IEA PIRLS 2021

CHAPTER 3

PIRLS 2021 Assessment Design

Michael O. Martin, Matthias von Davier, Pierre Foy and Ina V. S. Mullis

In I.V.S. Mullis & M. O. Martin (Eds.), *PIRLS 2021 Assessment Frameworks*.

IEA
TIMSS & PIRLS International Study Center
Lynch School of Education
BOSTON COLLEGE

CHAPTER 3

PIRLS 2021 Assessment Design

Michael O. Martin, Matthias von Davier, Pierre Foy and Ina V. S. Mullis

Overview

PIRLS 2021 is a comprehensive assessment of fourth grade students' reading literacy achievement. Conducted on a regular five-year cycle, with each assessment linked to those that preceded it, PIRLS provides regular data on trends in students' reading literacy on a common achievement scale. PIRLS 2021 begins the transition from paper-and-pencil to digital format, with about half the countries choosing to administer the digital format (digitalPIRLS) and half the paper format (paperPIRLS). To ensure comparability across formats, digitalPIRLS and paperPIRLS have the same content in terms of reading passages and questions, although digitalPIRLS takes advantage of some features and item types not available in paper and pencil mode. Countries choosing digitalPIRLS will also administer ePIRLS, an assessment of online reading to acquire and use information first conducted as part of PIRLS 2016. Similar to previous PIRLS assessments, PIRLS 2021 includes a series of contextual questionnaires to gather information about community, home, and school contexts for developing reading literacy.

Based on PIRLS experience in earlier assessment cycles, it is clear that achievement levels in reading comprehension vary widely both across and within countries, posing a challenge in matching the difficulty of the assessment to the reading ability of the students in every country. When an assessment is much too difficult or too easy for a population, little information is available to accurately measure performance. To address this challenge, PIRLS in the past has offered less difficult versions of its assessment materials that countries could choose to administer, beginning with prePIRLS in 2011 and following up with PIRLS Literacy in 2016. These efforts were successful in expanding PIRLS coverage of students at the lower end of the ability distribution but required separate (although linked) versions of PIRLS and did not address the need for more challenging material for higher achieving students.

PIRLS 2021 addresses the need for a broader range of assessment difficulty and better targeting of student ability by adopting a single unified assessment based on a new group adaptive assessment design (Click here for the rationale underlying group adaptive designs). The new design is based

on having three levels of passage difficulty—difficult, medium, and easy—that are combined into two levels of booklet difficulty. More difficult booklets are composed of two difficult passages or one medium and one difficult passage while less difficult booklets consist of an easy and a medium passage or two easy passages. Each country administers the entire assessment, but the balance of more difficult and less difficult booklets varies with the reading achievement level of the students in the country. For example, a country with higher average reading achievement (average score of 550 or above) could assign the more difficult booklets to 70% of its students and the less difficult booklets to 30% of its students, whereas a country with lower achievement levels (average score of 450 or below) could assign the more difficult booklets to 30% of its students and the less difficult booklets to 70%.

The group adaptive design improves the match between assessment difficulty and student ability in each country's population by having a greater proportion of more difficult booklets in countries with relatively high achievement and a greater proportion of less difficult booklets in countries with relatively low achievement. Accordingly, the new design maximizes the information obtained from the assessment while minimally changing existing procedures and time requirements.

Although the group adaptive design was developed to provide a better match between assessment difficulty and student ability at the country level, it also is possible to apply the approach within a country, provided the country has clearly defined subpopulations that differ substantially in student achievement.

Student Population Assessed

PIRLS assesses the reading literacy of children in their fourth year of formal schooling. This student population was chosen for PIRLS because it is an important transition point in children's development as readers. Typically, at this point, students have mastered the basics of learning to read and are now reading to learn. In many countries, this also is when students begin to have separate classes for different subjects, such as mathematics and science.

PIRLS defines the fourth year of formal schooling according to the International Standard Classification of Education (ISCED) developed by the UNESCO Institute for Statistics.[1] The ISCED classification provides an international standard for describing levels of schooling across countries, and covers the full range of schooling, from early childhood education (Level 0) to doctoral study (Level 8). ISCED Level 1 corresponds to primary education, or the first stage of basic education, and is considered to be the first stage of formal schooling.

The target population for PIRLS is defined as follows:

> The PIRLS target grade should be the grade that represents four years of schooling, counting from the first year of ISCED Level 1.

