Adaptive testing: an introduction

Assessment Research and Innovation

Yaw Bimpeh and David West

Key points

- A computer adaptive test (CAT) is an assessment administered on a computer that selects the difficulty level of each question depending on whether a student answered previous questions correctly.
- The benefits of computerised adaptive testing include a tailored experience for students and flexibility of assessment.
- Drawbacks of computerised adaptive testing include limited applicability to mainstream qualifications and the fact that a student's raw test scores are no longer a direct indicator of performance. CATs are not suitable for all subjects and cannot assess all skill types.
- In Scotland, children take adaptive tests in literacy and numeracy for the Scottish National Standardised Assessments, which provide teachers with diagnostic information and policymakers with national data on student progression. In Wales, statutory personalised assessments in reading and numeracy use adaptive tests and are available to take throughout the school year for learners in Years 2 to 9.
- GCSE and A-level subjects in England that may be more likely to adopt CATs in future – assuming assessments were in on-screen format – include Computer Science, Mathematics, Science, Economics, and Geography.

1. Introduction

A computer adaptive test (CAT) is an assessment administered on a computer that matches the difficulty level of each question or item to the ability level of the candidate. CATs have potential benefits over conventional fixed-form assessment, whether delivered via pen and paper or in an on-screen format.

In future, it is anticipated that more large-scale educational assessments may be delivered using CATs, potentially including GCSEs and A-levels. However, there are still complexities and challenges involved in significantly increasing the use of CATs in the English education system.

This briefing explores the following questions.

- What is adaptive assessment?
- What is the history of adaptive tests?
- How do computer adaptive assessments work?
- What types of questions can make use of adaptive assessment?
- What are the benefits of adaptive assessment?

- What are the challenges of adaptive assessment?

2. What is adaptive assessment?

Adaptive assessment 'involves the selection of test items [questions] during the process of administering a test so that the items administered to each individual are appropriate in difficulty for that individual'.[1]

There are two main approaches to adaptive assessment implemented in existing assessment systems. One is based on so-called item response theory (IRT), which uses a statistical model to estimate a numerical value for each individual's level of proficiency in a subject. The other uses what is known as knowledge space theory: a mathematical way of representing the learning objectives for a subject and then estimating a learner's 'location' on the various paths between no knowledge of the subject and mastery of the subject. The rest of this briefing focuses on adaptive assessments based on IRT.

Items are selected by a computer from a large item bank – a database of standardised, quality-assured questions. From this resource, the computer designs a bespoke assessment in real time to suit each individual student.

The computer's selection of the next question presented to a student depends on the current estimate of the student's ability, based on their answers to previous questions.

2.1 Examples of CATs in education

In Scotland and Wales, children take adaptive tests in literacy and numeracy:

- the Scottish National Standardised Assessments (SNSA) provide Scottish teachers with diagnostic information on aspects of reading, writing and numeracy
- in Wales, personalised assessments in reading and numeracy are currently statutory for learners in maintained schools from Years 2 to 9.

Outside the UK, well-known CATs in educational settings include the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT) and the Test of English as a Foreign Language (TOEFL). In these tests, individuals are presented with questions from an item bank, with each one determined by their preceding answer. Other examples of individualised CATs include:

- American Society of Clinical Pathologists-Board of Registry Certification Examinations (ASCP), USA
- A Flexible Testing System in Mathematics Education for Adults (MATHCAT), Netherlands
- Computerized Assessment System for English Communication (CASEC), Japan
- Business Language Testing Service (BULATS) computer test, offered by Cambridge English

[1] Weiss, D. J. (1983). *New Horizons in Testing*. Academic Press.

- North American Pharmacist Licensure Examination (NAPLEX), USA

3. What is the history of adaptive tests?

The idea of tailoring test difficulty to each individual was first developed by Alfred Binet, a French psychologist who played a key role in the development of IQ testing. Binet's adaptive test was paper based, requiring the administrator to estimate the examinee's ability prior to the start of testing. However, it was soon abandoned due to its complexity, and the concept lay dormant until the early 1950s.

The early 1970s saw adaptive testing become computerised adaptive testing, in which an algorithm selected questions for each individual tailored to their ability level.

