








REVIEW

# A narrative review of adaptive testing and its application to medical education [version 1; peer review: 4 approved]

Steven A. Burr <sup>1</sup>, Thomas Gale<sup>1</sup>, Jolanta Kisielewska <sup>1</sup>, Paul Millin <sup>1</sup>, José M. Pêgo <sup>2</sup>, Gergo Pinter <sup>1</sup>, Iain M. Robinson<sup>3</sup>, Daniel Zahra<sup>1</sup>

<sup>1</sup>University of Plymouth, Plymouth, England, UK

<sup>2</sup>University of Minho, Braga, Portugal

<sup>3</sup>Lancaster University, Lancaster, England, UK

**V1** First published: 24 Oct 2023, 13:221  
<https://doi.org/10.12688/mep.19844.1>

Latest published: 24 Oct 2023, 13:221  
<https://doi.org/10.12688/mep.19844.1>

## Abstract





Adaptive testing has a long but largely unrecognized history. The advent of computer-based testing has created new opportunities to incorporate adaptive testing into conventional programmes of study. Relatively recently software has been developed that can automate the delivery of summative assessments that adapt by difficulty or content. Both types of adaptive testing require a large item bank that has been suitably quality assured.



Adaptive testing by difficulty enables more reliable evaluation of individual candidate performance, although at the expense of transparency in decision making, and requiring unidirectional navigation. Adaptive testing by content enables reduction in compensation and targeted individual support to enable assurance of performance in all the required outcomes, although at the expense of discovery learning.

With both types of adaptive testing, candidates are presented a different set of items to each other, and there is the potential for that to be perceived as unfair. However, when candidates of different abilities receive the same items, they may receive too many they can answer with ease, or too many that are too difficult to answer. Both situations may be considered unfair as neither provides the opportunity to demonstrate what they know. Adapting by difficulty addresses this. Similarly, when everyone is presented with the same items, but answer different items incorrectly, not providing individualized support and opportunity to demonstrate performance in all the required outcomes by revisiting content previously answered incorrectly could also be considered unfair; a point addressed when

## Open Peer Review

Approval Status 

	1	2	3	4
<b>version 1</b>				
24 Oct 2023	<a href="#">view</a>	<a href="#">view</a>	<a href="#">view</a>	<a href="#">view</a>

1. **Kathy Cullen** , Queen's University Belfast, Belfast, UK
2. **Joanne Somerville**, The University of Western Australia, Perth, Australia
3. **Steven Capey**, Swansea University, Swansea, UK
4. **S Helen Oram** , King's College London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

adapting by content.

We review the educational rationale behind the evolution of adaptive testing and consider its inherent strengths and limitations. We explore the continuous pursuit of improvement of examination methodology and how software can facilitate personalized assessment. We highlight how this can serve as a catalyst for learning and refinement of curricula; fostering engagement of learner and educator alike.

### Keywords

Assessment, adaptive testing, personalised, progress testing, fairness, different questions, reliability, compensation

**Corresponding author:** Steven A. Burr ([steven.burr@plymouth.ac.uk](mailto:steven.burr@plymouth.ac.uk))

**Author roles:** **Burr SA:** Conceptualization, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Gale T:** Writing – Review & Editing; **Kisielewska J:** Writing – Review & Editing; **Millin P:** Writing – Review & Editing; **Pêgo JM:** Conceptualization, Writing – Review & Editing; **Pinter G:** Writing – Review & Editing; **Robinson IM:** Conceptualization, Writing – Review & Editing; **Zahra D:** Conceptualization, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The author(s) declared that no grants were involved in supporting this work.

**Copyright:** © 2023 Burr SA *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Burr SA, Gale T, Kisielewska J *et al.* **A narrative review of adaptive testing and its application to medical education [version 1; peer review: 4 approved]** MedEdPublish 2023, 13:221 <https://doi.org/10.12688/mep.19844.1>

**First published:** 24 Oct 2023, 13:221 <https://doi.org/10.12688/mep.19844.1>

## Introduction

Tests that adapt by difficulty have historically been referred to as computerized adaptive testing, or CAT (Collares & Cecilio-Fernandes, 2019; Gershon, 2005; Linden *et al.*, 2000; Meijer & Nering, 1999; Wainer *et al.*, 2000). Test items representing a range of difficulties are pre-calibrated (Rice *et al.*, 2022). During delivery, the adaptive algorithm aims to select items with a 50/50 chance of candidates responding correctly. If a candidate answers an item incorrectly then the next item they receive is easier; and if they answer correctly then the next item is more difficult. The difficulty of items thus fluctuates to find the level of ability of the candidate, at which point they will be answering half of the items they receive correctly and half incorrectly. It is implicit that candidates will each receive different items, and all expect to reach a 50% probability of getting the correct answer and thereby achieve an overall score approaching 50%. Decisions about the adequacy of performance by individual candidates are then based on the level of difficulty of the items they have gravitated toward. With most items pivoting around the level of ability of the individual candidate, it follows that the measure of their achievement is more reliable (Lord, 1980). The main advantages of CAT are, therefore, increased reliability and more precise measurements (Lord, 1980); with fewer test items needed to estimate the ability of an individual candidate. This improves efficiency but also decreases test item exposure which is a relevant factor in maintaining the security and stability of the test item bank underlying this technology. Additionally, since each candidate receives an individualized test, test delivery can be asynchronous for candidates. Thus, candidates from the same cohort can be assigned to different time slots. This can be useful when cohorts are too large for available venues as physical space and computers can be reused. The main criticism that is raised relates to the potential unfairness of the test administration, since candidates are presented with different subsets of items that may bias the sampling of content and stimulus difficulties (Denny *et al.*, 2019). To mitigate this, large banks of calibrated items need to be set up, to ensure a diverse range of content and difficulties that match a predetermined blueprint. This increases the workload, time demand, and cost to develop and maintain such item banks. Another criticism of CAT by difficulty is that it not only requires large banks of pre-calibrated items, but the item response theory models on which difficulty categories are based require large numbers of candidates in order to be robust (> 200 for even the simplest models). Thus, the successful examples cited have usually been national exams with many thousands of candidates (Sands *et al.*, 1997).

