Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and

Statement of Research Intent

Project Number: 52

Applications of Computer Adaptive Testing in Educational Assessment

Oscar Li

Hajin Kim

Yu-Cheng Tu (Supervisor)

Andrew Meads (Co-supervisor)

13/04/2022

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Signature: 

Name: Oscar Li

ABSTRACT: Assessment plays a vital role in all educational settings. As a consequence, it is of utmost importance to have tests that accurately measure a student's ability. Computer adaptive testing is a testing paradigm that differs from standard fix-length tests in hopes of achieving greater measurement precision while also requiring fewer test items to be administered to the user. This is achieved through the use of an algorithm that dynamically selects future test items based on the student's response to previous items. Until now, educational assessment has mainly been conducted through fixed-length tests due to ease of implementation and administration. However, technological advances and the gradual adoption of electronic devices in the classroom has meant that computer adaptive testing is becoming increasingly attractive for teachers and students. This literature review explores various applications of computer adaptive testing and the theoretical models behind them to understand the potential for further adoption in educational domains.

1. Introduction

Testing plays a crucial role in the educational process. There are two forms of testing, formative and summative. Formative tests are used to attain an understanding of the student's current knowledge to inform the teaching and learning process [1]. Summative tests, on the other hand, focus on measuring performance after the student has completed a programme [1]. In both cases, good test design is crucial to get an accurate assessment of student ability. This is important for a multitude of reasons including but not limited to: helping the student improve by identifying gaps in their knowledge, providing feedback to teachers, evaluating the relative performance of students as well as the overall performance of a particular population [2]. However, the current most popular testing paradigm of fixed length testing is not without issue. The key limitations of which are test length and test difficulty [3]. Test creators must strike a delicate balance of having a test that is comprehensive enough to give an accurate measure of ability while also trying to maximise testing efficiency. Another key facet is test difficulty. To get an accurate measure of performance for a wide range of ability levels, the test needs to have questions of various difficulties. An inefficiency of this approach is the poor measurement accuracy for students of outlying abilities [4]. High ability students must spend extraneous time answering easy questions and likewise for low ability students with difficult questions. As a result, the topic of computer adaptive testing has been a major area of interest for researchers and educational institutions. The results of their efforts and the progress made so far will be discussed in later sections.

2. Traditional Computer Adaptive Testing Theory

Computer adaptive testing differs from traditional fixed length tests in that it adapts to the user's ability. The goal of this approach is to maximise assessment precision while requiring the same amount or fewer test items. Generally, CAT can

be seen as being composed of six primary components: item response model, item pool, entry level, item selection, scoring method, and termination criterion [3].

2.1. Item Response Model

An item response model is used to identify the probability that a given test taker will answer a test item correctly. As a result, calculation using a model requires certain parameters such as item difficulty, item discrimination (how effective the item is at determining the ability of a test taker) and the chance of guessing a correct result. Most item response models are based on item response theory (IRT) and can be classified as one (1PL), two (2PL) or three (3PL) parameter logistic models [3]. In 1PL the only parameter considered is item difficulty, 2PL expands upon this by adding a discriminating factor and 3PL also adds another parameter to account for guessing. The models described are all dichotomous which are targeted at test items that only have two possible scores: correct or incorrect. There also exist polytomous models such as Rating Scale Model [5] and Partial Credit Model [6] which are used for items with more than two possible scores such as Likert-type items.

2.2. Item Pool

Based on the previous information about the item response model, it logically follows that a computer adaptive test must utilise a pool of questions that have been calibrated with the correct IRT parameters. Based on the literature, a pool of 100 items is usually sufficient to give acceptable results [3]. Additionally, it is highly important to have questions of varying difficulty and that the questions can discriminate between test takers of different ability.

2.3. Entry Level

Adaptive testing allows test takers to begin the test at different difficulty levels without having a significant effect on the measurement outcome [3]. This is because an adaptive test will gradually adjust to the test taker's ability level as they answer questions. However, to provide the most precise measurement within a certain number of items, test takers should begin at a level matching their ability [7].

