An Introduction to Computerized Adaptive Testing

Nathan A. Thompson, Ph.D.
Adjunct Faculty, University of Cincinnati
Vice President, Assessment Systems Corporation

Assessment Systems Corporation
www.assess.com
Welcome!

- CAT: tests that adapt to each examinee
- The purpose of this webinar is to provide an introduction to:
  - Item Response Theory as used in CAT
  - CAT algorithms
  - Implementing CAT
Welcome!

- There will be four parts:
  - Intro to item response theory (IRT)
  - Basic principles of CAT (the five components)
  - Benefits of CAT
  - Implementing CAT
Part 1

Introduction to item response theory
What is IRT?

- There are two psychometric theories: Classical and IRT
- IRT offers distinct advantages, the most important with regards to CAT is that *items and examinees are on the same scale*
What is IRT?

- IRT assumes that we can specify a mathematical function that models the probability of getting an item correct
- The item response function
- The following presents a figure from a classical analysis
Classical item statistics

- The line for the correct answer (blue) should go up while the distractor lines go down
- The line for the correct answer is usually of primary importance
Classical item statistics

- What if we had 10 groups?
Classical item statistics

- The general idea of IRT is to find a mathematical model for the line of the correct response (previous slide)
- A special form of regression - need a curve rather than a line
The item response function

- Reflects the probability of a given response as a function of the latent trait
The item response function

- The x-axis is the standard z score you learned in statistics classes
- IRFs can slide left or right, which defines item difficulty
  - Left is easy
  - Right is difficult
The item response function

• The location of an item is where the middle of the IRF is with respect to the x-axis

• So therefore, both items and examinees are on the z scale
Part 2

Basic principles of CAT
(The Five Components)
What is CAT?

- A Computerized Adaptive Test (CAT) is a test administered by computer that dynamically adjusts itself to the trait level of each examinee as the test is being administered.
CAT Components

- 1. Calibrated item bank
- 2. Starting rule
- 3. Item selection rule
- 4. Scoring rule
- 5. Stopping rule

- Given 1 and 2, we repeat 3 and 4 until 5 is satisfied
- All CAT follows this basic format - we just modify the details for whatever testing situation we have
CAT Components

1. Calibrated item bank
2. Starting rule
3. Item selection rule
4. Scoring rule
5. Stopping rule

Given 1 and 2, we repeat 3 and 4 until 5 is satisfied.

All CAT follows this basic format - we just modify the details for whatever testing situation we have.

Algorithms inside your testing engine
1. Calibrated item bank

- While it is possible to design CATs with classical test theory (Frick, 1992), IRT is more appropriate because it puts items and examinees on the same scale.
- Therefore, the items need to be calibrated with IRT.
1. Calibrated item bank

- CAT algorithms work with any IRT model
- The choice of the model depends on characteristics of the test and your goals
1. Calibrated item bank

- The bank for the CAT should be constructed with the purposes of the test in mind
- Flat or peaked?
- If peaked, where?
1. Calibrated item bank
2. Starting rule

1. Can start everyone with the same theta estimate (e.g., theta = 0.0)
   - Everyone gets the same first item
   - Could be an exposure problem in a high stakes test

2. Assign a random theta estimate within an interval
   - E.g., between theta = -0.5 and +0.5
   - Improves exposure levels and has little effect on a properly implemented CAT
2. Starting rule

3. Use prior information available for a given examinee
   - Subjective evaluations, e.g., below average, above average
   - Theta estimates from tests previously administered in the same or a prior test session
   - Theta estimate from same test administered at a previous time
3. Item selection rule

- Items are selected to maximize information (how good an item is)
- Information is a function of the slope of the IRF
- An item provides more information where there is more slope
The item response function

![Graph of the item response function](image-url)
3. Item selection

- Example 5 items:
3. Item selection

- Also, there are usually practical constraints in item selection
  - Item exposure
  - Content area (domain)
  - Cognitive level
  - Etc.
4. Scoring rule

- IRT scores students with a form of *maximum likelihood estimation*
- Basically, IRFs are multiplied
4. Scoring rule

- IRT utilizes the IRFs in scoring examinees
- It is not done with number-correct scores
- If an examinee gets a question right, they “get” the item’s IRF
- If they get the question wrong, they “get” the (1-IRF)
- These curves are multiplied for all items to get a final curve called the \textit{likelihood function}
4. Scoring rule

- Here’s an example IRF
4. Scoring rule

- A “1-IRF”
4. Scoring rule

- We multiply those to get a curve like this...
4. Scoring rule

- Since we find the highest point of the likelihood function, it is called *maximum likelihood estimation*.
- There are also two Bayesian methods (MAP, EAP) and weighted MLE.
5. Stopping rule

