Adaptive Testing Trade-Offs

Claudia Tamassia
Sr. Director, Research Project Management, Educational Testing Service

Washington, DC
4 February 2020
Context of large-scale international assessments

• Design criteria
  • Broader concept of comparability
  • Strong emphasis on innovation and technology
  • Design complexity
    • Strong emphasis on the integration of contextual and cognitive instruments
    • International collaboration
  • These types of surveys are growing both in terms of coverage and number of participants with the goal of assessing a wide range of proficiency across a diverse set of countries/economies.
  • They demand innovation and expertise in key areas.
Common approach

• Linear tests
  - The **same forms** are implemented across the entire (often heterogeneous) student distribution.
  - This represents the **most traditional** assessment design but can be quite **inefficient**.

• Characteristics
  - **Limited alignment** between students’ proficiency distribution and the difficulty of the test form, which is mostly centered around the middle of the distribution.
  - **Precision is not uniform across the entire range of test scores.** Scores at the top and bottom of the test's range, for example, generally have more error associated with them than scores closer to the middle.
Test information and its importance

![Graph showing test information and proficiency distribution across different countries.]

- **Country A**
- **Information**
- **Country B**
- **Country C**

The graph illustrates the test information and proficiency distribution for students across different countries.
Adaptive testing

• Adaptive algorithms optimize the delivery of test items to match the characteristics of individuals, thereby allowing the test to provide more reliable information about skills in a relatively shorter period of time.

• Depending on the design, these algorithms could tailor both the difficulty and number of items to the characteristics of a respondent.

• Compared to a linear test, adaptive tests:
  • maximize precision of measurement across a larger range of ability
  • provide an opportunity to accumulate greater accuracy in making the decision
  • require more sophisticated methodologies for analysis and larger sample sizes

• Adaptive algorithms assume (in virtually all cases) that automatically scoreable items are used. Items that cannot be automatically scored are not usable.

• Levels of adaptivity and efficiency depend on how well measurement objectives described in the framework can be adequately scored by computer.
Benefits of adaptive testing

- **Improved measurement** without increasing testing time to a diverse population of test takers
  - At the **individual** level, it brings benefits in terms of time and efficiency of measurement by ensuring students answer items around their own levels of ability.
  - At the **country** (or sets of countries) level, it better targets the average ability of test takers within countries by improving differentiation between low and high performers.
- **Better and more uniform precision** across a wider range of proficiency. This improve fairness from the psychometric perspective.
### CATs vs MSATs

<table>
<thead>
<tr>
<th><strong>Item-based computer-adaptive tests (CATs)</strong></th>
<th><strong>Multi-stage adaptive tests (MSATs)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• CAT algorithm decides on the next item after each response</td>
<td>• MSAT algorithm decides on the next cluster of items after multiple responses</td>
</tr>
<tr>
<td>• Adaptation therefore often occurs at the item level</td>
<td>• Adaptation therefore occurs at the testlet level (i.e., a mini test)</td>
</tr>
<tr>
<td>• Assumption of automatically scored items only</td>
<td>• Able to incorporate both automatically and human-scored items</td>
</tr>
<tr>
<td>• Requires a very large item pool</td>
<td>• Requires a more manageable item pool</td>
</tr>
</tbody>
</table>
Considerations for CAT or MSAT

• Adaptive testing designs are not without constraints:
  
  • **Large pilot efforts** as preliminary stable item parameters are needed for the adaptive algorithm. Thus, pilots must ensure a sufficient number of responses per item.
  
  • A team of **psychometricians** to lead the complex data analysis and simulation efforts as well as appropriate analytical software
  
  • **Computer delivery platforms** that are flexible enough to accommodate the various assessment designs as well as the sophisticated scoring and adaptive algorithms
  
  • A change in the **standard testing approach** as these designs do not allow respondents to return to items already administered
Concrete examples of adaptive testing in international assessments

• **PISA for Development Strand C:** Implemented as a tablet-based survey to assess skills of out-of-school youth. Relied on automatically scored items only.

• **PIAAC:** Implemented as a computer- and interview-based survey to assess skills of adults age 16-65. Relied on automatically scored items only.

• **PISA:** PISA 2015 switched from a paper-based assessment (PBA) to a mostly computer-based assessment (CBA) to assess 15-year-old students. Its adaptive design focus on population estimates as well as trend over time. PISA relies on a combination of automatically and human scored items

“MSAT designs in ILSA were introduced with the aim of providing more accurate and efficient measures, given the heterogeneous characteristics of the participating groups and individuals, and the number and type of items (automatically vs. human scored) needed to represent the full construct.”
(EDU/WKP(2019)17)
PISA for Development Strand C:
Represents the simplest level of adaptation, where a core test is used to assess students’ basic skills. The results determines the path the test taker will proceed to in the next stage.

Designed for 25-45 minutes of assessment time.

Stage 1 adaptivity: Selection of items appropriate to the target population

Stage 2 adaptivity: Ensuring the respondent have basic set of skills
**PISA:**
The MSAT design for PISA consists of three stages: Core, Stage 1, and Stage 2

Selection of Core was random

Designed for 60 min of assessment time
Efficiency of multistage adaptive testing

The PISA MSAT design showed about a 4.5% precision gain on average. The MSAT contributed to the accuracy of the person ability estimator across all scale scores, in particular at the extreme performance levels of lower than 300 and higher than 700, with around 10% higher accuracy.
PIAAC:
The MSAT design for PIAAC represents the most comprehensive approach, with multiple levels of adaptivity within both paper and computer paths, with the help of interviewers.

Adaptivity occurred based on computer experience and set of skills. Additionally, there are two levels of adaptivity within the cognitive assessment: Stage 1 and Stage 2. Each stage contained a number of testlets varying in difficulty.
Efficiency of multistage adaptive testing

**PIAAC:** The PIAAC MSAT design was 15 to 47 percent more efficient, meaning we can obtain the same amount of test information as we might expect from a test that is 15 to 47 percent longer.

Final thoughts

• MSAT provides more information than linear tests of equal length optimized for the ability distribution of each country. There is no proficiency range where adaptive testing was less informative than a traditional linear test.

• The use of adaptive algorithms within household surveys brings even more benefits given the:
  • heterogeneous distribution of skills and diverse background of participants
  • need for the assessment to be as short as possible (as these are interview based surveys) while being sensitive to the respondent

• These computer-based adaptive designs require early planning and a higher level of technical expertise

• As computer-based assessments are not feasible for all contexts and programs, are there better ways to tailor instruments to the characteristics of students within the paper-based world of assessments?
Thank you!

ctamassia@ets.org