Tests that adapt by content have been developed more recently and are referred to as content adaptive progress tests or CAPT (Burr *et al.*, 2022), and are especially applicable to longitudinal approaches to knowledge testing such as the progress test format that is more commonly used in the education of health professionals. When adapting by content the knowledge required is retested in subsequent tests until every candidate has successfully demonstrated knowledge in all required content areas. If a candidate answers an item on particular content incorrectly then that content is more likely to be repeated (with a different item); and if they answer correctly then that content is less likely to be repeated. This effectively functions as

a progress test (Schuwirth & Van der Vleuten, 2012) since the growth of each candidate toward demonstrating knowledge in every required content area covered by the learning objectives of a programme can be monitored and supported over time.

Here we explore the history of adaptive testing, its evolution over time, and potential implications for using different types of adaptive testing in the future.

## History of adaptive testing

The roots of adaptive testing can be traced back to the work of Binet & Simon (1905). Binet developed an IQ test that used a bank of items asked by a psychologist. The psychologist would use their estimate of the candidate's ability to determine which items to start with, followed by branching to select which set of items they would ask next depending on the answers to the previous set. There was a pre-determined ceiling level of performance that would terminate the test, and the final score would be weighted by the chronological age of the candidate. These principles have continued to be utilized in IQ testing, including in the progression and termination rules of the widely used WAIS-IV (Wechsler, 2008). During WWII sequential testing was developed, as a subcategory of assessment that specifically ended data collection once a predetermined threshold had been achieved (Taylor & Russell, 1939; Wald, 1945), and this was then used in medical research (Armitage, 1950). However, the methods were impractical for large-scale testing due to the time and resources required.

It was not until the advent of sufficient computing power in the 1960s that adaptive testing became a practical method for large-scale testing. In 1968, Frederick Lord developed a set of rules that could form the basis of a computerized adaptive testing algorithm for the Educational Testing Service to administer the Graduate Record Examination in the USA (Lord, 1968). Subsequently, in the 1970s and 1980s, researchers began to develop new algorithms and models for adaptive testing. Sequential testing was intensively studied and applied to the assessment of candidates where candidates answered items one at a time until a predetermined level of performance was achieved (summarized by Weitzman, 1982). Item Response Theory (IRT) as a distinct area of study with practical applications was developed (Lord, 1980), which provided a framework for modeling the relationship between a candidate's ability and their performance on individual test items. The IRT model allowed for more accurate and efficient adaptive testing, as it could estimate the candidate's ability level with greater precision than previous methods (Weiss, 1985).

In the 1990s, the increasing accessibility of the internet and networked computing enabled the coordinated delivery of different items to different candidates in large cohorts. Researchers began to explore the use of adaptive testing in online learning and assessment. The Graduate Record Examination in the USA was the first large-scale high-stakes assessment to use adaptive testing (Sands *et al.*, 1997). However, by the turn of the century, adaptive testing had not become as commonplace as some had speculated (Linacre, 2000). Take-up for summative use more widely across the education sector was slow,

probably due to the need for a combination of: (1) commercial software (Sahin *et al.*, 2018), (2) computer venues, (3) academics willing to innovate, and (4) enlightened regulatory policymakers. Adaptive testing has been recognized to provide personalized learning experiences, as it can adapt the difficulty of items to the individual needs of each candidate. However, for a long time, it has been perceived to be unattainably complex to deliver summatively in comparison to established non-adapting methods of assessment. Perceptions of adaptive testing are also colored by the persistent cultural belief that for a test to be ‘fair’ everyone should receive the same items. In addition, there is the difficulty of making high-stakes decisions based on candidates answering different items. The expansion in use of adaptive testing has thus predominantly been through developing formative assessments, and this has evolved with the advent of mobile devices to support flexible personalized learning (Choi & McClenen, 2020; Conejo *et al.*, 2004; Huang *et al.*, 2009; Oppl *et al.*, 2017; Triantafillou *et al.*, 2008). Other applications based on the principles of adaptive testing have been developed to help in the evaluation of patients, for example in aspects of mental health (Gibbons *et al.*, 2008) and arthritis (Fries *et al.*, 2009). Another success, currently familiar to many is the popular “Duolingo” mobile application that requires successful content completion before the user can “level up” using adaptive release (Teske, 2017). There have also been relatively recent developments to deliver national assessments of proficiency in literacy and numeracy for school children using adaptive testing in Australia (Thompson, 2017) and Wales (Williams, 2017).

To date, the emphasis on adaptive testing has been to adapt by difficulty. Recently in 2018–22 the European Board of Medical Assessors coordinated the development of an online international progress test, that adapted by difficulty, involving 8 medical schools as partners across 5 countries. This revealed that the creation and delivery of sequential formative tests was feasible and achievable (Rice *et al.*, 2022). Other medical schools have also begun to develop tests that adapt by difficulty (Koşan *et al.*, 2019). The development of a progress test that adapts by difficulty provided the foundations to enable two of the original partner medical schools (at the universities of Plymouth and Minho) to then develop the means to adapt tests by content instead of difficulty, and to begin to practically implement this summatively (Burr *et al.*, 2022). This form of summative assessment is not only personalized to ensure performance in all required content is achieved, but also motivates, supports, and empowers candidate development, and thus functions *as* learning (Bennett, 2010).