The PIRLS target grade is four years after the beginning of Level 1, which is the fourth grade in most countries. However, given the linguistic and cognitive demands of reading, PIRLS wants to avoid assessing very young children. Thus, if the average age of fourth grade students at the time of testing would be less than 9.5 years, PIRLS recommends that countries assess the next higher grade (i.e., fifth grade).

Reporting Reading Achievement

The PIRLS assessment is designed to provide a complete picture of the reading literacy achievement of the participating students in each country. This includes achievement by reading purpose and comprehension process as well as overall reading achievement. Consistent with the goal of a comprehensive view of reading comprehension, the entire PIRLS 2021 assessment, digital and paper versions, consists of 18 reading passages and accompanying questions (known as items), half assessing reading for literary experience and half assessing reading to acquire and use information. In accordance with the group adaptive design, one third of the passages are relatively difficult, one third of medium difficulty, and one third relatively easy. Countries administering digitalPIRLS also administer ePIRLS, which consists of five tasks that assess online informational reading.

In order to keep the assessment burden on any one student to a minimum, each student is presented with just two passages, one literary and one informational, according to a systematic booklet assembly and rotation procedure, as described in the next section. In digitalPIRLS countries, some students also are presented with ePIRLS material, either two ePIRLS tasks, or one digitalPIRLS informational passage followed by one ePIRLS task. The PIRLS administration consists of two 40-minute sessions, one for each passage or task, separated by a short break, and followed by a 30-minute session for the student questionnaire. Following data collection, student responses to the assessment passages are placed on the PIRLS reading achievement scales using item response theory methods that provide an overall picture of the assessment results for each country.[2]

PIRLS was designed from the outset to measure trends over time in student reading achievement. Accordingly, the PIRLS reading achievement scale provides a common metric on which countries can compare their fourth grade students' progress in reading over time from assessment to assessment. The PIRLS achievement scale was established in 2001 so that 100 points on the scale corresponded to one standard deviation across all of the countries that participated in 2001, and the scale centerpoint of 500 corresponded to the international average across those countries. Using passages that were administered in both the 2001 and 2006 assessments as a basis for linking the two sets of assessment results, the PIRLS 2006 data also were placed on this scale so that countries could gauge changes in students' reading achievement since 2001. Following a similar procedure, the PIRLS 2011 and PIRLS 2016 data also were placed on the PIRLS scale, as will be the data from PIRLS 2021. This will enable countries that have participated in PIRLS since its inception to

have comparable achievement data from 2001, 2006, 2011, 2016, and 2021, and to plot changes in performance over this 20-year period.

The PIRLS reading achievement scale is an overall measure of reading proficiency that includes both reading purposes and processes of comprehension. However, in addition to the overall scale, PIRLS also provides separate achievement scales on the same metric for purposes for reading and for processes of comprehension. More specifically, there are two scales for reading purposes:

- Reading for literary experience; and
- Reading to acquire and use information.

In addition to these, there also are two scales for processes of reading comprehension:

- Retrieval and straightforward inferencing; and
- Interpreting, integrating, and evaluating.*

Countries participating in digitalPIRLS also administer ePIRLS; so, in addition to the usual PIRLS overall reading achievement results and results by reading purpose and comprehension process, in countries participating in digitalPIRLS, student achievement will also be reported for online informational reading. The ePIRLS online reading achievement scale was established in 2016 to enable countries to examine their students' online reading performance relative to their performance on the PIRLS reading achievement scales.

PIRLS 2021 Group Adaptive Design

Implementing the group adaptive design in PIRLS 2021 required grouping the assessment passages and items into three levels of difficulty—easy, medium, and difficult—with three literary and three informational passages at each level of difficulty. Of the 18 passages needed for the design, 12 were administered previously in PIRLS 2016 and were available to support the measurement of trends and six were developed and field tested for first time use in PIRLS 2021.

In 2016, the main PIRLS assessment was accompanied by PIRLS Literacy, which was a reading assessment similar to PIRLS in size and scope but less difficult, with shorter and less demanding passages and easier items. Taking PIRLS and PIRLS Literacy together, there were eight passages (four literary and four informational) that appeared only in PIRLS, four shared passages (two for each purpose) that appeared in both PIRLS and PIRLS Literacy, and eight passages (again four for each purpose) that appeared in PIRLS Literacy only. The 12 trend passages for 2021 were chosen from among these passages, with the difficult passages drawn from the PIRLS only passages, the medium passages from the passages shared between PIRLS and PIRLS Literacy (with one exception), and the easy passages from those used in PIRLS Literacy only.