IRT models were developed to separate the difficulty of a question from student ability when it comes to interpreting test performance. Importantly, IRT can help compare scores for students who have taken very different assessments, removing the need to control statistically for difficulty and to align standards between tests and cohorts.

The International Association for Computerized Adaptive Testing (IACAT) is dedicated to developing theory, techniques and technologies for adaptive testing to solve educational problems.

4. How do computer adaptive assessments work?

An item bank, containing high-quality items for different ability levels and covering a wide range of content, is essential for adaptive assessment to work.

The item bank must have been pretested to obtain statistics such as item difficulty and level of discrimination. ('Discrimination' is a technical concept that refers to how effectively a question can distinguish between two students of different ability.)

An adaptive assessment begins with random selection of a few mid-difficulty items. The student's responses to these allow an initial estimate of his or her ability. The subsequent items selected will be close to this difficulty estimate. If they are answered correctly, the next question is more difficult. If they are answered incorrectly, the next one is easier.

The computer continuously updates its estimate of the student's ability until the process is stopped. The assessment might finish because a maximum number of items have been administered, or the student's ability has been established, based on their response pattern. It could also be time restricted.

5. What types of questions can make use of adaptive assessment?

The first generation of CATs only used multiple-choice questions, in which students have to select the correct response from several options.

Advances in psychometric techniques and more powerful computers have enabled the use of other types of questions, including short narrative responses. Some modern foreign language assessments, such as listening and reading, can also use adaptive assessment.

However, setting essays or seeking longer narrative responses within adaptive tests presents designers with more of a technical challenge. More research is required into how best to extract ability information from more complex responses containing diagrams, data plots, tables or performance tasks, which usually need human intervention to provide a score.

5.1 Which GCSEs and A-levels have the most potential to partially or fully incorporate adaptive assessment?

CATs are not suitable for all subjects as IRT cannot be applied to all kinds of skills and assessment items. GCSE and A-level subjects that may be more likely to exploit computerised adaptive testing in future, assuming assessments were in on-screen format, include:

- Computer Science
- Mathematics
- Science
- Modern Foreign Languages (e.g. listening and reading test)
- Business Studies
- Economics
- Geography.

6. What are the benefits of adaptive assessment?

6.1 Benefits for students

Responses to relatively few items are needed to estimate ability from an adaptive assessment when compared with conventional fixed-form exams. As such, students can be presented with a shorter, bespoke sequence of exam questions. Individuals can also work at their own pace.

Adaptive assessment challenges a student at the top end of the ability spectrum without discouraging the weakest students[2], as students at both ends of the ability spectrum are presented with questions tailored to their level.

Nevertheless, sitting adaptive assessments can take some getting used to; for example, high-attaining students may be unaccustomed to getting questions wrong.

Due to the flexible delivery of adaptive tests, they can be taken in a calmer, quieter setting than a traditional exam hall.

6.2 Benefits for centres

Given that CATs draw on large item banks, the exam schedule can be more flexible because there is no single fixed paper, and the requirement for secrecy of question papers is relaxed.

[2] Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum

This makes it easier for students to plan for and take exams at their convenience. Responses can be processed digitally, instead of physical transportation of paper, and administering a unique set of items to each examinee can improve exam security.[3]

6.3 Benefits for assessment

As responses are entered on and marked by the computer, feedback for students can be provided during the assessment and immediately afterwards. This avoids both the delay and cost of human marking.

Digitally delivered tests can provide additional data such as the time it takes an examinee to answer an item. Computer-based assessment also removes the need for separate test equating or tiered exams.

6.4 Benefits for learning

Computer-based assessment provides educational benefits such as timely feedback to inform future teaching and learning.

For example, Welsh students in Years 2 to 9 who take the personalised assessments in reading and numeracy are provided with instant feedback after their tests, allowing their teachers to target their support.

7. What are the challenges of adaptive assessment?

A wide range of challenges have been highlighted in research on adaptive testing,[4] including the creation of large pretested item banks, question selection, sequencing and when to stop testing.

7.1 Raw test scores are no longer a direct indicator of performance

For conventional GCSE and A-level question papers, a raw mark can be found by tallying up correct responses. The total raw marks are usually converted into grades by establishing grade boundaries that are intended to be stable from year to year. However, it is difficult to conceive of a system that would allow students to estimate outcomes in a CAT setting.

