.1080/15305058.2013.825265

Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group-score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 371–390). Boca Raton, FL: Chapman and Hall/CRC.

Organisation for Economic Co-operation and Development (2013). *Technical report of the Survey of Adult Skills (PIAAC)*, Ch. 17 (pp. 406-438). Retrieved from http://www.oecd.org/site/piaac/Technical%20Report_17OCT13.pdf

Organisation for Economic Co-operation and Development (2016). *Skills matter: Further results from the Survey of Adult Skills*. doi: 10.1787/23078731

Ponsoda, V., Olea, J., Rodriguez, M. S., & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education, 12,* 167–184. doi: 10.1207/s15324818ame1202_4

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design* (Law School Admission Council Computerized Testing Report No. 97-02). Princeton, NJ: Law School Admission Council.

Robin, F., Steffen, M., & Liang, L. (2014). The multistage test implementation of the GRE® revised General Test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 325–341). Boca Raton, FL: Chapman and Hall/CRC.

Setzer, J. C., & Allspach, J. R. (2007). *Studying the effect of rapid guessing on a low-stakes test: An application of the effort-moderated IRT model.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Linden, W. J., & Hambleton, R. K. (2016). *Handbook of modern item response theory* (2nd ed). New York, NY: Springer.

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

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press.

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*. Amsterdam, Netherlands: Elsevier. doi: 10.1016/S0169-7161(06)26032-2

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report No. 74-5). Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing, 2*, 1–17. doi: 10.7333%2Fjcat.v2i0.30

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. doi: 10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183. doi: 10.1207/s15324818ame1802_2

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*(4), 15–27. doi: 10.1111/j.1745-3992.2010.00190.x

Yamamoto, K. (1997) Scoring, scaling, and statistical models for proficiency estimation of the IALS. *International Adult Literacy Survey Technical Report*. Ottawa, Canada: Statistics Canada.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC Cognitive Data. In Organisation for Economic Co-operation and Development (2013), *Technical Report of the Survey of Adult Skills (PIAAC)* (pp. 406–438), OECD Publishing. Retrieved from http://www.oecd.org/site/piaac/All%20PIACC%20Technical%20Report%20final.pdf

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155–174. doi: 10.3102/10769986017002155

Zenisky, A., Hambleton, R. K., & Luecht, R. (2010). Multistage testing: Issues, designs and research. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Berlin, Germany: Springer. doi: 10.1007/978-0-387-85461-8_18