LETTING THE CAT OUT OF THE BAG: HOW CAN WE MAKE MORE ASSESSMENTS ADAPTIVE?

NATHAN THOMPSON, PHD

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Welcome

1. What is CAT? + Background on psychometrics
2. 5 step model to develop CAT
3. Software and other hurdles
4. CATs in the wild

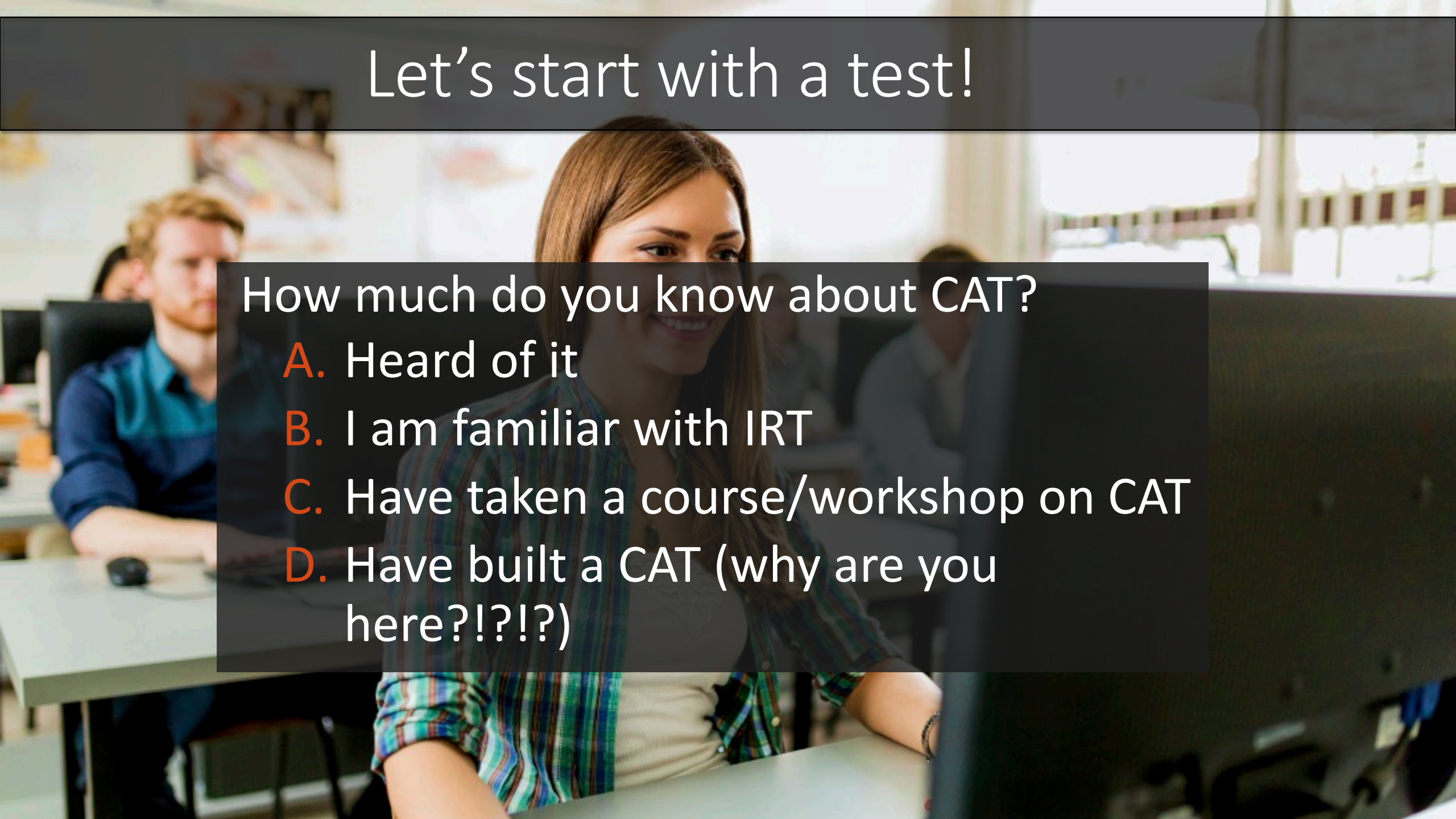
ASSESSMENT SYSTEMS
FOR GOOD2MEASURE™

Welcome!

