Paper review summaries

In this doc, write a short (1 - 2 paragraph) summary of each paper you review. Perhaps organize the papers by theme / topic.

[**Computerized Adaptive Testing: Theory, Applications, and Standards (Chapter 12)**](https://link.springer.com/content/pdf/10.1007%2F978-94-009-2195-5.pdf)

This paper discusses the background behind computerized adaptive testing (CAT) and the research that underpins computerized adaptive testing. In addition to the above, the chapter introduces item response theory and how it is used to create regular as well as adaptive tests. A computer adaptive testing system is composed of 3 parts:

* It must predict how an examinee will respond to test questions not yet administered based on the examinee’s previous responses.
* It must use the above information to decide on the test question to be administered next.
* It must assign a numerical score to the participant at the end of the test which represents their ability.

The overall goal of adaptive testing is to improve measurement precision while requiring less test items than a traditional test.

[**The GRE Computer Adaptive Test: Operational Issues**](https://link.springer.com/content/pdf/10.1007%2F0-306-47531-6_4.pdf?fbclid=IwAR37lwqfTVkaKCyrgWJuOZBFvgGKnqY24uKlH1ukDyucbMHwABOVHpZCBvw)

This paper discusses the Graduate Record Examinations (GRE) and the development of the GRE CAT. The paper also dives deeper into the three main issues that have arisen from the GRE CAT:

* Item overlap
* Maintaining item pool quality over time
* Scoring incomplete tests.

[**Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation**](https://www.sciencedirect.com/science/article/pii/S0360131511000418?fbclid=IwAR2GyqiC24yKcswEo-NVG0z14Oin0yAnf7W0hJEXDOIHoAub06N0NaaQxPU&via%3Dihub#bbib4)

The authors use an item response model based on the Elo rating system to assign Math problems to students and to measure their ability. They created an experiment where users and problems were given Elo ratings. Users and test items begin with a baseline rating and their ratings (both the students’ and the items’) self-adjust based on whether the student answers the questions correctly. The authors also incorporated a scoring rule for speed and accuracy where the faster a user answers a question, the larger the effect on their rating.

[**Applications of the Elo Rating System in Adaptive Educational Systems**](https://www.fi.muni.cz/~xpelanek/publications/CAE-elo.pdf)

The paper suggests that the Elo rating system should mainly be used for adaptive practice and low stakes testing. They found that while the Elo model can provide reasonable estimates for adaptive testing, the system does not provide statistical guarantees on estimated skills whereas well calibrated IRT models do. As such, the key use case for the Elo model is when an adaptive system is needed to be built quickly and cheaply as it does not require expert judgment to evaluate item difficulty, instead relying on data provided during the testing process.

[**How to build a Computerized Adaptive Test with free software and pedagogical relevance?**](https://www.researchgate.net/publication/326803834_How_to_build_a_Computerized_Adaptive_Test_with_free_software_and_pedagogical_relevance)

The paper details the process of creating a computer adaptive test utilizing free software packages available for R. Their aim was to create a test that has fewer items and more accuracy than a similar but non-adaptive test.

They tested five item selection methods (random, Maximum Fisher Information, Urry rule, Maximum Information with stratification, Progressive method, Proportional method) based on IRT described by the two-parameter logistic function. They also tested four proficiency estimation methods (maximum likelihood, weighted likelihood, modal bayesian estimator, expected a posteriori). However, they found that there were no significant differences between the item selection (excluding random) and proficiency selection methods.

They also had to develop a termination criteria for the test based upon 3 criterion:

* Absolute limits (at least 8 items, at most 20 items)
* Maximum error
* In their case specifically, when the proficiency and its confidence interval are fully contained inside a single proficiency level, among the five levels defined for Provinha Brasil.

[**A Calibrated Item Bank for Computerized Adaptive Testing in Measuring Science TIMSS Performance**](https://files.eric.ed.gov/fulltext/EJ1272239.pdf)

[**Computer-Adaptive Testing: A Methodology Whose Time Has Come**](https://www.cehd.umn.edu/EdPsych/C-Bas-R/Docs/Linacre2000_CAT.pdf)

The paper looks at CAT as an emerging testing paradigm. This paper was written in 2000.

* Multistage CAT is introduced as a ‘crude approach’ that usefully produces the same results as more sophisticated CAT techniques.
* Dichotomous vs polytomous items - writing and developing defensible scoring schemes for polytomous items can be difficult, but would provide more information and require fewer administered items.
* Asserts importance of unidimensionality of construct for use in CAT. An item that assesses more than 1 construct (eg literacy AND numeracy) will confound CAT, and prevent stochastic ordering. Psychometric model necessary to establish and maintain consistent stochastic ordering is the Rasch model.
* log(probability of success on item i/probability of success on item i) = test taker ability - difficulty of item. Test taker ability is the same as difficulty of item when probability of success is 50% (log(1) = 0)
* Provides source code for UCAT Implementation written in BASIC. UCAT is CAT that updates difficulty assigned to the items, and re-evaluates candidates ability based on the updated difficulty, thereby reducing the effect that taking test at different times has on accuracy of estimation.
* **Introduces simple implementation of CAT**
* Discusses stopping rules/design of CAT, and advantages of CAT as well as criticisms against CAT
* Quite comprehensive.
* Also details views against CAT at the time computer wasn’t prevalent

[**Computer adaptive testing, big data and algorithmic approaches to education**](https://sci-hub.hkvisa.net/10.1080/01425692.2016.1158640)

Critically discusses CAT in sociological / educational lens, focusing on the NAPLAN test (which will be implemented as an adaptive testing) as a case study. Keywords are datafication, big data and data infrastructure.

* It asserts CAT’s benefit beyond testing. It will help teachers to design learning paths/curriculums for individual students, continue tracking each student.
* By integrating more factors (such as time taken to answer a question), the algorithm can be designed to make sense of the data patterns, which better represents the subject through data.
* Discusses multistage adaptive testing design of NAPLAN as a method to shift the tests quicker and more accurate

[**Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subjective Test Experience**](https://www.researchgate.net/publication/316292082_Computer-Adaptive_Testing_Implications_for_Students%27_Achievement_Motivation_Engagement_and_Subjective_Test_Experience)

Discusses the implications of the results from a very recent implementation & pilot run of NAPLAN tests (national literacy & numeracy test) in 2016. In particular, it shows:

* Ability measurement precision is greater than that of fixed length tests
* CAT (implemented as multistage) results in higher test-relevant motivation & engagement (measured through a comprehensive model outlined in the paper)

It also:

* Discusses gap in the literature regarding implications of CAT in motivation & engagement (this present paper acts as a data to support positive argument for CAT on student’s motivation & engagement.
* Discusses implementation of MAT in practical details

[**Rapid CAT prototyping with OSCATS**](https://www.ideals.illinois.edu/handle/2142/27706)

Program written in C that supports mix-and-match of algorithms used for each stage of CAT (administering, ability estimation etc).

* Provides easy interface in Java/Python
* Identifies that not lots of people want to write CAT implementation for their research, hence provides codes that even less technical people can interface with.
* Provides ‘lego blocks’ for us to play around with to research how each different combinations of algorithms at each stage of CAT affects results.