- Depends primarily on purpose of the test: point estimation or classification?
  - Point estimation: we want an accurate score for each student
  - Classification: we do NOT need an accurate score, just a classification into pass/fail etc.
5. Stopping rule

- Point estimation methods involve actual scores, and stop when we have zeroed in enough.
- Classification methods check after every item to see if we can make a classification within a certain degree of accuracy.
5. Stopping rule

- For educational tests, this is usually point estimation
- Common stopping rule: stop the test when examinee reaches a certain level of error of measurement
- Means all examinees are scored with equal precision
5. Stopping rule

- Either type of CAT can be designed with a fixed number of items.
- But this is a bad idea from a psychometric perspective.
- *Variable-length testing* is much more efficient.
The big picture

1. Define Item Bank
2. Specify test-level parameters: Allow review, min/max length, etc.
3. Start test
4. Select Item
   - Things to consider: item information, has the item already been used, content domain, randomizing
   - Administer item to examinee
   - Score Item Correct/Incorrect
   - Evaluate Stopping Rule
   - Not Satisfied
     - Things to consider: psychometric criterion (e.g., SEM), min/max length, content distribution
5. Satisfied
   - Conclude Test and Print Score Report

www.assess.com
The big picture

- Item by item graph:

\[ \text{Cut} = 0.55 \]
Part 3

Benefits of CAT
Benefits of CAT

- **Efficiency:** CATs are more efficient than conventional tests: they generally reduce test length by 50% or more (Weiss & Kingsbury, 1984)
  - See research for examples
  - Simulations can estimate for you
  - Even more efficient for classification CATs - average test length in single digits
Benefits of CAT

- Control of measurement precision: A properly designed CAT can measure or classify all examinees with the same degree of precision.
Benefits of CAT

- Equal precision is impossible with conventional tests
- So the question is: is it more fair that all students see the same items, or that they are measured with the same accuracy?
Benefits of CAT

• Added security
  ◦ If everyone receives a standard test with the same 50 items, the items will become well known
  ◦ This effect is decreased when everyone receives a different set of items
  ◦ We can also make multiple forms, but is that better than CAT? Case by case...
Benefits of CAT

• Immediate score reporting
  ◦ P&P testing requires the question papers to come back and be scored
  ◦ If immediate feedback for students is desirable, then P&P testing is not an option
Disadvantages of CAT

• Public relations
  ◦ Need to explain to examinees/parents why certain things can happen, like failing after only 10 questions, or passing with a 50% correct score
Disadvantages of CAT

• Sophistication
  ◦ Requires specially designed software
  ◦ Requires a lot of expertise and effort – so often out of reach for small testing programs
  ◦ Some say too expensive, but really: ~$3000 for an administrator and testing center?
  ◦ The major cost in test development is the same for CAT and P&P: item development
Disadvantages of CAT

• Item Exposure
  ◦ Some items will be used far more often than others, which needs to be addressed
  ◦ Plenty of methods have been suggested, but they decrease the efficiency of the CAT process
Part 4
Implementing CAT
So, you want a CAT?

- Well, you’ve decided to use CAT, and you’ve built a nice item bank, what next?
- You need a test development system and delivery engine that does CAT
- I’ll show you what it looks like in FastTEST Pro
- Late this year there will be a FastTEST Web
FastTEST Pro

- Common source of confusion:
  - FastTEST is the item banker and test development system
  - FastTEST Pro is that plus the delivery engine
FastTEST Pro

- Common source of confusion:
  - FastTEST is the item banker and test development system
  - FastTEST Pro is that plus the delivery engine
FastTEST Pro:  1. Bank items

Each year one college football player receives a trophy rewarding his outstanding play during his college football career. The name of this award is:

1. Stanley Cup
2. Heisman Trophy
3. Golden Globe
4. CMVP Award
FastTEST Pro

2. Design pool for your CAT
FastTEST Pro

3. Define CAT modules
FastTEST Pro

3. Define CAT modules

- Select items based on current theta estimate...
- By maximum value of item information at current theta estimate throughout the test.
- Randomly from the 0 items with maximum item information, for the first 0 items in the test.
FastTEST Pro

- 3. Define CAT modules

### Terminate the test after...

- **exactly** 0 items have been administered
- the standard error of the theta estimate is less than, or equal to **0.000**
- the theta estimate +/- **0.000** standard errors, is above or below a cutoff theta value of **0.000**
- administer a minimum of **0** items prior to using this calculation (optional)
- a maximum of **1** items have been administered (optional)
FastTEST Pro

- Now I’ll show a real CAT with FastTEST Pro
- You can download and use free for 30 days at
Thank you!

- Questions?
- Any questions in the future:
  - nthompson@assess.com
Resources

- PARE online: [http://pareonline.net/](http://pareonline.net/) - see Vol 12, #1