- ✓VP of Psychometrics at ASC
- ✓Twin Cities, MN
- ✓PhD in Psychometrics from U of MN
- ✓My career: getting organizations out of the 1950s and into modern assessment with real psychometrics

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Let’s start with a test!



How much do you know about CAT?

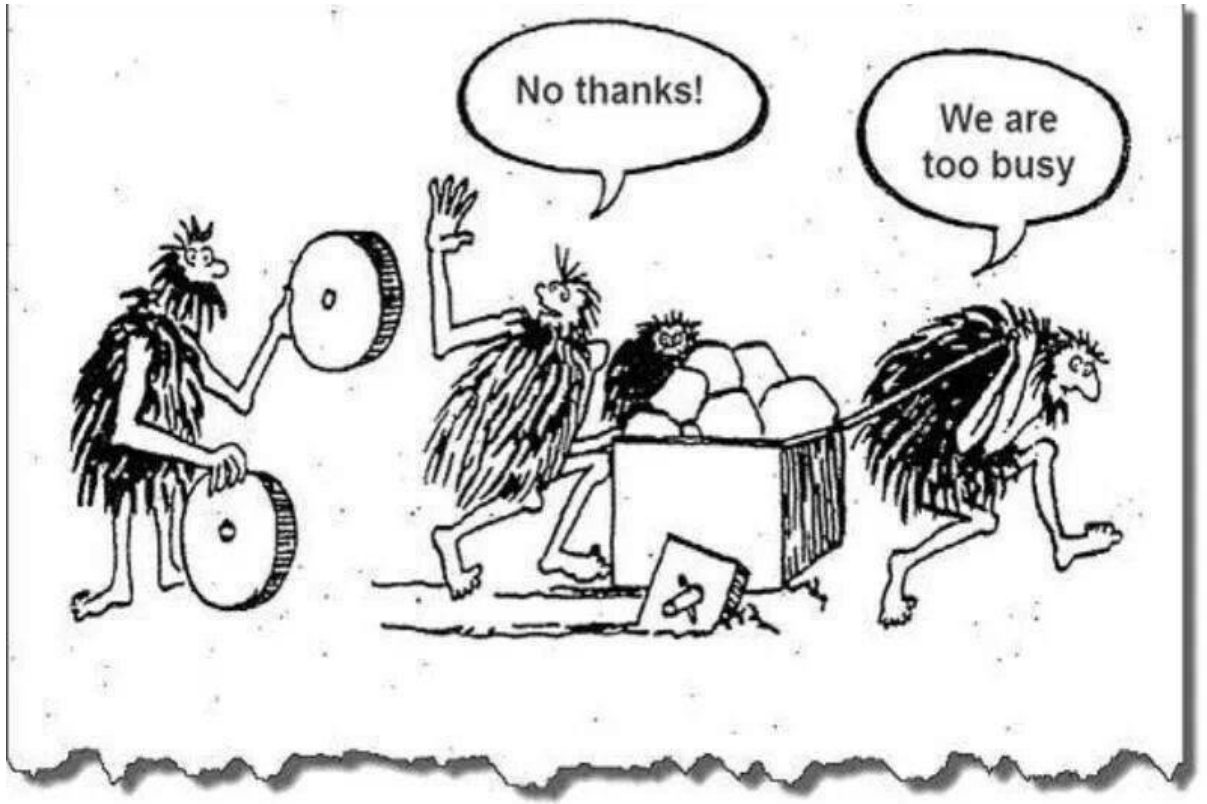
A. Heard of it
B. I am familiar with IRT
C. Have taken a course/workshop on CAT
D. Have built a CAT (why are you here?!?!?)