* Retrieval and straightforward inferencing combines items from the Focus on and Retrieve Explicitly Stated Information and Make Straightforward Inferences comprehension processes. Similarly, interpreting, integrating, and evaluating is based on items from the Interpret and Integrate Ideas and Information and Examine and Critique Content and Textual Elements processes.

Exhibit 1: Reading purpose and difficulty level for 18 PIRLS 2021 passages

Passages

Reading Purpose	Difficulty Level	Passage Label*	Passage Name*
Literary	Difficult	LitD1 (06)	Shiny Straw (06)
		LitD2 (16)	Oliver and the Griffin (16)
		LitD3 (21)	New LitD3 Passage (21)
	Medium	LitM1 (16)	Pemba Sherpa (16)
		LitM2 (21)	New LitM2 Passage (21)
		LitM3 (11)	The Empty Pot (11)
	Easy	LitE1 (21)	New LitE1 Passage (21)
		LitE2 (11)	The Summer My Father Was 10 (11)
		LitE3 (16)	Library Mouse (16)
Informational	Difficult	InfD1 (11)	Where's the Honey? (11)
		InfD2 (16)	Icelandic Horses (16)
		InfD3 (21)	New InfD3 Passage (21)
	Medium	InfM1 (16)	How Did We Learn to Fly? (16)
		InfM2 (21)	New InfM2 Passage (21)
		InfM3 (06)	Sharks (06)
	Easy	InfE1 (21)	New InfE1 Passage (21)
		InfE2 (11)	Training a Deaf Polar Bear (11)
		InfE3 (16)	Hungry Plant (16)

* The number in parentheses is the assessment year in which the passage was first used.

Exhibit 1 shows how the existing trend passages fit into the purpose-by-difficulty level scheme, and also where the new passages and items belong. Of the six new passages, three will be literary and three informational, with one of each in each of the difficulty categories. The exhibit also includes a passage label for each passage to facilitate assignment of passages to booklets.

Passage Difficulty Level

For the design to be effective, it is necessary that there be distinct differences between the average difficulties of the passages in each of the passage groups. For example, reasonable difficulty goals in terms of average percent correct across the student population would be 40% for the difficult group, 60% for the medium group, and 80% for the easy group. New passages developed for PIRLS 2021 will aim for these difficulty levels, but there is less flexibility with the existing, trend passages, which make up two-thirds of the passage total.

Exhibit 2: Average difficulties of existing trend passages from 2016 and target difficulties for 2021 (average percent correct)

Passage Level	Difficulty of Trend Passages from 2016	Target Difficulty for 2021
Difficult	55%	50%
Medium	68%	65%
Easy	80%	80%

As shown in Exhibit 2, the difficulties of the existing medium and difficult trend passages (68% and 55%, respectively) are higher than the long-term goals (60% and 40%, respectively), especially for the difficult passages. However, by combining the existing passages with new passages developed to be closer to the target difficulty, it will be possible to make progress in 2021 toward these long-term goals. Exhibit 2 shows the interim target difficulties for 2021, which should be attainable given the mix of existing and new passages. Further progress will be made toward the long-term goals in PIRLS 2026, as more of the trend passages are replaced by new, more targeted passages.

Booklet Design

In the PIRLS assessment, each student is randomly assigned a test booklet (or booklet equivalent in digitalPIRLS) consisting of two passages and their items. In PIRLS 2021, the 18 passages are arranged into 18 booklets of two passages each, with each passage appearing in two booklets and paired with a different passage each time. Exhibit 3 summarizes the passage pairs that make up each booklet. The direction of the arrows shows which passage comes first in the booklet. For example, an arrow points from passage InfM1 to LitD1, indicating that these two passages share a booklet, with InfM1 preceding LitD1. Note that when passages of different difficulties are paired in the same booklet, the easier of the two always comes first.