### Advantages of adaptive testing

One of the key advantages of adaptive testing is its ability to provide a more accurate assessment of a candidate’s knowledge and skills compared to other assessment methods. In a study by Wang and colleagues, adaptive testing was shown to have higher validity, in addition to reliability, compared to traditional testing methods (Wang *et al.*, 2012). The authors concluded that adaptive testing could provide a more accurate and fair assessment of a candidate’s abilities. In another study, the use

of an adaptive testing method significantly reduced the testing time while maintaining the same level of accuracy as traditional testing methods (Weiss & Kingsbury, 1984). Thus, adaptive testing can provide a more efficient testing experience, which could be improved further by considering the use of alternative statistical models to assess candidate knowledge (Boyd *et al.*, 2010). Another advantage of adaptive testing is its ability to provide personalized feedback. In a study by Martin and Lazendic (2018), candidates were assessed using an adaptive testing method that provided personalized feedback based on their performance. The results showed that candidates who received personalized feedback had higher levels of motivation and engagement compared to those who received generic feedback. It has also been suggested that adaptive testing improves the candidate’s motivation and engagement by presenting items that are challenging but not too difficult, although benefits will not accrue without curriculum alignment (Griff & Matter, 2013). Furthermore, adaptive testing can provide valuable feedback to both the student and the educator by identifying areas of strength and weakness and guiding the learning process (Martin & Lazendic, 2018). In contrast to tests that adapt by difficulty, tests that adapt by content are intrinsically aligned to the principles and advantages of programmatic assessment (van der Vleuten *et al.*, 2012), offering an alternative to portfolio-based assessment, with linkage between tests. A programmatic approach to assessment decreases the use of high stakes assessments for making progression decisions and instead uses continuous assessment methods which build up a profile of a learner’s growth in knowledge and skills.

### Potential limitations of adaptive testing

One of the main limitations is the potential for item exposure, where candidates may share or memorize the items and compromise the validity of the assessment (Persky & Fuller, 2021). To minimize this risk, CAT requires a large pool of items that can be randomly selected and rotated across candidates. Nering and Ostini (2011) have previously noted that the development of large item banks can be time-consuming and costly. However, the advent of automatic item generation (Lai *et al.*, 2009) and generative Artificial Intelligence (Falcão *et al.*, 2022; Falcão *et al.*, 2023; Sun & Hoelscher, 2023) has the potential to dramatically augment the process of creating large numbers of items. Item security could be further improved if the same item could be rewritten in different ways using generative Artificial Intelligence.

An additional requirement for tests that adapt by difficulty is the need for sound psychometric models and algorithms to facilitate decisions about the setting of performance thresholds (van der Linden & Hambleton, 2013). For example, the application of IRT and Rasch modelling to make pass-fail decisions (Kuravsky *et al.*, 2017; van der Linden & Pashley, 2009) could be considered difficult to understand and as a result, for decisions about candidates to lack transparency. Adapting by difficulty requires the selection of items ‘on the fly’ within a test, and this also means that candidates cannot be permitted to go back and change their answers to items previously viewed. Thus, navigation must be unidirectional if

adapting by difficulty within a test (whereas navigation can be bidirectional if adaptation is by content between tests). When adapting by difficulty, if a candidate receives an easier item, then this indicates they have answered the previous item incorrectly. With bidirectional navigation candidates could spend time during an examination trying to evaluate the relative difficulty of sequential items and thus whether they should change their answers. The potential for differential cueing effects on easy and hard items (Schuwirth *et al.*, 1996) would become more complex and undesirable. In contrast, when adapting by content the items can be preselected before the start of each test, and bidirectional navigation retained, because adaptation occurs between tests rather than within each test.

An additional requirement for tests that adapt by content is the need for sequential test opportunities with intervals between them to give the opportunity for weaknesses to be addressed before re-testing (Ricketts & Bligh, 2011). The potential to benefit from such feedback mandates that an indication of the content (although not necessarily the items) which could be asked is transparently known by candidates in advance; a fact that can undermine the long-established benefits of discovery learning (Bruner, 1960; Castronova, 2002). When adapting by content, all items also need to have equal value to each other, when in practice there may be no such consensus. Furthermore, there needs to be an acceptance that the different aspects of content being tested can be acquired in any order over the series of tests, and so the assessment of some content may be out of alignment with its teaching for some candidates. However, this can be considered beneficial where there are multiple teaching sites or multiple classes within the same year, or where different placement rotations are experienced, as tests which adapt by content effectively adjust.

In comparison, a test that adapts by both difficulty and content would need to prioritize one over the other, and whichever isn't prioritized would suffer a degree of compromise. Practically, a test cannot contain the same range of difficulty in items for the same range of content, for all candidates, and still adapt by both simultaneously; a trade-off is required (Luecht *et al.*, 1998). If prioritizing difficulty, then candidates will need to receive items covering different aspects of content in order to balance the overall difficulty of the test. Similarly, if prioritizing content, a test cannot cover as wide a range of difficulty and provide assurance all content is achieved to the same standard. To completely adapt by both difficulty and content, all content areas would need to be covered at each difficulty increment, or vice versa, requiring additional items, such that it becomes impractical to vary both factors within a restricted test length. The balancing of content within tests which adapt by difficulty is a well-known issue with several approaches available to control it (Leung *et al.*, 2003; Veldkamp & van der Linden, 2002). An equivalent combination of adapting by difficulty whilst simultaneously evaluating cognitive attributes has also been developed (Gierl & Zhou, 2008), although the number of domains is limited and decision making more complex.