What is CAT?

First: What is Computerized Adaptive Testing (CAT)?

A test that adapts the length and/or difficulty of the test for each examinee *individually*

More than 40 years old and still underutilized!



What is CAT?

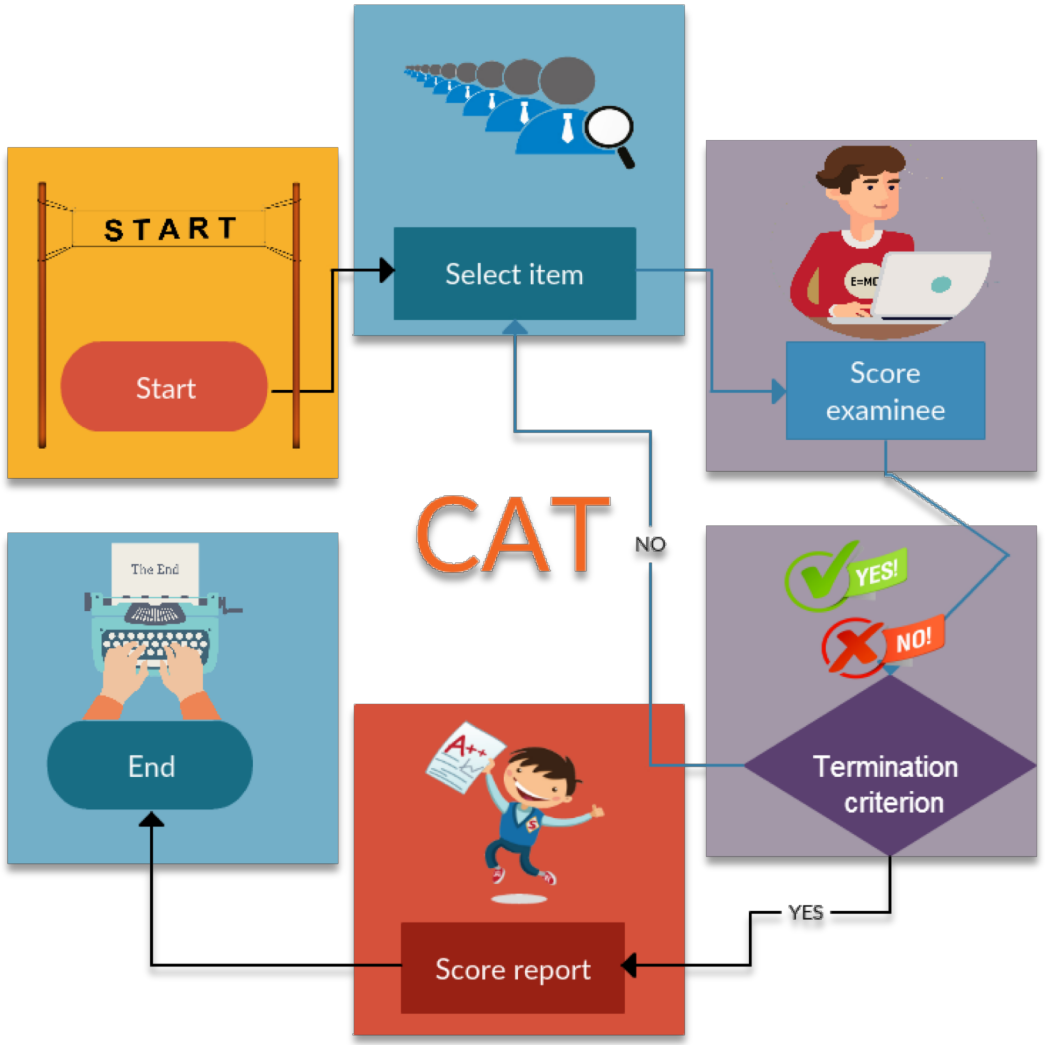
I’m talking today about how CATs are built, so that you can build your own if you want.

