

Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and
Statement of Research Intent

Project Number: 52
Computerised Adaptive
Testing for Educational
Assessment

Hajin Kim

Oscar Li


Yu-Cheng Tu (Supervisor)

Andrew Meads (Co-supervisor)

15/04/2022

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: Hajin Kim

ABSTRACT: Adaptive testing is a relatively new paradigm that seeks to replace fixed-length tests. Adaptive testing delivers the most relevant questions for the candidate's ability, which leads to the benefit of reduced test length and accurate assessment of students at extreme ends of the ability spectrum. As computer capabilities have grown exponentially, implementing sophisticated adaptive testing strategies has become increasingly feasible. This literature review aims to explore and compare different computer adaptive testing (CAT) strategies in both theoretical basis and case studies and investigate the potential for computer adaptive testing to improve educational tests given the current state of the art technologies and frameworks.

1. Introduction

Tests in an educational context are usually administered to measure a student's ability and thereby construct an ordering between students' ability in a particular subject domain [1]. The result from a test provides feedback to students regarding what knowledge gaps they have and provides feedback to instructors on how they might teach their students next. Fundamental to test design is an assumption that every student has a measurable latent trait - ability - that correlates with their performance in a particular test [2]. Therefore, a student's performance in a test is an index of their trait [2] [3], and it is in the best interest of everyone involved in the test that the test accurately measures the student's ability.

Computer Adaptive Test (CAT) is an attempt to alleviate the difficulty of the Fixed Length Test in accurately measuring a student's ability at any ability level. It is a long-known issue that the Fixed Length Test does not accurately evaluate a candidate's ability at high and low ability levels [4]. This is because every candidate receives the same questions regardless of their ability. In contrast, CAT administers questions that best fit to students ability while iteratively adjusting its ability estimate as the student responds. Hence, CAT extracts maximum information about the student regardless of their ability and requires fewer items to reach an accurate ability estimation than the Fixed Length Test.

Much of the theoretical and practical basis for the Computer Adaptive Test was developed in late 1960 when Fred Lord started his research on Item Response Theory (IRT) [5]. IRT is a central theory to CAT. It allows estimation of ability independent of test item selection - that is, the ability estimation based on two different sets of questions is comparable [6]. CAT is a testing paradigm that leverages the power of interactive, powerful computers and decades of research around adaptive testing and item response theory.

This literature review will focus on the goals and theoretical background of CAT, existing efforts in CAT in terms of case studies, and frameworks that support the development and research of CAT. Then, we will identify patterns and gaps across the papers and suggest a direction for our research.

2. Computer Adaptive Test (CAT) Theory

The goal of CAT is to reduce test length while offering an increased precision on the ability estimates. In this section, we explore CAT theory and its various components that enable successful CAT implementation.

2.1. Theoretical background

Typical components of CAT are the item response model, item pool, entry-level, scoring method, item selection, and terminating criterion [7]. It is essential to consider each of these components to construct a well-designed CAT. Here, we briefly discuss each component and how they relate to effective CAT design.

2.1.1. Item response model (IRT model)

The item response model maps the current ability estimate to a candidate's likely response to an item – usually the probability of getting it correct. Item response model determines what parameters will be taken into account. The 3-parameter logistic (3PL) model considers three parameters: discrimination, difficulty, and guessing. Less general variants (1PL, 2PL) are also widely used as IRT models in CAT implementation and are respectively constrained by one (discrimination) and two (discrimination, difficulty) parameters [7]. Less information on particular parameters would lead to choosing the logistic model with fewer parameters. Estimating parameters for these logistic models is discussed further in the upcoming section (2.3 Estimating parameters for IRT models).

2.1.2. Entry-level

Entry-level is the initial question administered to a candidate. Entry-level is vital because it can impact the psychological state of the candidate [8]. If too difficult, the candidate might perform worse due to anxiety; if too easy, the candidate can become careless and create mistakes. Gershon [9] suggests initially administering an accessible item to minimise the effect of test anxiety on a candidate's performance. If some information about the candidate's ability is known, entry-level can consider that information for the initial item selection [10].