Exhibit 3: Passage pairings for each assessment booklet

Passages and Links

Reading Purpose	Difficult Passages & Items	Medium Passages & Items	Easy Passages & Items
Literary	LitD1 (11)	LitM1 (16)	LitE1 (21)
	LitD2 (16)	LitM2 (21)	LitE2 (11)
	LitD3 (21)	LitM3 (11)	LitE3 (16)
Informational	InfD1 (11)	InfM1 (16)	InfE1 (21)
	InfD2 (16)	InfM2 (21)	InfE2 (11)
	InfD3 (21)	InfM3 (11)	InfE3 (16)

The 18 booklets are divided into two levels of difficulty, as follows:

- More difficult booklets (9) composed of either two difficult passages or one medium and one difficult passage
- Less difficult booklets (9) composed of two easy passages or one easy and one medium passages.

Exhibit 4 shows the passage assignments for the 18 booklets, with booklets 1-9 being the more difficult booklets and booklets 10-18 the less difficult ones.

Exhibit 4: Assessment booklets with passage assignments

Student Assessment Booklets		Part 1	Part 2
More Difficult Booklets	Booklet 1	InfM1 (16)	LitD1 (11)
	Booklet 2	LitD3 (21)	InfD2 (16)
	Booklet 3	LitM1 (16)	InfD1 (11)
	Booklet 4	InfM2 (21)	LitD2 (16)
	Booklet 5	LitD1 (11)	InfD3 (21)
	Booklet 6	LitM2 (21)	InfD2 (16)
	Booklet 7	InfM3 (11)	LitD3 (21)
	Booklet 8	InfD1 (11)	LitD2 (16)
	Booklet 9	LitM3 (11)	InfD3 (21)
Less Difficult Booklets	Booklet 10	LitE1 (21)	InfM1 (16)
	Booklet 11	InfE2 (11)	LitM2 (21)
	Booklet 12	InfE1 (21)	LitE3 (16)
	Booklet 13	LitE2 (11)	InfM2 (21)
	Booklet 14	InfE3 (16)	LitM3 (11)
	Booklet 15	LitE1 (21)	InfE2 (11)
	Booklet 16	LitE3 (16)	InfM3 (11)
	Booklet 17	InfE1 (21)	LitM1 (16)
	Booklet 18	LitE2 (11)	InfE3 (16)

Exhibit 5 also presents the passage assignments for each booklet, this time showing where the 12 existing trend passages belong and where the new passages developed for 2021 will go.