If you buy them off the shelf, you don’t need all this, but you’ll learn how your vendor is developing their CATs.

build or buy?

What is CAT?

What is CAT?



ASSESSMENT SYSTEMS
FOR GOOD&MEASURE™

Why? Benefits of CAT

- Shorter tests with no loss of precision
  - Can substantially reduce costs!
- Improves security
- Increases fairness
- Enables equiprecision
- Reduces fatigue
- Increases engagement/motivation
- Frequent retesting/confirmation

Benefits: efficiency

Consider this scenario

- Paying $25/hour for seat time (or away from work costs)
- Current test is 2 hours; CAT cuts to 1

Annual examinees	Seat cost savings
1,000	$25,000
10,000	$250,000
100,000	$2,500,000

Why not cat?

Economics

- Volume too low to realize the savings
- Lack of staff: CAT experts are super rare
- Heavy costs to do it well

Why not cat?

Test characteristics/situation

- Large sample sizes needed for IRT (100 to 1000 depending on model)
- Item bank must be robust
- Items must be objective
- Items must be scored in real time

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Why not 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
- You'll need resources like NCLEX

Why not cat?

Item exposure can be a problem

- But better than fixed forms!

Need for strict content coverage or other controls

- Hence move by some agencies to multistage testing (MST) and massively parallel forms