## Comparison with other assessment methods

Adaptive testing is a method of assessment where each item presented to a candidate is based on their previous answers. Adaptive testing is important for developing personalized assessments: (1) to gain more reliable measures of individual performance, (2) to ensure acquisition of all the required learning outcomes by all candidates, and (3) to tailor individual support for learning. Tests can be designed to adapt by item difficulty, content, or both.

Except for some specific circumstances such as test-equating, other methods of assessment typically present all candidates with the same items: (1) When considering difficulty this means that high ability candidates will receive many items that they find easy, while other low ability candidates will receive many items that they find hard. A lower proportion of items are close to the level of ability of any given candidate and so the reliability of the measures of performance will be lower than if the test adapts. (2) When considering content this means that, unless the required performance threshold is 100%, there will be a degree of compensation whereby some of the content will not need to be demonstrated. Candidates can thus avoid certain content areas and still achieve a pass. When considering these points, adaptive testing should work more efficiently to practically facilitate learning by providing support in accord with the theory of the zone of proximal development (Vygotsky & Cole, 1978).

## Why it is fairer to ask candidates different questions

What is fairness? Is it that candidates are all treated the same? Is that possible, or reasonable? Is it fairer for candidates to be asked the same questions or different questions? At face value, it may appear fairer for everyone to be judged by their answers to the same set of questions. However, a moment's reflection might call to mind complexities such as those parodied by the Traxler cartoon of a monkey, elephant, goldfish, and other animals being assessed 'fairly' by all being asked to climb a tree (Traxler, 1983). This equality approach to assessment presupposes all candidates have the same preparatory experience, as well as being matched in all other variables. If candidates cover the content of the curriculum at different times through different learning experiences, but are given the same set of questions at the same time with which to achieve a particular pass mark, then the unfairness in this equality approach should be apparent. In such circumstances, it becomes fairer to ask candidates different questions to each other.

The idea of asking candidates different questions to each other might be countered by claiming that candidates receive the same learning and teaching within a programme. However, it is typical of most contemporary curricula in higher education (Clegg *et al.*, 2010; although admittedly less so in school education; McNeil, 2014), for candidates to have different amounts of time on different content. This can occur: in small group work, with different peers; by choosing different coursework assignment titles; by selecting or been limited to

different placement opportunities; and due to a greater emphasis on self-directed learning. Thus, undergraduate candidates may cover content in different orders, and to different extents (or not at all), compared to other candidates in the same cohort. This variation in experience is compounded by differences in attendance through illness and other valid extenuating circumstances, leading to alternative replacement coursework, resit examinations, or even the repeating of parts of a programme with a different cohort (Burr *et al.*, 2018). Thus, traditional assessment strategies already allow that candidates on the same programme will be graduated (or not) based on a different set of assessments and assessment items to each other. It is therefore not currently the case, nor a reasonable expectation of new assessment designs, that candidates all be judged only by their response to the same questions during a programme of study. With this in mind, programmes are often designed with the end-point being for candidates to demonstrate their achievement in covering a set of learning objectives.

The traditional model of assessment presents all candidates with the same set of items covering all learning objectives, and requires them to achieve a particular pass mark with those items (Holzinger *et al.*, 2020). This has as its graduation criteria the achievement of a particular pass mark across assessments, rather than the demonstration of all the learning objectives. Thus, candidates can compensate for missing some learning objectives by excelling in others (Brinkman *et al.*, 2017), unless all assessments have a 100% pass mark (Herold *et al.*, 2016). This can even be to the point of completely ignoring one or more objectives or content areas. Furthermore, achievement of the pass mark is based on the same items set for all candidates. Within a curriculum where learning experiences differ, this means that not all candidates are given the same opportunities to develop and demonstrate their knowledge and meet the pass mark. Unmet objectives are never revisited if the pass mark has been achieved, and as such, the end-point for candidates is different (Bierer *et al.*, 2008). Some candidates will graduate with a broad range of knowledge, whereas others will graduate with exceptional knowledge of some areas and less (or no) knowledge of others (Cummings *et al.*, 2019).

In contrast, requiring candidates to demonstrate achievement of knowledge in each content area by the end of the programme is unquestionably more aligned to the stated aim of creating graduates that have demonstrated achievement of *all* of the learning objectives. Unlike the previous situation, in this case, all students have the same end-point of demonstrating achievement of each and every learning objective, and are given repeated opportunities to demonstrate this through assessment items that adapt by content. The selection of items is personalised to each candidate, with tailored feedback and support to meet objectives that remain unmet after each assessment. This adaptation, to ensure the same end-point is met by each graduate, necessitates each candidate receiving different assessment items over the course of the programme. Candidates have been

shown to perceive alternative assessment types as fairer where there is the opportunity to react to feedback and stimulate deeper learning (Struyven *et al.*, 2005). In addition, the approach fosters self-directed learning, a form of individualised learning which is realisable with asynchronous delivery (Svedberg, 2010). Individualised learning can then also be assessed, with the common end-point of meeting all objectives, through adapting items (Burr *et al.*, 2022).

Considerations of fairness also tend to focus on the potential for cheating, which can be ameliorated by presenting different questions to candidates in the same cohort (Denny *et al.*, 2019). It has long been established that fairness also depends on candidate perceptions that assessments are authentic, reasonable, realistic, developmental, and beneficial (Sambell *et al.*, 1997). It is clear that fairness is a multidimensional construct, perceived through many different lenses. Contrary to the recurrent, yet impractical, position that fairness depends on everyone facing the same questions, it can in fact be fairer, as explained here, to ask candidates different questions. This is particularly true where repeated personalised assessments are possible, giving a fairer opportunity for all candidates to demonstrate achievement of all the learning objectives, and to ensure the same standard is achieved.