Furthermore, with item selection and ability estimation taking place in real time, computer algorithms become both test designer and grade awarder with little or no human involvement. Any ambiguity about test difficulty could undermine public confidence and would require a clear explanation of how adaptive assessment works.

---

[3] Thompson, N. A. (2010). *Adaptive testing: Is it right for me?* https://assess.com/docs/Thompson_(2010)_-_Adaptive_Testing_Right.pdf

[4] Mills, C. N., & Stocking, M. S. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9,* 287–304. https://www.tandfonline.com/doi/abs/10.1207/s15324818ame0904_1

7.2 High demand for new assessment items

It is estimated that an item bank for adaptive administration of an assessment needs to contain at least six times the number of items found in a conventional paper-based exam[5].

For high-stakes exams such as GCSE and A-levels, writing six times the current number of assessment items could be a very expensive process, and new item development at this rate could be unsustainable.

7.3 Security

Pre-exposure of the content of the item bank could result in examinees memorising items and sharing them with future examinees. This represents a risk to exam secrecy. Most existing security measures to address item exposure only address overexposure.

7.4 Constraint on the assessment items that can be used

As noted above, current adaptive assessments restrict the types of items offered in order to use IRT to estimate difficulty. More often than not, this excludes the more authentic and complex item types.

7.5 Fairness among students presented with different assessments

Many researchers have raised concerns about fairness in CATs. For example, two maths questions considered to be at the same level of difficulty may actually require very different response times [6]. If the item-selection algorithm does not consider a typical response time, then some students will face more time-consuming questions than others working at a similar ability level. Although test lengths could be adjusted to take account of such differences, this may itself result in a perception of unfairness.

Important questions have also been raised as to how appropriately computerised adaptive testing measures the ability of students with disabilities[7]. This is a factor that merits much more in-depth exploration.

7.6 Student experience

Conventional question papers allow a student to review the whole paper, to skip more difficult items initially and to return to these later. Computer-based adaptive testing may require exam questions to be answered in the order in which they are presented.[8]

---

[5] Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (ETS Research Report RR-94-5). Educational Testing Service.

[6] Bridgeman, B., & Cline, F. (2000). Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assessment (Graduate Record Examination Board 96-20P; ETS RR-00-07). Educational Testing Service.

[7] Johnstone, C. J., Altman, J., Thurlow, M. L., & Thompson, S. J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004* (Technical Report 45). University of Minnesota, National Center on Educational Outcomes.

[8] Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised General Test. *Educational and Psychological Measurement*, *75*(6), 1002–1020.

8. Conclusion

Assessments that use CATs are now an established pillar of different systems around the world. However, as with any form of assessment, computerised adaptive testing has pros and cons.

While there has been recent interest in computerised adaptive testing in the context of the English education system,[9] there would inevitably be limits on its scope and applicability, particularly in relation to high-stakes qualifications such as GCSEs and A-levels.

Policymakers may wish to consider the following questions.

- How would trust be maintained in any 'black box' algorithm to control assessments?
- Will all centres have the technological infrastructure to support computer adaptive assessment?
- Will the cost of computer adaptive assessments exceed that of the current paper-based fixed exams?
- How will computer adaptive assessment data be used to maintain an academic standard and inform policymaking?
- How will teachers, head teachers, and administrators be trained to use and understand computer adaptive assessment?

[9] For example, see the Ofqual corporate plan 2022 to 2025.

Further reading

Bimpeh, Y. (2019). *Exploring the use of item response theory models for analysing high-tariff items*. AQA.

Larkin, K. C., & Weiss, D. J. (1974). *An empirical investigation of computer-administered pyramidal ability testing* (Research Report 74-3). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2009). Innovative items for computerized testing. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). Springer. https://doi.org/10.1007/978-0-387-85461-8_11

Reckase, M. (2011). Computerized adaptive assessment (CAA): The way forward. In Policy Analysis for California Education and Rennie Center for Education Research & Policy, *The road ahead for state assessments* (pp. 1–12). Rennie Center for Education Research & Policy.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review*, *36*, 263–277.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, *21*(4), 365–389.

van Lent, G. (2009). Risks and benefits of CBT versus PBT in high-stakes testing. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 83-91). Joint Research Centre of the European Commission. https://op.europa.eu/en/publication-detail/-/publication/e4915bce-5d43-4530-84be-1fcc61b7d397

Verschoor, A. J., & Straetmans, G. J. J. M. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 137–149). Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27.