2.1.3. Scoring method (Ability estimation)

The ability estimation algorithm is principal to CAT, as it estimates ability at each candidate's response and ultimately provides the final estimate. Models such as ML (maximum likelihood), BM (Bayes modal), WL (weighted likelihood), and EAP (expected a posteriori) estimators are used [11], and the idea is that the ability of the candidate is a value that

maximises or minimises a particular objective function. For example, the ML estimator estimates the ability to be a value that maximises the maximum likelihood function. Details of those models are well described in this famous paper [4].

2.1.4. Item selection

The item selection algorithm determines what item will be administered to the candidate during the test. Algorithms and administrative models for item selection are discussed further in the next section (2.2 Computer adaptive testing administrative models).

2.1.5. Item pool (Item bank)

The item bank is a collection of questions that can be administered to candidates. Estimating a candidate's ability requires enough questions of a difficulty that match the candidate's ability [10]; that is, difficult items are most informative to generate estimation about a candidate with high ability, and vice versa. However, little is known about what constitutes a 'large enough' number of questions, and the number varies from a hundred to thousands [10] [12].

2.1.6. Terminating criterion

Theoretically, CAT should stop when we are confident enough about ability estimates. This can be done in three ways: stop when a certain test length is reached, stop when the standard error of ability estimate becomes less than a set threshold, and stop when a pass/fail condition is reached [7].

2.2. Computer adaptive testing administrative models

Item selection rules differ depending on what CAT administrative model is chosen. Test administration model is an important decision as it affects measurement efficiency, test security, and design complexity. Three administration models often discussed in case studies or applications are the Multistage Adaptive Test (MST), Question-Adaptive Model (QAM), and Computerised Classification Test [6]. The computerised Classification Test will not be discussed since this administrative model is catered for pass/fail tests which are not our main focus of discussion.

The Question Adaptive Model (QAM) determines the next question based on the ability to estimate a student's performance level. Two popular methods for item selection are Maximum Information (chooses a question that can maximise the information known about the student at a given ability estimate) and Bayesian item selection (selects a question that minimises the 'posterior belief distribution' - or uncertainty - about candidate ability estimation. Initial GRE CAT used this administrative model [13].

In the Multistage Adaptive Test (MST), a candidate does a *routing* test that is usually a small set of questions. Depending on their performance, they are assigned to different testlets (predetermined sets of items, each set classed by difficulty). Depending on the candidate's combined performance on both the first routing test and second test, they can be further routed to different tests (or MST can end here). Ability estimation takes account of all the testlets the candidates have completed. MST is popular among educational CAT implementations such as the revised GRE (Graduate Record Examination) [13] and NAPLAN (National Assessment Program - Literacy and Numeracy) [14]. MST is usually compared against QAM, and this is discussed further in the case studies section.

2.3. Estimating parameters for IRT models

One major practical concern for implementing CAT is calibrating item pools for parameters of logistic models. In the conventional CAT, such calibration needs to be done prior to testing, which is cumbersome for a large number of items [10]. Known calibration methods are Conditional Maximum Likelihood [15], Marginal Maximum Likelihood [16], Bayesian Model Estimation [17], and Joint Maximum Likelihood [18], which are all computationally demanding.

ERS, a relatively new model in CAT, can be put into production without any prior calibration as it supports on-the-fly item difficulty estimation [19]. There have been a handful of empirical studies on the use of ERS in the context of CAT. A major hurdle to using ERS is that ERS estimates students less accurately than the standard CAT model [19] due to the time required for the difficulty estimation to reach a stable value. Some literature considers a hybrid approach where ERS is used in the initial phase of CAT to calibrate the item pool and more computationally heavy methods are used later in time (e.g. JML, MML) [20].