Exhibit 5: Assessment Booklets with Trend and New Passage Assignments

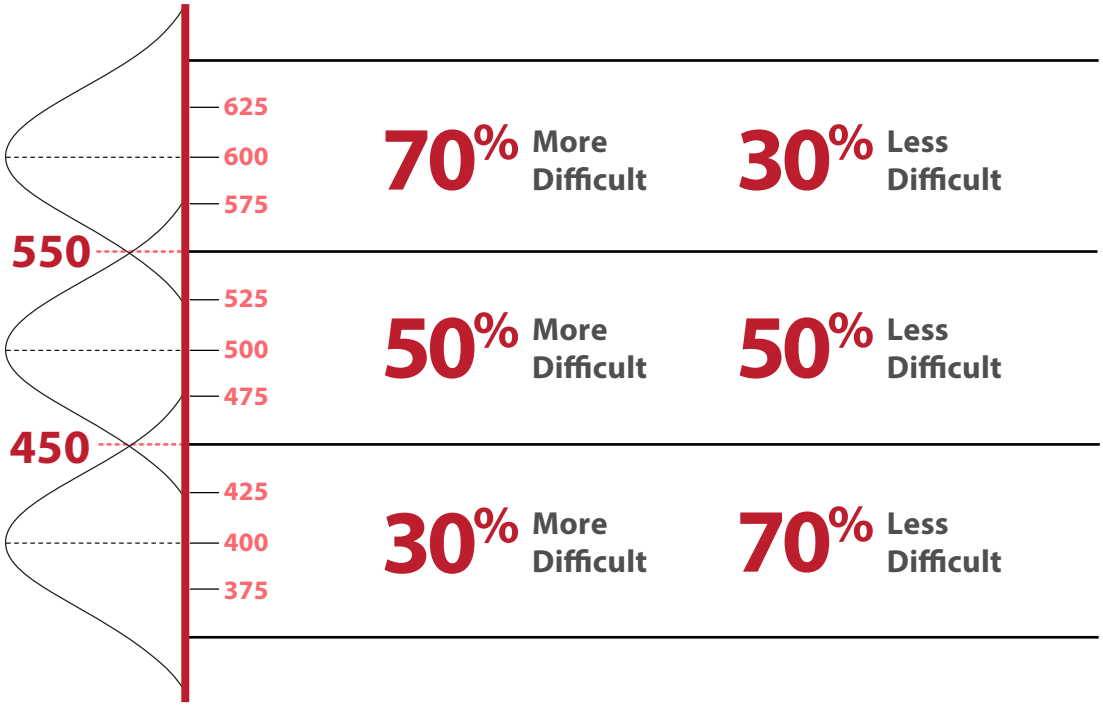
Student Assessment Booklets		Part 1	Part 2
More Difficult Booklets	Booklet 1	How Did We Learn to Fly? (16)	Shiny Straw (06)
	Booklet 2	New LitD3 Passage (21)	Icelandic Horses (16)
	Booklet 3	Pemba Sherpa (16)	Where's the Honey? (11)
	Booklet 4	New InfM2 Passage (21)	Oliver and the Griffin (16)
	Booklet 5	Shiny Straw (06)	New InfD3 Passage (21)
	Booklet 6	New LitM2 Passage (21)	Icelandic Horses (16)
	Booklet 7	Sharks (06)	New LitD3 Passage (21)
	Booklet 8	Where's the Honey? (11)	Oliver and the Griffin (16)
	Booklet 9	The Empty Pot (11)	New InfD3 Passage (21)
Less Difficult Booklets	Booklet 10	New LitE1 Passage (21)	How Did We Learn to Fly? (16)
	Booklet 11	Training a Deaf Polar Bear (11)	New LitM2 Passage (21)
	Booklet 12	New InfE1 Passage (21)	Library Mouse (16)
	Booklet 13	The Summer My Father Was 10 (11)	New InfM2 Passage (21)
	Booklet 14	Hungry Plant (16)	The Empty Pot (11)
	Booklet 15	New LitE1 Passage (21)	Training a Deaf Polar Bear (11)
	Booklet 16	Library Mouse (16)	Sharks (06)
	Booklet 17	New InfE1 Passage (21)	Pemba Sherpa (16)
	Booklet 18	The Summer My Father Was 10 (11)	Hungry Plant (16)

Booklet Assignment within Countries

To ensure that the same assessment is conducted in every country, all 18 passages are distributed in every country, but with varying proportions of the more and less difficult booklets depending on the average reading ability of the student population. This is estimated based on performance in prior PIRLS assessments, or in the field test for countries participating for the first time. Higher performing countries sample proportionally more of the more difficult booklets while lower performing countries sample proportionally more of the less difficult booklets, with the goal of a better match between assessment difficulty and student ability in each country.

Exhibit 6 illustrates the differential booklet assignment plan for higher, middle, and lower performing countries. Countries with higher average performance, above 550 on the PIRLS achievement scale, would randomly assign proportionally more of the more difficult booklets, e.g., 70%, and fewer of the less difficult booklets, e.g., 30%. Countries with performance between 450 and 550 would assign equal proportions of more and less difficult booklets, and countries with lower average performance, below 450 on the PIRLS scale, would assign proportionally fewer of the more difficult booklets (30%) and more of the less difficult booklets (70%).

Exhibit 6: **Booklet assignment plan for higher, middle, and lower performing countries**



Integrating ePIRLS with digitalPIRLS

The ePIRLS assessment of online informational reading in 2021 consists of five tasks presented by computer or tablet. With the guidance of a teacher avatar, students navigate within and across webpages to answer questions, explain relationships, and interpret and integrate information. Three of the tasks were administered as part of ePIRLS 2016 and are available for measuring trends and two were developed for PIRLS 2021. As a step toward the future further integration of PIRLS and ePIRLS, in digitalPIRLS countries ePIRLS tasks are included in the booklet assignment rotation scheme, as are a number of "hybrid" booklets consisting on one digitalPIRLS informational passage followed by one ePIRLS task. This means that students participating in digitalPIRLS may be presented with one of three booklet types: a regular booklet with two digitalPIRLS passages, an ePIRLS booklet with two ePIRLS tasks, or a hybrid booklet with one digitalPIRLS informational passage followed by one ePIRLS task.

digitalPIRLS Booklet Assignment Rotation Schemes

The basic assignment rotation scheme for digitalPIRLS is designed to match that of paperPIRLS, and so has the same 18 passages arranged in the same 18 booklets as paperPIRLS. In paperPIRLS these 18 booklets are distributed among the students in sampled classes using a systematic random

assignment process that ensures that the proportions of more and less difficult booklets conform to the rates established for the country. digitalPIRLS follows a similar assignment process for the 18 regular booklets, but also includes provision for the ePIRLS and hybrid booklets.