2.4. Item Selection

Item selection plays a key role in CAT as the goal at each stage of the test is to select the item that will reveal the most information about the test taker's ability [3]. The two most used and researched item selection procedures are the Maximum Information [8] and Bayesian [9] procedures. Maximum Information seeks to find an item that will give the most information at the test taker's current estimated ability whereas Bayesian seeks to minimise the variance of the

current estimated ability. Most CATs are conducted with a single procedure, however, there has also been research into using multiple procedures in the same test with promising results [10].

2.5. Scoring method

A computer adaptive test must provide a new estimate of the test taker's ability after each item has been answered. Generally, this means a correct answer will result in an increased estimated ability and vice versa. However, the exact method used varies and can be done using either a maximum likelihood estimation or a Bayesian estimation method [3].

2.6. Termination criterion

A stopping criterion is required to inform a CAT system when no further items are required to be administered to the test taker. Generally, the stopping criterion is specified by the test maker as an acceptable measurement error whereby the test concludes once the measurement error of the estimated ability falls within the specified range [3].

Each component discussed above needs to be considered when designing a computer adaptive test. However, there is significant flexibility in how each component is implemented which can be seen in the following section.

3. Innovations in Computer Adaptive Testing Theory

Research has also been conducted in CAT that deviates from traditional approaches. The main motivation of these methods is to address specific issues with traditional CATs such as the difficulty of assigning item parameters and the need for significant mathematical computation due to recalculation of ability after every test item.

3.1. Elo Rating System

A new item response model is proposed in [11] is based off the Elo system which is used to rank players in zero-sum games such as chess. The authors of the paper identified three key issues with the traditional CAT models based on item response theory. The first issue with traditional CAT was that the item parameters such as difficulty and discrimination had to be calculated in advance before the item could be used in a test environment which is both time-consuming and costly. The second issue discussed was test taker motivation regarding the difficulty of the administered items. A traditional CAT is most precise when the administered test item has a 50% chance of being answered correctly by the test taker, however, this success rate is likely to be perceived as discouraging. The last issue is the failure of traditional CATs to account for the trade-off between speed and accuracy which can be highly valuable for ability estimation. To address these issues, the authors developed their novel Elo Rating System for CAT. Their new CAT allowed for on-the-fly item calibration which addressed the first issue by removing the need for pre-testing. This was because students and test items

started with a baseline rating which was automatically adjusted based on whether the student answered the question correctly. They also integrated speed into the scoring system which addresses the second and third issues as it allowed for easier test items to be used with only a minor loss of measurement precision. This was achieved by using response time as a score multiplier i.e. a quick answer has a larger effect on score.

3.2. Multistage Adaptive Testing (MST)

Multistage adaptive testing differs from the traditional CAT, also known as the question adaptive model (QAM), in the way test items are administrated. While QAM selects a new test item after each item has been answered, MST instead does this in sets of questions known as testlets or stages [12]. Similar to QAM, the next set of questions administered to the participant in MST depends on their performance on previous item sets. Research has been conducted comparing the effectiveness of both approaches and results suggest that MST has greater measurement precision compared to QAM when using one-parameter and two-parameter logistic models and they performed equally for three parameter logistic models [13]. Another benefit of this approach is that test creators have more control over content balancing. The issue with content balancing occurs in QAM because examiners generally want tests to contain sufficient content coverage from multiple categories. In QAM, there is no guarantee that items will have balanced exposures without some constraints being imposed which may limit the measurement efficiency [14]. This issue is mitigated by the coarse-grained nature of MST because examiners can decide the individual items within a set and hence have more control over content coverage. It also becomes easier to enforce limits on question set reuse and item overlap which improves test security [15]. Such characteristics makes MST more favorable in practice as tests are expected to be secure while covering a wide range of topics. Consequently, the popularity of MST is reflected in the case studies discussed below.