Some researchers have also discovered that item parameters do not require the highest accuracy estimation: 'educated guess is good enough' [21]. It is agreed upon that experts can assign relative difficulties to items in a fairly robust manner [8]. Such statements open doors for practical implementations of CAT in that teachers can, for example, make a good estimate of parameters themselves and make use of the model instead of carrying out heavy mathematical computations such as Bayesian model estimation. It will be worthwhile to explore the feasibility of implementing CAT with a minimal calibration effort.

3. Existing CAT efforts: Case Studies

Case studies on CAT reveal what factors are considered in designing a CAT to best serve the assessment goals. In this section, we discuss the case studies of NAPLAN and GRE.

3.1. Case study: NAPLAN

NAPLAN (National Assessment Program - Literacy and Numeracy in Australia) is a test that recently (in 2017) shifted from Fixed Length pen and pencil test to an online CAT. The following addresses NAPLAN's implementation of CAT.

3.1.1. *Choosing CAT administrative model*

When NAPLAN made a transition to CAT, it was imperative that CAT covers the same breadth of domains, regardless of what test path a candidate takes. Consequently, ACARA (Australian Curriculum, Assessment, and Reporting Authority) chose a multistage adaptive test (MST) design as opposed to QAM (Question Adaptive Model). The benefit of multistage testing is that the test-makers have a greater control over the test structure, the test administration pathways, and the exposure of items. It also allows for students to review their responses and change their responses. [22]

3.1.2. *Implementation of MST NAPLAN*

MST is implemented as follows (here, a panel refers to a set of questions):

- 3-stage, with one panel at the first stage, two panels at the second stage, and three panels at the third stage.

In addition to this, NAPLAN embeds a unique feature to MST. When a candidate severely underperforms, they are taken to the easiest test panel in the third stage then are routed to the easier test panel in the second stage. This novel design was introduced to ensure that the underperforming students are encouraged to engage with the rest of the test by being exposed to the easiest test panel early. A simulation conducted to test the feasibility of this tailored design suggested that this tailored design considerably improves the measurement precision of the student achievement [22].

3.2. Case study: GRE

GRE (Graduate Record Examination in the United States) has been CAT since 1993 and is a famous case study for CAT. This case study summarises GRE's rationale for choosing MST as their test administrative model, as well as the implementation details.

3.2.1. *The choice between QAM and MST*

GRE's choice of administrative question model was QAM in 1993, but they switched to MST when they redesigned GRE in 2014. QAM-based CAT does not allow test takers to review their answers once each candidate answers a question. In addition, it bears a high test development cost because of the requirement that every test path should measure each

candidate equally well. MST-based CAT fixes both issues because 1) full evaluation of each testlet (or *panel* in their term) is possible before delivery (hence ensures each testlet meets various GRE test specifications), 2) limit can be imposed on testlet reuse and test item overlap between testlets thereby improving test security 3) test-takers can revisit questions before moving onto next testlet and change their answers. [13]

3.2.2. *Implementation of revised MST GRE*

Implementation of the current MST version of GRE consists of

- 2-stage, one panel at the first stage and three panels (easy, medium, difficult) at the second stage
- Each panel is 20 questions
- 2PL model is used for score estimation ('number correct')

After the routing panel in the first stage, each candidate is routed to the second stage. Scoring model uses 2PL instead of 3PL, because the redesign of GRE involved reducing multichoice questions, which consequently meant guessing was not as likely. The scoring mechanism was 'number correct' rather than 'pattern scoring' because 'number correct' tends to be more robust in covering suboptimal test conditions (e.g. anxiety, poor time management). [13]

4. Existing CAT efforts: frameworks

Some researchers have developed open CAT frameworks that provide implementations of many commonly used CAT routines to aid those intending to create a rapid prototype of CAT. We explore two frameworks – OSCATS and catR.

4.1. Framework 1: OSCATS (Open-Source Computer Adaptive Testing System)

OSCATS is a framework developed in C that provides ready-made implementations of a variety of algorithms that can be selected for each step in CAT [23]. OSCATS provides a framework that is modular (mix-and-match different features of CAT, such as models and algorithms), extensible, familiar (bindings in other languages, such as Python and Java), relieving researchers from concerns of implementations. The framework currently supports several IRT models (1PL, 2PL, and 3PL) and item selection algorithms (random, closest difficulty, MFI, and KL).