Similar to the ePIRLS assignment rotation scheme used in 2016, each of the five ePIRLS tasks for 2021 is paired with each of the others, with each task appearing in both the first and second positions in each booklet. This results in 20 distinct ePIRLS booklets. In the hybrid booklets, each of the nine digitalPIRLS informational passages is paired with each of the five ePIRLS tasks, resulting in 45 hybrid booklets. In each one, the digitalPIRLS passage precedes the ePIRLS task.

Combining the 18 digitalPIRLS booklets, the 20 ePIRLS booklets, and the 45 hybrid booklets results in a total of 83 booklets to be rotated among the sampled students. However, including each booklet type in the rotation at the same rate would result in too few digitalPIRLS booklets relative to the other booklet types, and especially relative to the number of such booklets in paperPIRLS countries. To avoid this imbalance, digitalPIRLS uses a 27 booklet rotation that includes all 18 digitalPIRLS booklets, six ePIRLS booklets, and three hybrid booklets. The assignment of booklets to individual students is conducted by IEA's WinW3S within school sampling software to ensure accurate implementation of the rotation scheme.

Beginning with a booklet chosen at random by WinW3S, the first iteration of this rotation involves all 18 digitalPIRLS booklets, booklets 1 to 6 of the 20 ePIRLS booklets, and booklets 1 to 3 of the hybrid booklets. The second iteration includes the 18 digitalPIRLS booklets again, together with ePIRLS booklets 7 to 12 and hybrid booklets 4 to 6. This pattern continues with the next iteration including the 18 digitalPIRLS booklets once again, together with ePIRLS booklets 13 to 18 and hybrid booklets 7 to 9, and so on throughout the country's entire student sample. This booklet assignment scheme results in two thirds of the student sample responding to digitalPIRLS booklets, two ninths to ePIRLS booklets, and one ninth to hybrid booklets.

The PIRLS 2021 group adaptive design represents a change from the design of previous PIRLS assessments, where booklets of approximately equal difficulty were distributed at random among students in sampled classes with equal probability in each country. By allowing booklets that vary in difficulty to be assigned at country specific rates, the group adaptive design aims to improve the accuracy of measurement in countries participating in PIRLS.

References

1 UNESCO. (2012). *International Standard Classification of Education ISCED 2011*. Retrieved from uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

2 Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 Achievement Data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in PIRLS 2016* (pp. 12.1-12.38). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-12.html

APPENDIX A

Rationale for Group Adaptive Designs in International Large Scale Assessment

Different test forms (booklets) are commonly used in large scale international assessments such as PIRLS to balance respondent burden and content coverage. Country level group adaptive assessment designs extend this approach through targeted sampling of booklets to provide better coverage of the diverse range of ability distributions encountered in such assessments. This can increase student motivation and reduce item level nonresponse. The PIRLS 2021 approach is designed to minimally change existing procedures and time requirements, while using prior data about country performance to maximize the information obtained from the assessment.

The basic idea behind adaptive assessment is that in order to enable any type of measurement, tasks must not be too easy or too difficult for the target population. If the tasks given to a sample of test takers are too difficult, nobody (or almost nobody) will be able to solve them. Similarly, if the tasks are too easy, everybody will answer them all correctly. In each of these situations all test takers receive the same observed scores, even if they are known to differ with respect to relevant skills.

For this reason, in educational and psychological measurement we try to craft test questions that match the ability of the targeted population of test takers, and quantify differences among test takers by eliciting responses that differentiate between higher and lower skilled respondents. A series of tasks that matches the skills of test takers will likely result in some correct and some incorrect responses. Mathematically, the variability of such a binary response (choice of correct versus incorrect option) is maximized when there is a 50% chance to get the tasks right. This 50/50 criterion leads to different requirements for different test takers. More proficient test takers require more challenging questions in order to have (only) a 50% chance, while less proficient test takers require a series of easier tasks to arrive at a 50% chance of correct responses. In order to achieve this optimal match for all test takers, it would be necessary to adjust the test difficulty for each individual respondent. However, since this is only possible if the exact difficulty of all items is known (or can be estimated well with little error), many testing programs instead rely on a variation of this individual level adaptivity and adapt their tests according to the known, or estimated, average ability levels of pre-defined groups, rather than individuals.