## Conclusions

Whilst the ideas underpinning adaptive testing have been around since the early 1900's, examples of its use are scarce until the later part of that century, when the widespread availability and use of computers made it possible to implement adaptive assessment methods on a large scale. Today, the recent development of software to comprehensively deliver tests that adapt by difficulty and content facilitates easier adoption of more personalized forms of summative assessment, and with it, the advantages of increased assessment reliability, and the assurance that each candidate is meeting all the required learning outcomes, with tailored individual support for learning and feedback where it is needed. These technical developments to facilitate the delivery of adaptive testing do not necessarily overcome other barriers such as the misguided perception that it is unfair for candidates to be required to answer different questions. However, in cases where assessment across a wide range of content is paramount, the use of different questions could become accepted if there are significant benefits. This could certainly be the case for adapting testing by content since this can eliminate compensation and thereby categorically demonstrate attainment in all of the knowledge required.

---

## Data availability

There are no data associated with this article.

## Acknowledgements

None.

## References

- Armitage P: **Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis.** *J R Stat Soc Series B Stat Methodol.* 1950; **12**(1): 137–144.  
[Publisher Full Text](#)
- Bennett RE: **Cognitively Based Assessment of for, and as Learning (CBAL): A Preliminary Theory of Action for Summative, and Formative Assessment.** *Measurement.* 2010; **8**(2–3): 70–91.  
[Publisher Full Text](#)
- Bierer SB, Dannefer EF, Taylor C, et al.: **Methods to assess students' acquisition, application and integration of basic science knowledge in an innovative competency-based curriculum.** *Med Teach.* 2008; **30**(7): e171–e177.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Binet A, Simon Th: **Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux.** *Annee Psychol.* 1905; **11**: 191–244.  
[Publisher Full Text](#)
- Boyd A, Dodd B, Choi S: **Polytomous Models in Computerized Adaptive Testing.** In: *Handbook of Polytomous Item Response Theory Models*, 1st ed. Routledge, 2010; 28.  
[Reference Source](#)
- Brinkman DJ, Tichelaar J, Okorie M, et al.: **Pharmacology and Therapeutics Education in the European Union Needs Harmonization and Modernization: A Cross-sectional Survey Among 185 Medical Schools in 27 Countries.** *Clin Pharmacol Ther.* 2017; **102**(5): 815–822.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bruner JS: **The Process of Education.** 1960.  
[Reference Source](#)
- Burr SA, Kisiielewska J, Zahra D, et al.: **Personalising knowledge assessments to remove compensation and thereby improve preparation for safe practice - developing content adaptive progress testing.** 2022.  
[Publisher Full Text](#)
- Burr S, Morrison JM, Salih VM: **When another assessment attempt is bad for progress [version 1].** *MedEdPublish.* 2018; **7**(3): 1–11.  
[Publisher Full Text](#)
- Castronova JA: **Discovery learning for the 21st century: What is it and how does it compare to traditional learning in effectiveness in the 21st century.** *Action research exchange.* 2002; **1**(1): 1–12.  
[Reference Source](#)
- Choi Y, McClenen C: **Development of adaptive formative assessment system using computerized adaptive testing and dynamic Bayesian networks.** *Appl Sci.* 2020; **10**(22): 8196.  
[Publisher Full Text](#)
- Clegg S, Stevenson J, Willott J: **Staff conceptions of curricular and extracurricular activities in higher education.** *High Educ.* 2010; **59**: 615–626.  
[Publisher Full Text](#)
- Collares CF, Cecilio-Fernandes D: **When I say... computerised adaptive testing.** *Med Educ.* 2019; **53**(2): 115–116.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Conejo R, Guzmán E, Millán E, et al.: **SIETTE: A web-based tool for adaptive testing.** *Int J Artif Intell Educ.* 2004; **14**(1): 29–61.  
[Reference Source](#)
- Cummings DL, Smith M, Merrigan B, et al.: **MSK30: a validated tool to assess clinical musculoskeletal knowledge.** *BMJ Open Sport Exerc Med.* 2019; **5**(1): e000495.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Denny P, Manoharan S, Speidel U, et al.: **On the fairness of multiple-variant multiple-choice examinations.** In: *Proceedings of the 50th ACM technical symposium on computer science education.* 2019; 462–468.  
[Publisher Full Text](#)
- Falcão F, Costa P, Pêgo JM: **Feasibility assurance: a review of automatic item generation in medical assessment.** *Adv Health Sci Educ Theory Pract.* 2022; **27**(2): 405–425.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Falcão FMV, Pereira DS, Pêgo JM, et al.: **Progress is impossible without change: implementing automatic item generation in medical knowledge progress testing.** *Educ Inf Technol.* 2023.  
[Publisher Full Text](#)
- Fries JF, Cella D, Rose M, et al.: **Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing.** *J Rheumatol.* 2009; **36**(9): 2061–2066.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gershon RC: **Computer adaptive testing.** *J Appl Meas.* 2005; **6**(1): 109–127.  
[PubMed Abstract](#)
- Gibbons RD, Weiss DJ, Kupfer DJ, et al.: **Using computerized adaptive testing to reduce the burden of mental health assessment.** *Psychiatr Serv.* 2008; **59**(4): 361–368.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gierl MJ, Zhou J: **Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment.** *Z Psychol.* 2008; **216**(1): 29–39.  
[Publisher Full Text](#)
- Griff ER, Matter SF: **Evaluation of an adaptive online learning system.** *Br J Educ Technol.* 2013; **44**(1): 170–176.  
[Publisher Full Text](#)
- Herold R, Schiekirka S, Brown J, et al.: **Structured Smoking Cessation Training for Medical Students: A Prospective Study.** *Nicotine Tob Res.* 2016; **18**(12): 2209–2215.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Holzinger A, Lettner S, Steiner-Hofbauer V, et al.: **How to assess? Perceptions and preferences of undergraduate medical students concerning traditional assessment methods.** *BMC Med Educ.* 2020; **20**(1): 312.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Huang YM, Lin YT, Cheng SC: **An adaptive testing system for supporting versatile educational assessment.** *Computers & Education.* 2009; **52**(1): 53–67.  
[Publisher Full Text](#)
- Koşan AMA, Nizamettin KOC, Elhan AH, et al.: **Developing an item bank for progress tests and application of computerized adaptive testing by simulation in medical education.** *International Journal of Assessment Tools in Education.* 2019; **6**(4): 656–669.  
[Publisher Full Text](#)
- Kuravsky LS, Artemenkov SL, Yuryev GA, et al.: **A new approach to computerized adaptive testing.** *Experimental Psychology (Russia).* 2017; **10**(3): 33–45.  
[Reference Source](#)
- Lai H, Alves C, Gierl MJ: **Using automatic item generation to address item demands for CAT.** In: *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Minneapolis, MN: IACAT, 2009.  
[Reference Source](#)
- Leung CK, Chang HH, Hau KT: **Computerized adaptive testing: A comparison of three content balancing methods.** *The Journal of Technology, Learning and Assessment.* 2003; **2**(5): 1–16.  
[Reference Source](#)
- Linacre JM: **Computer-adaptive testing: A methodology whose time has come.** MESA memorandum, 2000; (69): 58.  
[Reference Source](#)
- Linden WJ, van der Linden WJ, Glas CA, (Eds.): **Computerized adaptive testing: Theory and practice.** Springer Science & Business Media, 2000.  
[Reference Source](#)
- Lord FM: **Some test theory for tailored testing.** Research Bulletin RB-68-38. Princeton NJ: Educational Testing Service, 1968; (2): i–62.  
[Publisher Full Text](#)
- Lord FM: **Applications of item response theory to practical testing problems.** Mahwah, NJ: Erlbaum, 1980.  
[Reference Source](#)
- Luecht RM, De Champlain A, Nungester RJ: **Maintaining content validity in computerized adaptive testing.** *Adv Health Sci Educ Theory Pract.* 1998; **3**(1): 29–41.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Martin AJ, Lazendic G: **Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience.** *J Educ Psychol.* 2018; **110**(1): 27–45.  
[Reference Source](#)
- Meijer RR, Nering ML: **Computerized adaptive testing: Overview and introduction.** *Appl Psychol Meas.* 1999; **23**(3): 187–194.  
[Publisher Full Text](#)
- McNeil JD: **Contemporary curriculum: In thought and action.** John Wiley & Sons, 2014.  
[Reference Source](#)
- Nering ML, Ostini R, (Eds.): **Handbook of polytomous item response theory models.** Taylor & Francis, 2011.
- Oppl S, Reisinger F, Eckmaier A, et al.: **A flexible online platform for computerized adaptive testing.** *Int J Educ Technol High Educ.* 2017; **14**(1): 1–21.  
[Publisher Full Text](#)
- Persky AM, Fuller KA: **Students' collective memory to recall an examination.** *Am J Pharm Educ.* 2021; **86**(3): 8638.  
[Reference Source](#)
- Rice N, Pêgo JM, Collares CF, et al.: **The development and implementation of a computer adaptive progress test across European countries.** *Computers and Education: Artificial Intelligence.* 2022; **3**: 100083.  
[Publisher Full Text](#)
- Ricketts C, Bligh J: **Developing a "Frequent look and rapid remediation" assessment system for a new medical school.** *Acad Med.* 2011; **86**(1): 67–71.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sahin A, Hurtado Grooscors HAH, Góngora-Cortés JJ: **Review of FastTest:**