4.2. Framework 2: catR (R package for CAT)

catR is an R package that offers various options to customise CAT and simulate inputs and responses, making it a valuable tool for CAT researchers to simulate different CAT strategies and determine gains [24]. Unlike OSCATS, catR does not come with bindings in common languages such as Java and Python. However, it offers much more comprehensive, robust implementations of CAT models. catR makes it easy to simulate CAT and integrate it as a backend to the CAT administration platform. The framework does not support the calibration of the item bank and leaves it to packages such

as mirt [25]. catR library currently supports implementations of most IRT models, from models that go into simulations, data generation, item banking, item parameter generation, estimation, and scoring, making it an extremely versatile library for simulation and CAT platform implementation.

5. Discussion

5.1. CAT Model to use

NAPLAN and the revised GRE both use the MST over the QAM test administrative model to benefit from shorter length tests while minimising costs of maintaining a large pool of questions and ensuring that any subset of the question pool will fairly assess each candidate. Since we want to avoid large development costs, MST seems appropriate for our implementation. Given excellent recent case studies and analyses on the use of MST, it would be in our best interest to explore deeper into the use of MST. However, we will not restrict ourselves to MST, as there seems to be a gap in the use of QAM in the recent (post-2010) case studies, which our research could attempt to fill.

As for IRT models, we will most likely use the 2PL or 3PL parametric logistic model, as they have been frequently used across papers mentioned so far and are actively supported by catR and OSCATS frameworks. An interesting exploration could involve ERS model, which proposes iterative, 'on-the-fly' calibration.

5.2. Framework for CAT prototyping

Either framework (catR, OSCATS) will enable us to produce a working prototype for a CAT test. However, catR seems to have been more widely used and actively maintained than OSCATS; catR supports *concerto* - one of the largest open-source CAT implementations, and a recent major update was made in 2017. OSCATS has not had any commit since 2012. It has been reported that OSCATS has substantial deprecation of dependencies, rendering OSCATS unideal for our purpose of easily building a CAT.

5.3. What and who to test

Despite its general applicability, the development and implementation of CAT have primarily been in the realms of education assessment. GRE uses CAT not just for multichoice questions but for questions of various formats, while NAPLAN uses it solely for multichoice questions. The consensus seems to be that multichoice questions are easy to develop due to their dichotomous nature. Most research and implementations have focused on primary students, and it is of our interest to test it on an adolescent population - Junior / Senior high school students.

6. Conclusion

Accurate and efficient assessment of students' ability is important, yet most examinations center around a single testing paradigm: the Fixed Length Test. This literature explored CAT, an alternative testing paradigm with a robust ground in the measurement theory. With various high-stake examinations such as NAPLAN and GRE making a transition to CAT, CAT is no longer an experimental testing paradigm but a valid alternative to the Fixed Length Tests. It seems timely that we contribute to the research body by demonstrating practical implementation of CAT based on solid theory and frameworks, which would accelerate CAT adoption to wider populations.

7. Statement of Research Intent

We want to learn more about the feasibility of implementing a CAT system using tools and software accessible to the broader population and using a relatively straightforward process. There is a gap in research showing how CAT can be implemented at small scale targeting high school students. There is also a gap in the literature showing how one might architect and implement a software architecture for CAT. We aim to create a prototype based on multiple models presented in CAT literature and explore the feasibility of employing CAT for testing purposes in the high school students' demographics. The prototype implementation and testing will involve the following:

- We will use a CAT framework for our implementation to enable rapid prototyping. We are currently inclined to use catR as it is widely used across research and is actively maintained.
- Depending on the feasibility and time, we will consider both the MST and QAM test administrative models and use 3PL due to its generality – though we could also consider the 1PL, 2PL, and Elo model.
- We aim to test our prototype on high school students, as most research has focused on primary students (NAPLAN) and tertiary students (GRE).