Existing Approaches

Country level adaptive assessment designs target particular booklets to specific populations in order to match ability distributions with the distribution of booklets. There are various approaches and assessment designs that adapt the assignment of tests to differences in target populations with respect to the distribution of skills, which may be estimated by previous routing instruments or inferred from variables such as age or educational attainment. The following section describes major approaches for adapting the difficulty of tests to the ability of the test taking populations.

Starting Rules and Discontinue Rules

In intelligence testing for individuals, for adult as well as child and adolescent populations, it is common to design tests that present items in the order of increasing difficulty (e.g., the Stanford Binet Intelligence Scale[1]). When first applying these tests to different age groups, it was soon noticed that the first few questions were not much of a challenge for older test takers, as they would get the first handful or so questions right in almost all cases. This led test administrators to skip these first few very easy items as they 'knew' (i.e., made an inference on cases observed so far) that older test takers would get these easy questions right. Along the same lines, it also became apparent that for younger test takers, there was a point in these tests where the remaining, harder questions were almost impossible to solve. This in turn led test administrators to stop presenting items that experience had shown to be too difficult.

Many tests of this type have a rule about how many items a test taker must get wrong consecutively before the testing session can be terminated. This number varies typically from 23 items for short forms to 5-6 items for longer IQ tests. It can be shown that the discontinued items (those for which no response was recorded after a pre-determined number of consecutive wrong responses) are missing data that is ignorable[2] and that the data only on what students actually took is sufficient to estimate ability.

Multistage Adaptive Testing

Multistage adaptive testing (MST) has been used in large scale international studies for adult populations[3] and can be understood as a flexible approach to assign test takers to a fixed number of test forms while aiming for a good, if not perfect, match between respondent ability and test difficulty.[4] In multistage adaptive testing, the completely randomized assignment of blocks to test takers (the previous practice in TIMSS, PIRLS, and PISA) is modified to take into account the performance of the test taker on a previous block, as well as the relative difficulty of the blocks contained in the test design.

In the beginning of the assessment, some form of preliminary ability estimate is required for each test taker so that he or she can be assigned item blocks that match their expected performance.

The assignment can be done deterministically, based on fixed cut-off scores, or probabilistically, based on a preliminary estimate of the student's ability distribution. Choosing the next block probabilistically ensures that at least some easy, medium and hard blocks are likely to be still available to all respondents at subsequent stages of the test. It also allows the assignment probability to be adjusted at each stage so that weak performance on a block is more likely to result in an easy block being assigned next, while it is still possible, with lower probability, to be presented with a medium or even hard block of items. Along the same lines, following strong performance on earlier blocks the probability of being administered a hard block of items increases, while the probability of easier blocks decreases.[5]

The drawback of most multistage adaptive designs is that the initial starting point is either not adaptive because nothing is known about test takers, or it requires an initial routing block that produces a very rough first estimate of proficiency based on a short block of items. This estimate is somewhat error prone, particularly in assessments administered to a wide and diverse set of populations, as it assumes that the item characteristics of the routing items are known without error. An alternative to this approach is to use prior information based on background data such as education and occupation or other socio-economic data.[6]

Adaptive Longitudinal Designs

Another example of how tests are adapted to different group level ability distributions is a design that is being used in longitudinal large scale skill surveys.[7] These designs use information on how test takers performed in prior assessment cycles to adaptively assign a more difficult test form to students who belong to a high performing group, and an easier test form to students who belong to a low performing group. These assessments are often 2 years apart,[8] so that the adaptation in this case uses information that dates back years in time. This approach turns out to be efficient as the performance at the group level is a reliable predictor of group performance at the next time point.

Pohl[9] describes these designs in more detail and discusses applications in multi-cohort longitudinal studies of student populations. Each assessment cycle determines which test form should be administered to which group based on information from prior data collections. Group membership is based on prior performance, which may itself have been estimated using a harder or easier form. Over assessment cycles this provides a sequence of test forms that are tailored to decrease the error of measurement in proficiency estimation. It does this by increasing the expected response variance by matching prior performance to test forms that elicit optimal levels of systematic, ability related responses variability in groups of test takers.