**A platform for adaptive testing.** *Measurement ( Mahwah N.J).* 2018; **16**(4): 256–263.

[Publisher Full Text](#)

Sambell K, McDowell L, Brown S: **'But is it fair?': an exploratory study of student perceptions of the consequential validity of assessment.** *Studies in Educational Evaluation.* 1997; **23**(4): 349–371.

[Publisher Full Text](#)

Sands WA, Waters BK, McBride JR: **Computerized adaptive testing: From inquiry to operation.** American Psychological Association, 1997.

[Reference Source](#)

Schuwirth LWT, Van der Vleuten CPM, Donkers HHLM: **A closer look at cueing effects in multiple-choice questions.** *Med Educ.* 1996; **30**(1): 44–49.

[PubMed Abstract](#) | [Publisher Full Text](#)

Schuwirth LWT, van der Vleuten CPM: **The use of progress testing.** *Perspect Med Educ.* 2012; **1**(1): 24–30.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Struyven K, Dochy F, Janssens S: **Students' perceptions about evaluation and assessment in higher education: A review.** *Assess Eval High Educ.* 2005; **30**(4): 325–341.

[Publisher Full Text](#)

Sun GH, Hoelscher SH: **The ChatGPT storm and what faculty can do.** *Nurse Educ.* 2023; **48**(3): 119–124.

[PubMed Abstract](#) | [Publisher Full Text](#)

Svedberg MK: **Self-directed learning and persistence in online asynchronous undergraduate programs.** Doctoral dissertation, Virginia Tech, 2010.

Taylor HC, Russell JT: **The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables.** *J Appl Psychol.* 1939; **23**(5): 565–578.

[Publisher Full Text](#)

Teske K: **Duolingo.** *CALICO Journal.* 2017; **34**(3): 393–401.

[Reference Source](#)

Thompson G: **Computer adaptive testing, big data and algorithmic approaches to education.** *Br J Sociol Educ.* 2017; **38**(6): 827–840.

[Publisher Full Text](#)

Traxler H: **Chancengleichheit.** In: Michael Klant (Ed.), *Schul-Spott: Karikaturen aus 2500 Jahren Pädagogik.* Fackelträger, Hannover, 1983; 25.

Triantafyllou E, Georgiadou E, Economides AA: **The design and evaluation of a computerized adaptive test on mobile devices.** *Comput Educ.* 2008; **50**(4): 1319–1330.

[Publisher Full Text](#)

van der Linden WJ, Hambleton RK, (Eds.): **Handbook of modern item response theory.** Springer Science & Business Media, 2013.

[Reference Source](#)

Van der Linden WJ, Pashley PJ: **Item selection and ability estimation in adaptive testing.** In: *Elements of Adaptive Testing.* New York, NY: Springer New York, 2009; 3–30.

[Publisher Full Text](#)

van der Vleuten CP, Schuwirth LWT, Driessen EW, et al.: **A model for programmatic assessment fit for purpose.** *Med Teach.* 2012; **34**(3): 205–214.

[PubMed Abstract](#) | [Publisher Full Text](#)

Veldkamp BP, van der Linden WJ: **Multidimensional adaptive testing with constraints on test content.** *Psychometrika.* 2002; **67**: 575–588.

[Publisher Full Text](#)

Vygotsky LS, Cole M: **Mind in society: Development of higher psychological processes.** Harvard University Press, 1978.

[Publisher Full Text](#)

Wainer H, Dorans NJ, Flaugher R, et al.: **Computerized adaptive testing: A primer.** Routledge, 2000.

Wald A: **Sequential tests of statistical hypotheses.** *Ann Math Stat.* 1945; **16**(2): 117–186.

[Publisher Full Text](#)

Wang S, McCall M, Jiao H, et al.: **Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis.** *J Educ Develop Psychol.* 2012; **3**(1): 88–100.

[Publisher Full Text](#)

Wechsler D: **Wechsler Adult Intelligence Scale—Fourth Edition Administration and Scoring Manual.** San Antonio, TX: Pearson, 2008.

[Reference Source](#)

Weiss DJ, Kingsbury GG: **Application of computerized adaptive testing to educational problems.** *J Educ Meas.* 1984; **21**(4): 361–375.

[Publisher Full Text](#)

Weiss DJ: **Adaptive testing by computer.** *J Consult Clin Psychol.* 1985; **53**(6): 774–89.

[PubMed Abstract](#) | [Publisher Full Text](#)

Weitzman RA: **Sequential testing for selection.** *Appl Psych Meas.* 1982; **6**(3): 337–351.

[Publisher Full Text](#)

Williams K: **Case study: Welsh government national assessments for school children.** 2017.

[Reference Source](#)



# Open Peer Review

Current Peer Review Status:    

---

## Version 1

Reviewer Report 22 December 2023

<https://doi.org/10.21956/mep.21260.r35254>

© 2023 Oram S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**S Helen Oram** 

King's College London, London, UK

The narrative review by Burr et al. delves into the intriguing realm of adaptive testing and its application in medical education. The authors meticulously trace the historical evolution of adaptive testing, shedding light on its lesser-known but significant history. The advent of computer-based testing, as discussed, has not only facilitated the integration of adaptive testing into conventional study programs but has also spurred the development of software capable of automating summative assessments that adapt by difficulty or content.

The distinction between adaptive testing by difficulty and content is eloquently presented. Adaptive testing by difficulty, often referred to as computerized adaptive testing (CAT), involves selecting items based on a candidate's performance. If an item is answered correctly, the subsequent one is more challenging, and vice versa. While this method enhances reliability and precision, concerns about fairness arise as candidates are presented with different subsets of items. On the other hand, adaptive testing by content, exemplified by Content Adaptive Progress Tests (CAPT), focuses on retesting knowledge until every candidate demonstrates proficiency in all required content areas. This approach ensures assurance of performance in all outcomes but may sacrifice some aspects of discovery learning.

The review thoroughly explores the educational rationale behind the evolution of adaptive testing, considering its strengths and limitations. The continuous quest for improving examination methodology, coupled with the role of software in facilitating personalized assessment, is eloquently discussed. The authors emphasize how adaptive testing can act as a catalyst for refining curricula, fostering engagement among learners and educators alike.

A significant portion of the review is dedicated to the historical progression of adaptive testing, tracing its roots back to Binet and Simon's work in 1905. The subsequent development of computerized adaptive testing in the 1960s, driven by increased computing power, laid the foundation for the more recent advancements discussed in the paper. The slow adoption of adaptive testing for summative purposes in education is attributed to a combination of factors, including the need for commercial software, suitable venues, innovative academics, and supportive regulatory policies.

The advantages of adaptive testing, both by difficulty and content, are thoughtfully presented. CAT, with its ability to provide individualized tests asynchronously, offers increased reliability and

efficiency, albeit with the challenge of potential unfairness due to varying item subsets. Content adaptive progress tests, on the other hand, provide targeted support and motivation for candidates to demonstrate knowledge in all required content areas, addressing some of the limitations of adaptive testing by difficulty.

The authors astutely discuss potential limitations of adaptive testing, including item exposure concerns and the need for large item banks. They recognize the role of technology, such as generative Artificial Intelligence, in addressing these challenges, showcasing a forward-looking perspective. The manuscript also delves into the psychometric models required for tests adapting by difficulty and the necessity for sequential test opportunities with intervals for tests adapting by content.

A distinctive aspect of the review is its exploration of the fairness of asking candidates different questions. The authors challenge the conventional belief that fairness equates to everyone facing the same questions, suggesting that candidates' varied learning experiences make this approach impractical. They argue that requiring candidates to demonstrate achievement in each content area, even if through different questions, aligns more closely with the aim of creating graduates who have met all learning objectives.

In conclusion, the manuscript offers a comprehensive exploration of adaptive testing, weaving together historical perspectives, methodological considerations, and a thought-provoking discussion on fairness. The authors provide a well-reasoned argument for the adoption of adaptive testing, backed by examples from medical education. The inclusion of practical examples, figures, or tables could enhance visual representation and further engage readers.

Overall, this review makes a significant contribution to the discourse on adaptive testing in medical education and invites further exploration of its nuanced facets.

**Is the topic of the review discussed comprehensively in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Is the review written in accessible language?**

Yes

**Are the conclusions drawn appropriate in the context of the current research literature?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Haematology; Medical Education; Assessments; Generative AI; Computer adaptive testing.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 24 November 2023

<https://doi.org/10.21956/mep.21260.r35252>

© 2023 **Capey S.** This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### **Steven Capey**

Swansea University, Swansea, Wales, UK

This review provides a detailed overview of the development and potential uses of adaptive testing. The authors consider two contrasting approaches to the delivery of computerised adaptive testing (CAT). The narrative is clear and covers the beginning of adaptive testing through to the advent of CAT in education and medical education. The advantages and disadvantages of CAT are clearly discussed and appropriate solutions to many of these challenges are offered by the authors. This review goes further than describing the methodology of CAT by considering some of the philosophical issues associated with testing, fairness, validity and reliability. The benefit of these approaches to testing in medical education and programmatic approaches to assessment. Clear conclusions are drawn from the literature that will help to shape opinions on this rapidly developing assessment methodology.

**Is the topic of the review discussed comprehensively in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Is the review written in accessible language?**

Yes

**Are the conclusions drawn appropriate in the context of the current research literature?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Assessment in Medical Education

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 23 November 2023

<https://doi.org/10.21956/mep.21260.r35249>

© 2023 **Somerville J.** This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Joanne Somerville**

The University of Western Australia, Perth, Western Australia, Australia

The narrative review describes the evolution of, and the educational application of the concept of adaptive testing in medical education, to tailor assessments to the individual learner's ability. This enhances the learner's experience and provides targeted feedback. The review discusses the strengths and weaknesses of the concept of adaptive testing and compares adaptive testing by difficulty and content. The introduction to the review is clear and comprehensive. The goals of the review and rationale for the review are clear and supported. The strengths and limitations of adaptive testing are comprehensively covered, analysed and supported with appropriate literature. Seminal studies are included to support the historical aspects of the review and literature from international experts in the field of programmatic assessment, progress and adaptive testing is appropriately cited, which verifies the discussion. The review is logically structured and clearly flows from the introduction to the discussion, application and conclusion. The authors take a non-biased view to the discussion, clearly presenting the strengths and limitations of adaptive testing and its iterations over time and to current developments.

The authors provide a balanced and unbiased review of the different types of adaptive testing drawing appropriate conclusions supported by relevant and current literature citations. The review utilises sources to support the author's discussion but fails to critically evaluate them, however, many of the citations are seminal works or undisputed peer-reviewed research in the field. Therefore a critical analysis is not necessarily necessary. The review synthesises the supporting literature to provide novel insights into the topic. An example of this, is when the review presents a discussion around the concept of fairness.

The main author has published in peer-reviewed journals on the topic of programmatic assessment and adaptive testing and therefore has demonstrated expertise in this area that gives credence to the depth and quality of the review. Further, the author remains objective in their synthesis of the appropriate use, adaptation and applicability of adaptive testing to current clinical assessment methods. The style and clarity of the review remain consistent in both language and style. References and citations are complete and all relevant works are cited. The conclusion accurately and appropriately summarises the key findings of the review and conflicts of interest are acknowledged.

The review proposes a new and innovative and current use of adaptive testing to enhance student learning, promote a continuous longitudinal assessment method, which is reliable, valid and individualised to the learner's ability. Therefore the review significantly adds value to the existing literature.

In my opinion, the review is suitable for indexing.

**Is the topic of the review discussed comprehensively in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Is the review written in accessible language?**

Yes

**Are the conclusions drawn appropriate in the context of the current research literature?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Medical education, assessment, professionalism, simulation.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 23 November 2023

<https://doi.org/10.21956/mep.21260.r35247>

© 2023 Cullen K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Kathy Cullen** 

School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, Northern Ireland, UK

This is a clearly written narrative review of adaptive testing. This is an area which is underrecognized and this review will be of interest to both undergraduate and postgraduate medical educators. It provides a succinct summary of the background and theory around adaptive testing and possible future applications. It outlines both advantages and disadvantages of this style of assessment.

**Is the topic of the review discussed comprehensively in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Is the review written in accessible language?**

Yes

**Are the conclusions drawn appropriate in the context of the current research literature?**

Yes

**Competing Interests:** One of the authors (JK) is External examiner for the Year one assessment at Queen's University where I work. I am overall assessment lead so the link is not direct.

**Reviewer Expertise:** Medical education - assessment

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---