7.1.1. Timeline

April: Complete Literature review (13th), Start CAT implementation, Ethics approval application

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

June: Complete an MVP on one branch of CAT (either MST or QAM), conduct testing on the SOFTENG cohort

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

August/September: Conduct a pilot study on high school students. Iterate upon MVP and document findings and results.

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

8. References

- [1] 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.
- [2] R. K. Hambleton and L. L. Cook, "Latent trait models and their use in the analysis of educational test data," *Journal of educational measurement*, pp. 75-96, 1977.
- [3] S. E. Embretson, "Multicomponent latent trait models for test design," *Test design: Developments in psychology and psychometrics*, pp. 195-218, 1985.
- [4] W. J. Linden, W. J. van der Linden, and C. A. Glas, *Computerized adaptive testing: Theory and practice*. Springer, 2000.
- [5] D. O. Segall, "Computerized adaptive testing," *Encyclopedia of social measurement*, vol. 1, pp. 429-438, 2005.
- [6] R. K. Hambleton and J. N. Zaal, *Advances in educational and psychological testing: Theory and applications*. Springer Science & Business Media, 2013.
- [7] 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.
- [8] J. M. Linacre, "Computer-adaptive testing: A methodology whose time has come," MESA memorandum, 2000.
- [9] R. Gershon, "Test anxiety and item order: new concerns for item response theory," in *Objective Measurement: Theory into Practice Vol 1*. Ablex, 1992, pp. 175-194.
- [10] D. Magis and G. Raïche, "Random generation of response patterns under computerized adaptive testing with the R package catR," *Journal of Statistical Software*, vol. 48, pp. 1-31, 2012.
- [11] D. Magis and L. Ippel, "Ability estimation (dichotomous and polytomous models)."
- [12] J. B. Bjorner, C.-H. Chang, D. Thissen, and B. B. Reeve, "Developing tailored instruments: item banking and computerized adaptive assessment," *Quality of Life Research*, vol. 16, no. 1, pp. 95-108, 2007.
- [13] C. Wendler, B. Bridgeman, and C. Ezzo, "The Research Foundation for the GRE revised General Test: A compendium of studies," *Educational Testing Service: Princeton, NJ, USA*, 2014.
- [14] L. L. Davis, "NAPLAN ONLINE RESEARCH AND DEVELOPMENT."
- [15] E. B. Andersen, "Asymptotic properties of conditional maximum-likelihood estimators," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 32, no. 2, pp. 283-301, 1970.
- [16] R. D. Bock and M. Aitkin, "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm," *Psychometrika*, vol. 46, no. 4, pp. 443-459, 1981.
- [17] H. Swaminathan and J. A. Gifford, "Bayesian estimation in the two-parameter logistic model," *Psychometrika*, vol. 50, no. 3, pp. 349-364, 1985.
- [18] F. M. Lord, *Applications of item response theory to practical testing problems*. Routledge, 2012.
- [19] S. Klinkenberg, M. Straatemeier, and H. L. 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.
- [20] M. Antal, "On the use of elo rating for adaptive assessment," *Studia Universitatis Babes-Bolyai, Informatica*, vol. 58, no. 1, pp. 29-41, 2013.
- [21] T. Yao and J. M. Linacre, "CAT with a Poorly Calibrated Item Bank," *Rasch Measurement Transactions*, 1991.
- [22] L. Goran, J.-A. Justus, and S. Rabinowitz, "The Tailored test design study 2013: Summary research report," *National Assessment and Surveys Online Program, ACARA*, 2013.
- [23] M. J. Culbertson, "Rapid CAT Prototyping with OSCATS: The Open-Source Computerized Adaptive Testing System," 2011.
- [24] D. Magis and J. R. Barrada, "Computerized adaptive testing with R: Recent updates of the package catR," *Journal of Statistical Software*, vol. 76, pp. 1-19, 2017.
- [25] R. P. Chalmers, "mirt: A multidimensional item response theory package for the R environment," *Journal of statistical Software*, vol. 48, pp. 1-29, 2012.