Group Adaptive Assessment in PIRLS 2021

Group adaptive assessment in PIRLS 2021 is implemented by dividing its 18 passages into three levels of passage difficulty – difficult, medium, and easy – and combining these into two levels of booklet difficulty:

- More difficult booklets (9) composed of difficult or medium and difficult passages
- Less difficult booklets (9) composed of easy or easy and medium difficult passages

In this approach, all countries administer all 18 passages, but in varying proportions. Higher performing countries will administer proportionally more of the more difficult booklets while lower performing countries will administer proportionally more of the less difficult booklets. The goal is a better match between assessment difficulty and student achievement in each country.

The group adaptive design in PIRLS 2021 involves changing from the procedure used in previous PIRLS cycles where booklets were randomly assigned to students at the same rate in each country to one where more or less difficult booklets are assigned at different rates in different countries. This change is intended to improve the accuracy of measurement in countries participating in PIRLS and provide some practical and operational advantages. More specifically, the PIRLS group adaptive design provides the following:

1. Better measurement at all achievement levels by matching booklet difficulty to student ability at the country level
2. All countries participate in the same assessment, maintaining full coverage of the reading construct while providing adaptivity at the population level
3. Minimal disruption of the PIRLS design as there is no need for a routing block under this approach
4. Improved student response rates, more student engagement, and less student frustration as passages are better aligned with target populations
5. Possibility of targeting subpopulations – although the PIRLS 2021 group adaptive design is intended to be implemented at the country level, it also could be implemented within countries that have clearly defined subpopulations that vary in student ability

As outlined in this paper there are ample examples of group-level adaptive approaches, from simple start/discontinue rules to elaborate longitudinal stage-based assessment designs. All these are based on group-level adaptivity that identifies groups of test takers which are to be assigned targeted test forms which are better aligned with the expected performance compared to complete random assignment or the use of only a single form.

The PIRLS group-adaptive design should benefit both high and low performing countries, in that students will be administered items that are either too difficult or too easy at a lower rate than in previous assessments. This improved targeting of the ability distributions will lead to more accurate measurement and will, as an intended side effect, likely also reduce item level non-response associated with administering too challenging or too easy items. Together, this is expected to lead to an overall improved database for reporting and secondary analyses.

References

1 Roid, G. & Barram, R. (2004). Essentials of Stanford–Binet Intelligence Scales (SB5) Assessment. Hoboken, New Jersey: John Wiley & Sons, Inc.

2 von Davier, M., Cho, Y., & Pan, T. (2019) Effects of Discontinue Rules on Psychometric Properties of Test Scores. Psychometrika, vol. 84, no. 1, 147–163. https://doi.org/10.1007/s11336-018-09652-3

3 Yamamoto, K., Khorramdel, L., & von Davier, M. (2013) Chapter 17: Scaling PIAAC Cognitive Data. In: Technical Report of the Survey of Adult Skills (PIAAC). Available at: http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf

4 Yan, D., von Davier, A., & Lewis, L. (2014; Eds.). Computerized multistage testing: Theory and applications (pp. 3–20). New York, NY: CRC Press.

5 Yamamoto, K., Chen, H., & von Davier, M. (2014) Controlling Multistage Testing Exposure Rates in International Large-Scale Assessments. Chapter 19 in: Yan, D., von Davier, A., & Lewis, L. (Eds.). Computerized multistage testing: Theory and applications (pp. 3–20). New York, NY: CRC Press.

6 Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Chapter 18: Scaling Outcomes. In: Technical Report of the Survey of Adult Skills (PIAAC). Available at: http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf Yamamoto et al. 2013

7 Rock D.A. (2017) Modeling Change in Large-Scale Longitudinal Studies of Educational Growth: Four Decades of Contributions to the Assessment of Educational Growth. In: Bennett R., von Davier M. (eds) Advancing Human Assessment. Methodology of Educational Measurement and Assessment. Springer.

8 ECLS: Rock, D. A. (2007a). A note on gain scores and their interpretation in developmental models designed to measure change in the early school years (Research Report No. RR-07-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02050.x

9 Pohl, S. (2014). Longitudinal Multistage Testing. JEM. https://doi.org/10.1111/jedm.12028

BOSTON
COLLEGE

pirls.bc.edu

IEA

TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE