CREATING A COMPUTER ADAPTIVE ASSESSMENT FOR USE IN SCHOOLS

C. Jellis

CEM, Cambridge University Press and Assessment (UNITED KINGDOM)

Abstract

Increasingly, assessments in schools and colleges are using computer adaptive models. Computer adaptive tests (CATs) have many benefits when compared with traditional paper and pencil assessments. CATs use a student's response to a question to decide on the next question to offer. If the student gets a question wrong, they may be offered a slightly easier question next time, if they get it right, they may be offered a slightly harder question. This means that assessments can be created that provide questions at an appropriate level of ability for individual students, enabling more questions to be offered within a specific difficulty range. Using CAT, less able students will not be offered questions that are too complex for them, and higher-ability students will not need to spend time answering questions that are too easy before they reach the questions that really stretch them.

Creating robust computer adaptive assessments for use on a large scale, which enable comparisons of performance across schools and populations, is a complex process. It requires extensive trialling of question items. Moreover, question banks need to be very large to allow for an appropriate balance of individual questions across the ability range. Similarly, rigorous statistical methods are used to ensure that the constructs within the assessment are measured appropriately and are not prone to gender, cultural or any other kind of bias. Assessments of this kind tend to be used across a wide range of school types and geographical areas and must be psychometrically robust enough to measure accurately under these circumstances.

In addition to using large-scale CATs, there may be certain circumstances when an individual school or college may wish to create a small-scale CAT for use with its own students. In the past, creating such assessments has been complex and expensive, involving commercial solutions and the requirement for a mix of complex technical and statistical skills. However, platforms are now available which can ease the burden of assessment creation, moving the focus to the creation of specific assessments that are not constrained by technical requirements. Similarly, access to the statistics required to calculate the relative difficulty of test items has also been made simpler through the greater availability of free statistics analysis software. Whilst the process of assessment creation has become more straightforward in these respects, CAT designers must still address challenging questions around the curriculum or other constructs to be assessed, the ability range to be covered, and the extent to which questions that have already been written could be repurposed.

In this paper I detail the processes involved in creating a suitable question bank and designing a simple computer adaptive assessment using a freely available platform and statistical analysis environment, in the context of schools, colleges, or other educational institutions. I also explore some advantages, risks, and considerations around this approach for aspiring CAT designers to think through.

Keywords: Computer adaptive assessment.

1 INTRODUCTION

There are many reasons to choose a computer adaptive test over a flat test, the main reason being to align the items (the word 'item' is often used as a synonym for 'question') asked with the ability range of the test taker. This can have beneficial effects such as making a test shorter, more pleasant to sit and focused more precisely on the test taker's abilities. However, there are also drawbacks to this approach. Unlike a flat test, the relative difficulty of each item needs to be measured and recorded, which can be achieved by setting students of similar ability all the questions as a flat test, although it could also be done by adding additional trial items to an existing CAT. Another drawback is the relative number of questions required. A CAT presents questions around a certain difficulty range and only stops when enough questions have been asked to establish a measure of test taker ability. Typically, this is around 10 to 15 questions, so even for a simple test, 100 or more questions may be required to represent the whole ability range.

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)

The trialling of question items for use on a large scale can be both expensive and time consuming but is vitally important to establish reliable difficulty estimates for individual items. Trial groups need to be carefully matched with the final test taking group by key features such gender, ability, culture, and age. Use of a suitable trial group and analysis using appropriate statistical models ensures that the measures obtained of each construct are free of gender, cultural or any other kind of bias. For example, InCAS, a CAT for primary school children developed by the Cambridge Centre for Evaluation and Monitoring (CEM), was initially standardised on approximately 4,300 pupils in 71 primary schools. It is also important to decide how the information provided by the assessment will be used as that may affect the items selected. For instance, a CAT might be created to predict future performance The Computer Adaptive Baseline Test (CABT) which is the basis of a number of CEM assessments, is designed to provide predictions of the results of GCSE and IGCSE examinations taken up to four years later. Predictive CATs tend to be more complex as the items selected need to be not only effective in terms of providing valuable information about what a student knows and can do, but also as good predictors of future performance. For this reason, small scale CATs are not appropriate for high stakes testing, predictions, or indeed for year-on-year comparisons. Despite this, there are simpler approaches that may be adopted in small scale CATs for use in schools.

2 HOW DOES A CAT WORK?

When a student sits a computer adaptive test several things need to have been considered beforehand. A CAT requires a well-defined item bank, a method for selecting the first item, a method for selecting subsequent items, a scoring scheme, and a method for deciding when enough questions have been asked to decide upon a result.

The result is a system that presents a student with an appropriate question and, based on the response to that question, selects another question. Concurrently the system also updates a score for that student and decides, based on the number of questions asked and the current score, whether enough questions have been asked to reach a stable ability measure for the student. Finally, when enough questions have been asked to establish a measure of student ability the assessment stops and reports the student's score.

3 A SHORT HISTORY OF CAT

Although development of personal CAT tests has been accelerated by the introduction of the personal computer, testing using a subset of a larger item bank of questions has a much longer history. Indeed, as early as the turn of the last century, Lewis Termon of Stanford University derived a version of a test created by the French psychologist Alfred Binet and his student, Theodore Simon. The Stanford-Binet intelligence test, now in its fifth edition [1] was one of the first that used different ranges of questions for different subgroups of test takers.

In 1970, Frederic Lord [2,3] published his Flexilevel tests that asked questions based on previous answers. Lord acknowledged that although at that time such tests could be administered by computer, the cost and availability of suitable hardware was prohibitive. Instead, he suggested using a test booklet based on a double page spread that had easier questions on the left-hand side and harder questions on the right-hand side. The test taker was asked to answer the question and mark their answer in an answer book. The answer book was designed to inform the test taker whether their answer was correct or incorrect. If correct, the test taker would then answer the next question on the right-hand side of the page. If incorrect, they would answer the next question on the left-hand side of the page. By this process the test taker would end up being presented with questions appropriate to their ability and the test would end when a given number of questions had been answered.

As computer access became more freely available, implementations of CAT systems were widely developed. In 1974, Carl Jensema [4] published a CAT 'Bayesian Tailored Testing' which used Bayesian methods to calculate the standard error of the ability estimate and utilised that to indicate when a stable ability measure had been reached. In the same year, Mark Reckase [5] published a FORTRAN CAT program based on the one parameter logistic Rasch [6] model. In 1986, DeAyala and Koch [7] published a computerised version of Lord's Flexilevel tests running on an IBM PC and compared it directly to the Bayesian Tailored test. DeAyala and Koch ran simulations of both tests and found that they both worked equally well resulting in the same rank ordering of students whichever method was used.

In the next decade, as computer systems became more affordable, more widely available, and easier to use, CAT systems have become more ubiquitous and have moved from ability testing to use in areas

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)
This version should not be distributed since it may change prior to final publication

such as medicine and psychology. As more people have developed CATs, preferred methods have been developed. In 1994 Spray and Reckase [8] suggested improved item selection algorithms whereas the establishment of item difficulties has tended to centre around the use of item response theory (IRT) and the Rasch model. In 2000, John Linacre [9] discussed CATs in detail and the use of IRT and Rasch models in their development. Linacre's paper included the source code (in BASIC) to a simple CAT he wrote in 1986.

Nowadays, CATs are widely used although they still have their opponents. Arguments against the use of CATs include the inability to go back and revisit past questions, creation of item bank statistics mean that the items will have to have been seen before, and exposure overflow, where similar students see exactly the same test items. Notwithstanding these concerns, the advantages that CATs bring has resulted in their being sufficiently well regarded for use in high stakes testing such as the American Graduate Management Admission Test (GMAT), the Test of English as a Foreign Language (TOEFL) and the Programme for International Student Assessment (PISA). CATs are also routinely used in standardised baseline testing, such as the Alis, Yellis, MIDYIS and InCAS assessments offered by Cambridge CEM and in Cambridge English's Liguaskill.

4 THE ITEM BANK

An item bank is a series of questions covering a range of specific constructs, each in difficulty order. For large scale CATs, the relative difficulty of the question is usually defined by a suitable Item Response Theory (IRT) measure such as that proposed by Rasch [6]. For each construct, the range of question difficulties needs to be wide enough to encompass the ability range of the test taker population, and the questions themselves should provide a stepwise rise in difficulty without large jumps. In addition, each question needs to stand alone, without reference to any other question in the bank, such that the system can select and present any question of suitable difficulty from the item bank.

In a school developed CAT these requirements are not so rigorous. Item banks can be created from questions from tests that have been set in the past, or newly created for the purpose. If the CAT is to be used for an individual year group only, then the ability range of the item bank does not need to be so large. For a simple CAT it is best to use questions that are marked either right or wrong and present them as multiple choice with two or three distractors (alternative incorrect responses). Although CATs can be created using a difficulty measure only, some systems can accommodate other sophistications which may be used to improve item selection, such as parallel item banks or multi-part questions.

There are two ways to assign difficulty values to each question, cognitively and systematically [10]. The cognitive approach relies on the perceptions of educational professionals to decide on the difficulty of questions relative to each other. The systematic approach requires the calculation of item difficulties from assessment data.

4.1 The Cognitive Approach

For the first attempt at creating a CAT, the cognitive method could yield a useful set of question difficulties without too much effort. The set of questions should be examined and sorted into difficulty order using methods such as rank ordering [11] or features of the questions that range in difficulty such as number bonds or reading complexity. Once they are in order, the midpoint of the questions should be determined. The midpoint isn't necessarily the 50th question if there are 100 questions, it will be the point where an average ability student has an equal chance of getting a question right or wrong.

In a simple CAT this midpoint will be the starting question and when the CAT is running, the system will select a question of higher difficulty if the student gets this question right, and a lower difficulty if they get it wrong. For this reason, at this point it is a good idea to make sure that there are enough questions both above and below the midpoint for the CAT process to work well. With the questions in rank order, a difficulty value can now be assigned to each question. The midpoint question will have a value of 0 (typically indicating a 50:50 chance of an appropriately able student answering correctly), and the harder questions will have values in the range 0 to 3. Easier questions will have values in the range 0 to -3. The difficulty values will be decimals and do not have to have a fixed step. The result of this process will be a range of questions each with a difficulty value and this may be used directly in the CAT software to run the CAT.

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)
This version should not be distributed since it may change prior to final publication

4.2 The Systematic Approach- Calculating Item Difficulties

The second approach is to use Item Response Theory (IRT) statistics to calculate item difficulties. To calculate item difficulties, it is necessary to obtain responses from as many students as possible across the ability range to the questions you wish to use in order to obtain a grid of item level data. A typical set of data is a spreadsheet with a column for each question and a row for each student's answer. A correct answer should be indicated with a 1, and an incorrect answer with a 0.

Once you have item level data, statistical methods can be used to calculate the relative difficulty of each question. Two methods that are easily available and more importantly, free, are Ministeps [12], or RStudio [13]. Ministeps is free to download and very easy to use but is limited to 25 questions and 75 student rows. RStudio is not limited but requires a small amount of programming skill. RStudio uses libraries of functions to carry out statistical analyses and suitable libraries for calculating item difficulties are readily available. Generally, R packages are very well documented and easy to use.

The Ministeps program can be used to read a student response file and create an item measure table. Fig.1 shows a typical item measure table.

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
9	2	20	2.29	.77	.74	-.31	.44	-.49	.51	.24	90.0	89.9	Question9
8	4	20	1.40	.59	.98	.04	1.02	.22	.29	.30	85.0	81.0	Question8
4	5	20	1.08	.55	1.03	.18	.83	-.27	.34	.33	70.0	76.7	Question4
5	7	20	.54	.50	.93	-.30	.83	-.50	.45	.36	65.0	68.1	Question5
10	9	20	.05	.49	.91	-.57	.96	-.10	.46	.38	75.0	64.8	Question10
2	10	20	-.18	.48	1.02	.21	1.04	.26	.35	.38	60.0	65.7	Question2
3	11	20	-.41	.49	1.19	1.09	1.20	.89	.19	.39	65.0	66.8	Question3
6	13	20	-.90	.51	.94	-.21	.91	-.21	.45	.38	75.0	71.9	Question6
7	14	20	-1.17	.53	1.13	.59	1.21	.66	.22	.38	75.0	74.6	Question7
1	18	20	-2.70	.77	.99	.16	.98	.28	.26	.27	90.0	89.9	Question1
MEAN	9.3	20.0	.00	.57	.99	.09	.94	.07			75.0	74.9	
P.SD	4.7	.0	1.35	.11	.12	.46	.21	.45			10.0	8.9	

Figure 1 Ministeps Item measure table

Using this approach, Ministeps calculates the Joint Maximum Likelihood Estimate (JMLE) of the questions in the file. The JMLE column represents the relative difficulty of each question in the test the relative difficulty of each question in the test. The lowest number (-2.70) represents the easiest question and the highest number (2.29) the hardest question. So, for this data set, Question1 is the easiest (18 students out of 20 got this question right) and Question9 is the hardest (2 students out of 20 got it right).

4.3 Building the assessment

Although there are a number of commercial CAT packages available, one package that is readily available, and more importantly, free, is Concerto, produced by The Psychometric Centre of Cambridge University [14]. To use it, simply visit the psychometric centre website, download, and install the package.

Concerto is an implementation of the catR library [15,16], a set of functions for use with the R programming language that explicitly relate to the creation and running of a CAT. This means that the Concerto platform is not just a 'black box'. Reference to the catR papers allows the user to understand exactly which statistical processes are being used within the application, and to choose methods which are most appropriate for their test. Since CATs are not used only for educational purposes, but in other fields such as psychology or medicine, many of the functions offered by the Concerto package are not appropriate for education-based CATs.

The Concerto system uses a very simple interface to create assessments. Each question is added to the system and the screen the student will see is created using a simple HTML editor. The correct answer and two or more incorrect answers are added (the distractors), and the difficulty value assigned to the question. It is also possible to bulk load questions from a spreadsheet.

The Concerto interface incorporates a simple 'wiring diagram' model that allows rapid creation of assessments. Right clicking on the screen allows the user to select an item to appear as an element of the assessment which is placed between the 'test start' and 'test end' boxes. This includes accepting username and passwords, assessment, and reporting screens. Adding an 'assessment' element makes the screen as shown in Fig. 2.

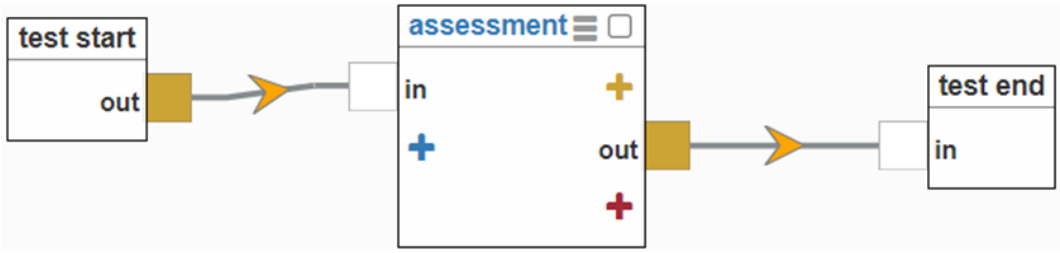

Figure 2 Concerto test construction screen

Questions, and for a simple multi choice assessment, distractors, are added to the system along with a difficulty value. The first few questions and associated data for a fictional assessment might look like that in Table 2.

Table 1 Fictional CAT data

Question	*Correct answer*	*Distractor 1*	*Distractor 2*	*Distractor 3*	*Difficulty*
1 + 1 =	2	1	3	4	-2.5
2 + 1 =	3	5	2	4	-2
2 + 2 =	4	3	5	2	-1.5
3 + 2 =	5	3	6	4	-1
4 + 4 =	8	4	6	10	0
5 + 5 =	10	5	12	8	1
10 + 5 =	15	10	50	5	1.5
10 + 100 =	110	1000	90	100	2
100 + 200 =	300	30	400	250	2.5
1000 + 200 =	1200	300	600	2000	3

Note that 1 + 1 = (difficulty -2.5) is the easiest question, and 1000 + 200 = (difficulty 3) is the hardest question. This is enough information to run a simple CAT.

4.4 First Item Selection

When presenting the first item to a student, some decisions need to be made. The first item must be appropriate for the student and fall into their ability and possibly age range. In some systems, a question is selected from a parallel item bank to avoid the student having seen the question before. Commonly, the first item offered has a difficulty of 0 or lower. Ideally, it should be a question that most test takers will find comparatively easy. In more sophisticated systems, the first item offered could be based on the test taker's age or ability measured on a previous test. In the Concerto system the CAT may be started anywhere, simply by selecting one of the items. If the test is trialled and that point is too hard or too easy, it can be easily changed.

4.5 Subsequent Item Selection

Once the response to the first question is known it is evaluated to see whether the answer was correct or incorrect. If the answer was correct, another question of higher difficulty should be selected. If the answer was incorrect, a question of lower difficulty should be offered. Some simple CATs may offer the next question in the item bank, either easier or harder as required, but doing so might not be an efficient way of establishing student ability. Using this simple method, students could find themselves answering many more questions than necessary before their ability level is established as they work their way

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)
This version should not be distributed since it may change prior to final publication

through the questions in the bank. In most cases it is better to jump to a new point in the range of items in the item bank relating to the construct being assessed and ask a question at that point. The decision as to which question should be offered will depend on several factors such as the size of the item bank, the desired jump (or drop) in difficulty, and the availability of questions of suitable difficulty. How the next item is offered is based on the next item selection algorithm. This can vary, but by default, the Concerto system uses the Maximum Fisher Information (MFI) algorithm, which is appropriate for most educational cases. This set of decisions will then be repeated until a specific stopping rule comes into play.

4.6 Stopping Rules

The CAT finishes when a certain condition (or conditions) is met. Commonly the CAT stops when a stable ability score is reached, but the test may also finish if a given number of questions has been asked, a particular time limit reached or, in extreme circumstances, the item bank is exhausted.

The Concerto system has provision for time, ability score and number of questions to be used singly or in combination as a stopping point. If selected, the Concerto system will automatically calculate a current ability value (theta) and the standard error of measurement (SEm), a measure of convergence around a specific ability level. The standard error of measurement estimates how repeated measures of a student on the same assessment tend to be distributed around their "true" score. The true score is always an unknown because when dealing with human subjects there is always a degree of variation. Although this variation can be accounted for, no measure can be constructed that provides a perfect reflection of the true score. Therefore, in terms of reaching a specified ability measure, the standard error of measurement (SEm) is commonly used with a suitable cut off point. As more information is known from the answers given to previous questions, the SEm tends to fall and provides a good measure for reaching a suitable ability estimate. A SEm measure of less than 3 is an appropriate level for convergence. SEm is directly related to the reliability of a test; that is, the larger the SEm, the lower the reliability of the test and the less precision there is in the measures taken and scores obtained.

4.7 Assessment Completion

Once one of the stopping rules above has been reached, the assessment can end, and the result calculated and presented. In some cases, such as a time-out, it may not be possible to calculate a score with enough accuracy, and in those circumstances the student could be advised to take the assessment again. Other cases that might result in an inability to calculate a score might be too few questions answered, or an inordinate amount of guessing where the results set is effectively random. Fig. 3 shows the result set for a 100 item CAT which starts at 50 where the test taker has an ability around 67.

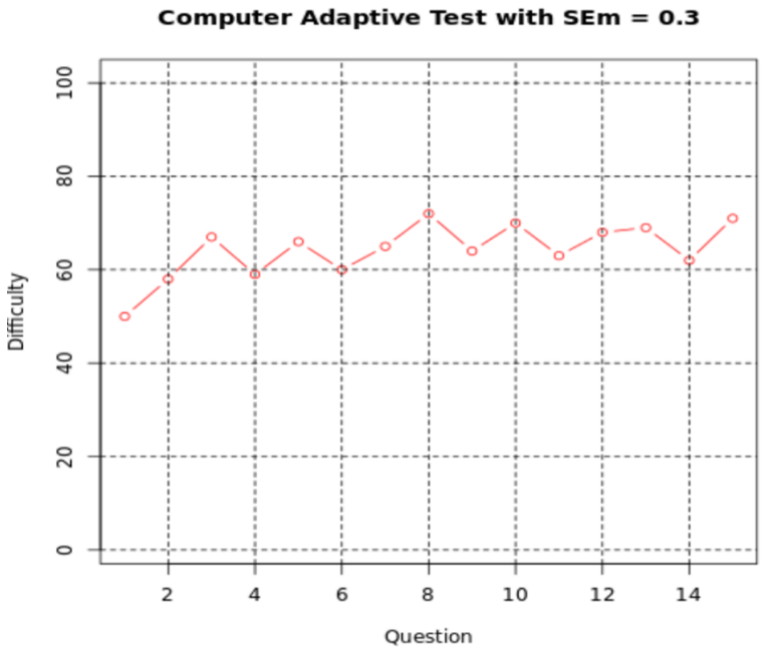

Figure 3 CAT results for a 100 item test where the ability of the test taker is around 67

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)
This version should not be distributed since it may change prior to final publication

Fig. 4 shows how the calculated SEm decreases with each successive response.

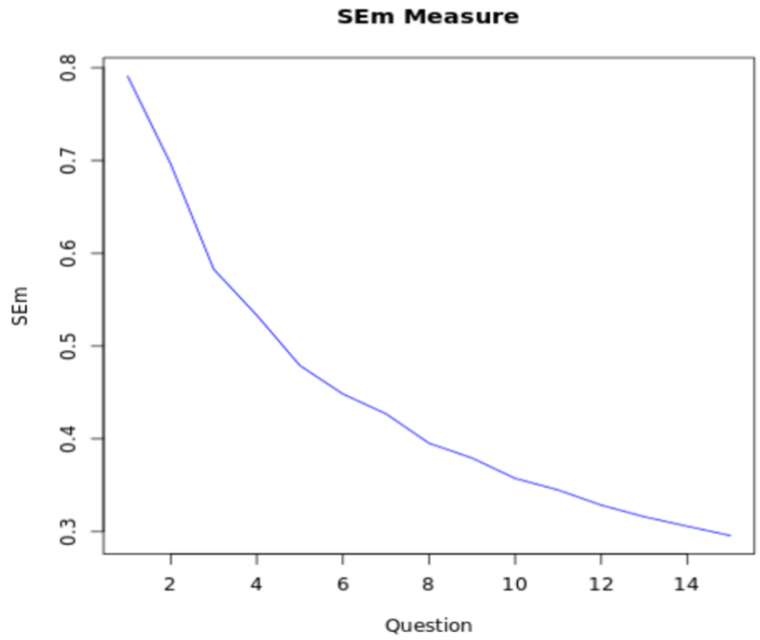

Figure 4 SEm plotted over a CAT test

4.8 Reporting

Once the CAT has completed, a score of some type is normally calculated from the final ability estimate, theta. The raw theta is a measure on the difficulty scale which may be of value, but it may be better to convert it into a more understandable score. The Concerto system allows the user to add a scoring box to the system to convert the theta value to a standardised score based on the distribution of difficulties in the item bank.

5 DISCUSSION

Although CAT testing is a closed book to many people, that doesn't necessarily need to be the case. As CATs are used more frequently in education and increasingly in high stakes examinations, it is important that they are understood by those who are educating students who will use them. Of course, building a CAT for use internationally, by many different students studying various curricula is a complex issue and much planning, effort and analysis is required to achieve those goals. For example, building a CAT which is required to cover a wide range of constructs or where language use is restricted to a specific CEFR level can be very demanding. The statistical IRT models used can also indicate gender bias, or questions where the wording is ambiguous to test takers, even if it is not apparent to the question setter. For these reasons, and others, if a school opts to carry out systematic testing of its students and wishes to compare the results over time, a school created CAT might not display the reliability, validity or robustness required for that purpose. For those reasons, particularly where academic interventions or other decisions will be made that could affect student's academic performance it is best to use a CAT from a trusted organisation such as Cambridge CEM.

Many papers [9, 17, 18] have been published detailing the use of CAT systems in a wide variety of domains, but a lot can be learnt about the process of CAT testing, the requirements of a good CAT and different application methods by studying a simple CAT such as the one detailed here. Indeed, creating a simple CAT and using it with students in schools is a good way to understand how it works and could also form the basis of a class based project.

As the use of AI becomes more established in schools and colleges, the next generation of CATs is likely to integrate some form of AI content, and AI techniques might be used to establish larger item banks for future adaptive tests. This is likely to increase the development of such tests and thereby increase their scope across the curriculum. There is no doubt that the benefits of computer adaptive testing are such that such tests will be a common part of a student's assessment as they follow their educational journey, and it is vitally important that students are well prepared to take on the challenge.

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)
This version should not be distributed since it may change prior to final publication

REFERENCES

[1] Bain SK, Allin JD. Book review: Stanford-Binet intelligence scales. Journal of Psychoeducational Assessment. 2005 Mar;23(1):87-95.

[2] Lord FM. The self-scoring Flexilevel test. ETS Research Bulletin Series. 1970 Dec;1970(2):i-9.

[3] Lord FM. A theoretical study of the measurement effectiveness of Flexilevel tests. Educational and Psychological Measurement. 1971 Dec;31(4):805-13.

[4] Jensema CJ. The validity of Bayesian tailored testing. Educational and Psychological Measurement. 1974 Dec;34(4):757-66.

[5] Reckase MD. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods & Instrumentation. 1974 Mar;6(2):208-12.

[6] Rasch G. Probabilistic models for some intelligence and attainment tests. MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637; e-mail: MESA@ uchicago. edu; web address: www. rasch. org; tele; 1993.

[7] DeAyala RJ, Koch WR. A Computerized Implementation of a Flexilevel Test and Its Comparison with a Bayesian Computerized Adaptive Test.

[8] Spray JA, Reckase MD. The Selection of Test Items for Decision Making with a Computer Adaptive Test.

[9] Linacre JM. Computer-adaptive testing: A methodology whose time has come. MESA memorandum; 2000.

[10] AlKhuzaey S, Grasso F, Payne TR, Tamma V. A systematic review of data-driven approaches to item difficulty prediction. In International Conference on Artificial Intelligence in Education 2021 (pp. 29-41). Springer, Cham

[11] Bramley T. A rank-ordering method for equating tests by expert judgment. Journal of Applied Measurement. 2005 Jan 1;6(2):202-23.

[12] https://www.winsteps.com/ministep.htm

[13] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[14] https://concertoplatform.com

[15] Magis D, Raîche G. Random generation of response patterns under computerized adaptive testing with the R package catR. Journal of Statistical Software. 2012 May 24;48:1-31.

[16] Magis D, Barrada JR. Computerized adaptive testing with R: Recent updates of the package catR. Journal of Statistical Software. 2017 Jan 11;76:1-9.

[17] Merrell C, Tymms P. InCAS (Interactive Computerised Assessment System): Using individual diagnostic profiles in assessment for learning. In EARLI Conference, Nicosia, Cyprus 2005 Aug.

[18] Meijer RR, Nering ML. Computerized adaptive testing: Overview and introduction. Applied psychological measurement. 1999 Sep;23(3):187-94.

This article will be included in the INTED2024 Proceedings (ISBN: 978-84-09-59215-9)
It will be fully citable as soon as it appears in IATED Digital Library (library.iated.org)
This version should not be distributed since it may change prior to final publication