4. Case Studies

4.1. Graduate Record Examination (GRE)

The Graduate Record Examinations is a test created by Education Testing Service (ETS) and is required for admission into graduate schools in the US and Canada [16]. The test was originally designed to measure three aspects, each with their own section: analytical, quantitative, and verbal reasoning. A key feature of the GRE was the transition to computer adaptive testing in 1993. Questions were given in the form of four/five-option multiple choice questions and a 3PL IRT model was used for scoring. To ensure the reliability and validity of the new testing method during its initial introduction, there had to be evidence to show that the new CAT was comparable to the original test. Studies found that the verbal and

quantitative scores from the CAT were comparable to the linear test while the analytical scores were comparable after appropriate adjustments were made [17]. The GRE serves as a solid foundation for understanding the history of CAT design and has inspired much research into the adoption of CAT within other educational domains.

It is also important to note that the GRE CAT itself has had major changes since its initial introduction. One such change was the transition to multistage testing in 2014 [18]. One motivation for this change was to improve the test taking experience. The original CAT did not allow participants to revisit previously answered questions whereas MST allows participants to freely move between questions within the same section. Another major benefit was the reduction in complexity and development cost when changing to MST as it required a smaller overall item bank to achieve the same content coverage. The final implementation of the revised GRE consisted of a 2-stage test with the first stage being known as a "routing test" whose results were used to decide the difficulty of the second stage test. The scoring model used was a 2PL IRT model which was based only on question difficulty and discrimination. The revised GRE introduced a wider variety of question types over pure multichoice; hence the effect of guessing was minimized and a 3PL IRT model was not required.

4.2. National Assessment Program – Literacy and Numeracy (NAPLAN)

NAPLAN is a test administered by the Australian Curriculum, Assessment and Reporting Authority (ACARA) whose goal is to provide a snapshot of literacy and numeracy skills in Australian students [19]. The main interest in this case study lies in their transition from paper-based testing to computer adaptive testing. A key decision that had to be made was the choice between an item-level computer adaptive test or a multistage test. The key strength of traditional CAT is that it provides the highest possible measurement precision while using relatively few test items. MST, on the other hand, allowed for more control over content coverage and requiring fewer test items to implement (as described in previous sections). As a result, the ACARA ultimately decided to use MST. A tailored test consisting of three stages and two branching points was proposed [20]. Pilot testing of the tailored test found that the standard error of measurement for ability estimates in students was significantly lower compared to fixed linear testing. This method also improved testing motivation of struggling students as poor performance at the first stage resulted in easier questions in later stages.

5. Frameworks

Due to the interest in CAT, various researchers have developed open frameworks that implement common CAT components with the aim of facilitating rapid prototyping and testing.

5.1. Open-Source Computer Adaptive Testing System (OSCATS)

OSCATS is a CAT framework developed in C which aims to provide implementations of CAT routines such as item response models, item selection algorithms and statistical analysis tools [21]. The framework supports 1PL, 2PL, 3PL IRT models and various item selection models such as Maximum Fisher Information and Kullback-Leibler Index. Another benefit of OSCATS is modularity as it allows for different models and algorithms to be swapped in and out. This is achieved through object-oriented programming whereby each feature is represented by an object therefore allowing them to be easily replaced without significant modifications to the code. The use of object-oriented programming also allows for extensibility which gives developers the freedom to modify features for their own needs. Lastly, OSCATS aims to be highly accessible by providing bindings in other programming languages such as Python and Java which encourages developers to work in their most familiar language.

5.2. catR

catR is another useful CAT framework which is provided as an R package [22]. Much like OSCATS, catR provides implementations of common IRT models such as 1PL, 2PL, 3PL, and 4PL models and item selection algorithms such as Maximum Fisher Information, Kullback-Leibler. However, catR is more extensive in the models and algorithms it provides. For example, catR also provides polytomous IRT models such as Graded Response, Partial Credit, Generalised Partial Credit, Rating Scale, Nominal Response. It is important to note that catR does not support item bank calibration and instead recommends the use of another package, mirt, which is aimed specifically at analysing response data using latent trait models [15]. A limitation of catR in comparison to OSCATS is the lack of language support as catR can only be used with the R programming language.

6. Discussion

Based on the case studies examined above, it seems that although pure question-adaptive CAT provides more measurement precision in theory, it is often less feasible in practice due to content constraints and item pool sizes. Consequently, both NAPLAN and GRE have opted for a multistage testing approach. However, one key difference between the two tests is the number of stages involved. The NAPLAN contains three stages whereas the revised GRE only contains two. One possible reason for this may be due to the goals of each test. The NAPLAN aims to identify numeracy and literacy level of all Australian students and draw conclusions from the data. This means it is in their best interest to provide ability estimation for the widest possible ability range. This contrasts with the GRE which measures the ability of graduate students for college admissions and hence are more likely to be concerned with whether participants meet a certain standard. It is also important to note that the GRE consists of three separate sections, each with their own

multistage test; hence it may not be feasible to have more than two stages due to time constraints. In terms of question format, both the NAPLAN and GRE contain multiple choice questions but the GRE also includes short and long answer questions. In general, it seems that multiple choice questions are easier to develop question banks for as there is only a single correct answer and no partial answers. However, the scoring model in these tests needs to account for the effect of guessing which is why multiple choice CATs tend to favour the use of the 3PL IRT model. It is also of interest to discuss the demographic of each test. The NAPLAN is targeted towards primary and intermediate students whereas the GRE is for graduate students. This suggests that there is a relative lack of research on computer adaptive testing for adolescents i.e. high school students. Based on the research conducted various CAT frameworks, it seems that catR offers the most potential customisation due to their extensive library of IRT models and algorithms. Additionally, catR is more popular and actively maintained in contrast to OSCATS which has not been updated since 2011. An example is the online adaptive testing platform, Concerto, which is built on top of catR. The combination of these factors makes catR an attractive choice for future research in CAT.

7. Research Intent

Most of the research so far has been conducted by large institutions aimed at creating tests for thousands of participants and there is a relative lack of research discussing the feasibility of CAT at smaller scales. Additionally, a significant amount of research on CAT has been conducted on either primary school or university students. Based on these points, we aim to explore the feasibility of implementing a CAT system targeting high school students using tools and software available to the wider population such as catR. Our plan is to achieve this is by implementing a prototype using the CAT models described in the literature and measuring its effectiveness by testing on high school students.

To successfully build and test the prototype, we will need to consider the following:

- Choice of CAT framework to be used for implementation. Based on our findings, catR is a good choice due to its popularity and the high quality of documentation provided.
- Whether to use a question adaptive or multistage model. However, there is potential to try both approaches if time permits.
- The test format and item response model. We are currently inclined to follow a multichoice format due to abundance of sample questions available and ease of marking. As a result, the 3PL IRT model is a superior choice compared to 1PL and 2PL since it accounts for guessing.
- The test demographic. Based on the gap in literature identified above, we aim to design our prototype for testing on high school students (year 9-11).

7.1. Timeline

April: Complete Literature review (15th), Ethics approval application, Begin CAT implementation

May: Complete Early Seminar (14th), Prepare a calibrated item pool, Continue CAT implementation

June: Complete the CAT MVP, Conduct testing on the SOFTENG cohort

July: Progress video (22nd), Iterate upon CAT MVP, Design an empirical study on high school students.

August: Conduct a pilot empirical study on high school students.

September: Document findings and results from study, Improve CAT MVP based on findings

October: Finish Final Report (14st), Project Compendium (17th), Project Seminar (20th)

8. Conclusion

Computer adaptive testing is an attractive topic in the domain of education due to its potential benefits to formative and summative assessment. Traditional computer adaptive testing is composed of six main components: item response model, item pool, entry level, item selection, scoring method, and termination criterion. However, there is also promising research that innovates on traditional CAT such as multistage testing or using an Elo rating system. This interest has led to the use of CAT in assessments such as the GRE and NAPLAN which are targeted towards graduate and primary/middle school students, respectively. Research in this area has also led to the creation of CAT prototyping tools such as OSCATS and catR. However, there are still gaps in the literature which form the basis of our project. Our aim is to develop a CAT software prototype aimed towards high school students with the goal of identifying the feasibility of CAT being adopted for small scale testing.

References

[1] B. Ussher and K. Earl, "'Summative' and 'Formative': Confused by the Assessment Terms?" *New Zealand Journal of Teachers' Work*, vol. 7, no. 1, pp. 53-63, 2010.

[2] H. L. Roediger III, A. L. Putnam, and M. A. Smith, "Ten benefits of testing and their applications to educational practice," *Psychology of Learning and Motivation*, vol. 55, pp. 1–36, 2011.

[3] D. J. Weiss and G. G. Kingsbury, "Application of computerized adaptive testing to educational problems," *Journal of Educational Measurement*, vol. 21, no. 4, pp. 361–375, 1984.

[4] W. J. Linden, W. J. van der Linden, and C. A. Glas, Computerized adaptive testing: Theory and practice. Springer, 2000.

[5] E. B. Andersen, "The Rating Scale Model," *Handbook of Modern Item Response Theory*, pp. 67–84, 1997.

[6] G. N. Masters, "A Rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, 1982.

[7] S. L. Wise and G. G. Kingsbury, "Practical issues in developing and maintaining a computerized adaptive testing program," *Psicologica*, vol. 21, pp. 135-155. 2000.

[8] D. J. Weiss, "Improving measurement quality and efficiency with adaptive testing," *Applied Psychological Measurement*, vol. 6, no. 4, pp. 473–492, 1982.

[9] R. J. Owen, "A bayesian approach to tailored testing," *ETS Research Bulletin Series*, vol. 1969, no. 2, pp. i-24, 1969.

[10] V. M. G. Jatobá, J. S. Farias, V. Freire, A. S. Ruela, and K. V. Delgado, "Alicat: A customized approach to item selection process in computerized adaptive testing," *Journal of the Brazilian Computer Society*, vol. 26, no. 1, 2020.

[11] S. Klinkenberg, M. Straatemeier, and H. L. J. van der Maas, "Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation," Computers & Education, vol. 57, no. 2, pp. 1813–1824, 2011.

[12] A. Zenisky, R. K. Hambleton, and R. M. Luecht, "Multistage testing: Issues, designs, and research," Elements of Adaptive Testing, pp. 355–372, 2009.

[13] O. Rotou, L. Patsula, M. Steffen, and S. Rizavi, "Comparison of multistage tests with computerized adaptive and paper-and-pencil tests," ETS Research Report Series, vol. 2007, no. 1, pp. i-27, 2007.

[14] T. J. H. M. Eggen, Overexposure and underexposure of items in computerized adaptive testing. Arnhem Netherlands: Citogroep, 2001.

[15] D. Magis, D. Yan, and A.A. von Davier, *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR*. Springer, 2017.

[16] C. N. Mills and M. Steffen, "The GRE computer adaptive test: Operational issues," *Computerized Adaptive Testing: Theory and Practice*, pp. 75–99, 2000.

[17] G. A. Schaeffer, M. Steffen, M. L. Golub-Smith, C. N. Mills, and R. Durso, "The introduction and comparability of the computer adaptive GRE general test," *ETS Research Report Series*, vol. 1995, no. 1, pp. i-48, 1995.

[18] C. Wendler and B. Bridgeman, *The Research Foundation for the GRE Revised General Test: A Compendium of Studies*. Education Testing Service, 2014.

[19] "NAPLAN – general," *National Assessment Program*. [Online]. Available: https://www.nap.edu.au/naplan/faqs/naplan--general. [Accessed: 13-Apr-2022].

[20] "Tailored test design study 2013: Summary research report," *National Assessment Program*. [Online]. Available: https://nap.edu.au/docs/default-source/default-document-library/tailored_test_design_study_2013_summary_research_report.pdf. [Accessed: 13-Apr-2022].

[21] M. J. Culbertson, "Rapid CAT Prototyping with OSCATS: The Open-Source Computerized Adaptive Testing System," 2011.

[22] D. Magis and J. R. Barrada, "Computerized adaptive testing with *R*: Recent updates of the package *catR*," *Journal of Statistical Software*, vol. 76, 2017.