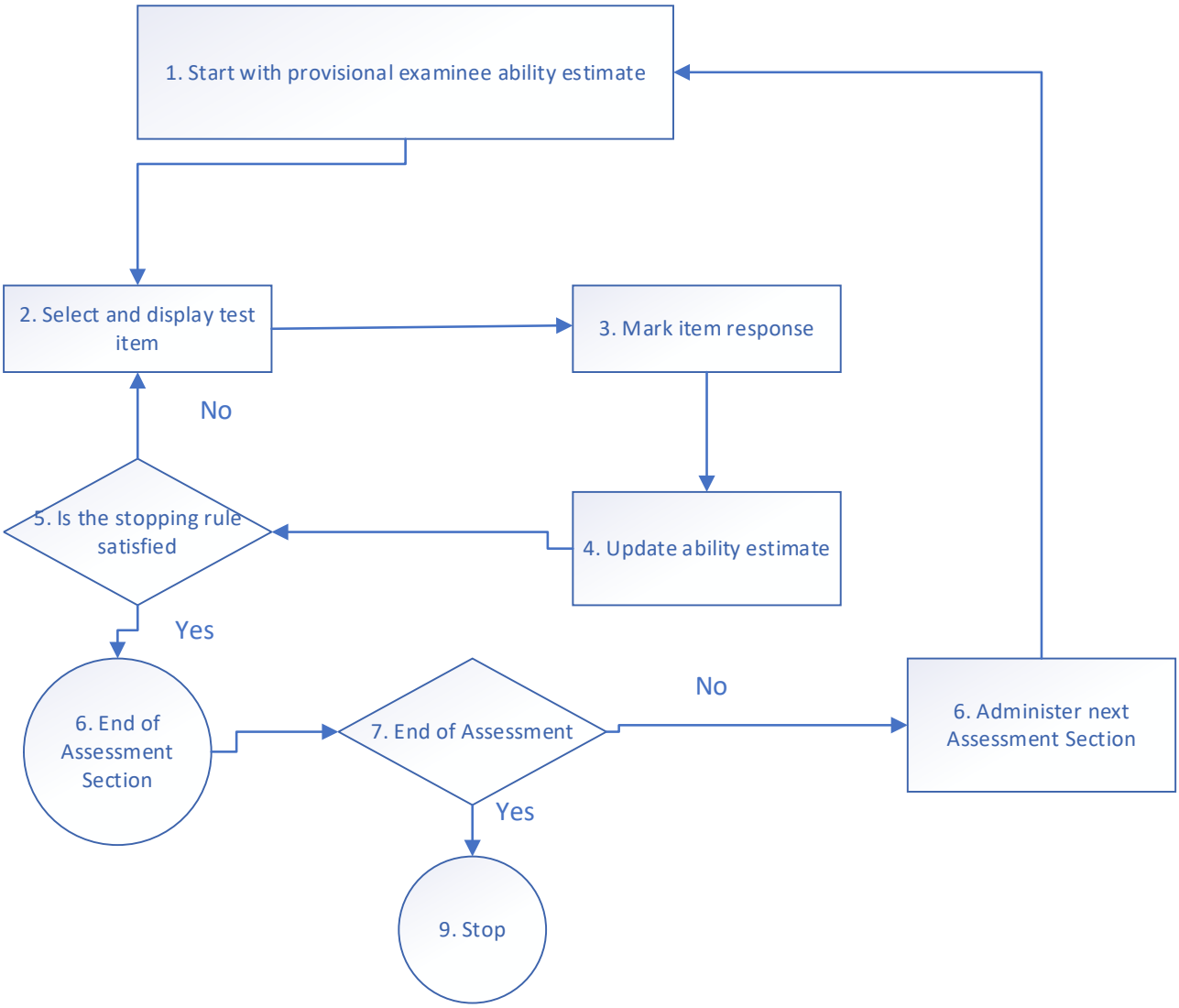

Figure 1 Illustration of adaptive assessment