CAT Components

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

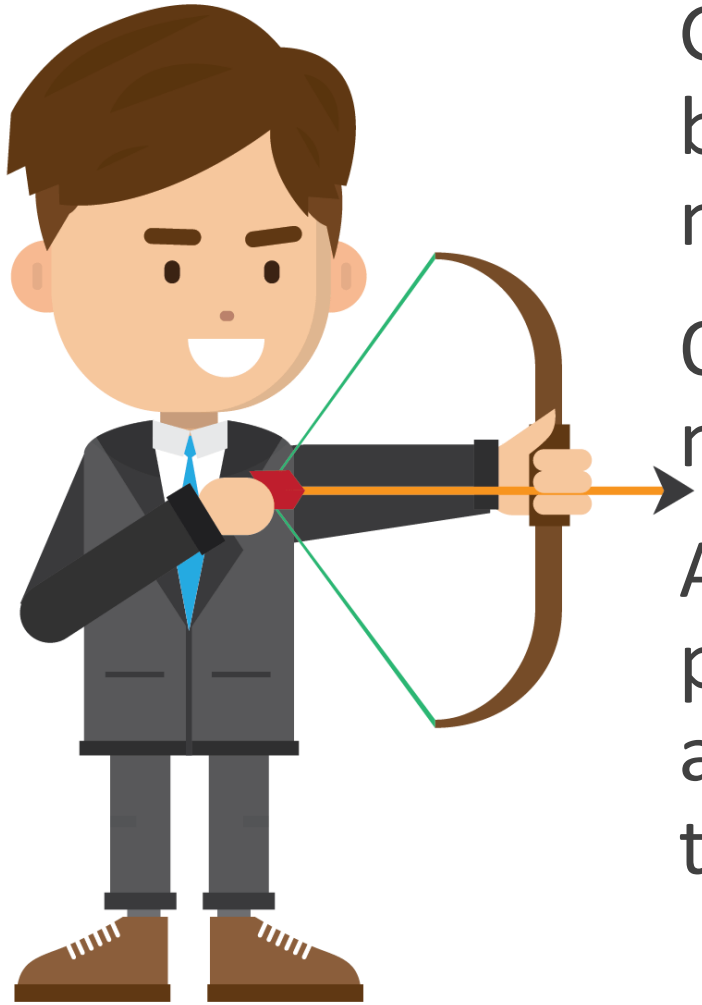
Test development side

Algorithms inside your testing engine

We must provide validity documentation on each

Weiss & Kingsbury, 1984

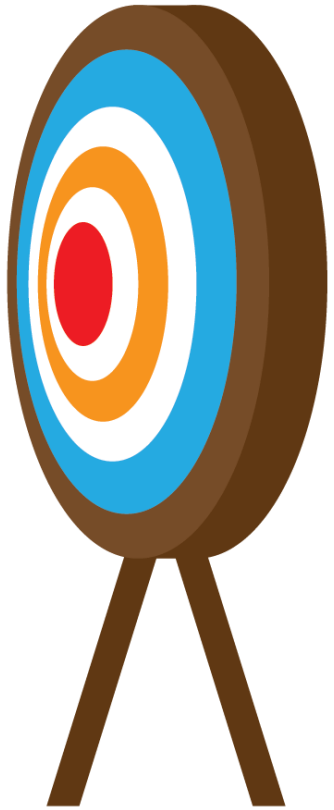
1. Calibrated item bank



Goal: develop an item bank that meets the needs of your CAT

Q: What are the needs?

A: Defined by the purpose of your CAT and item response theory

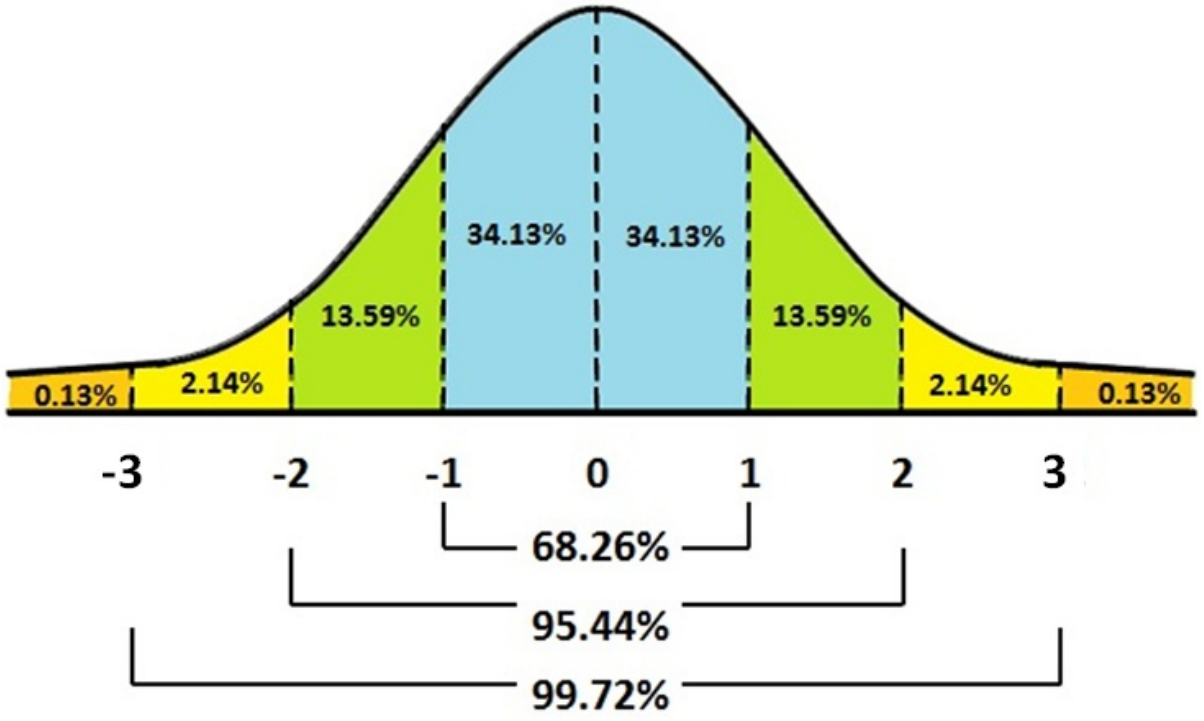


ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

1. Calibrated item bank

IRT: placing people on the normal bell curve

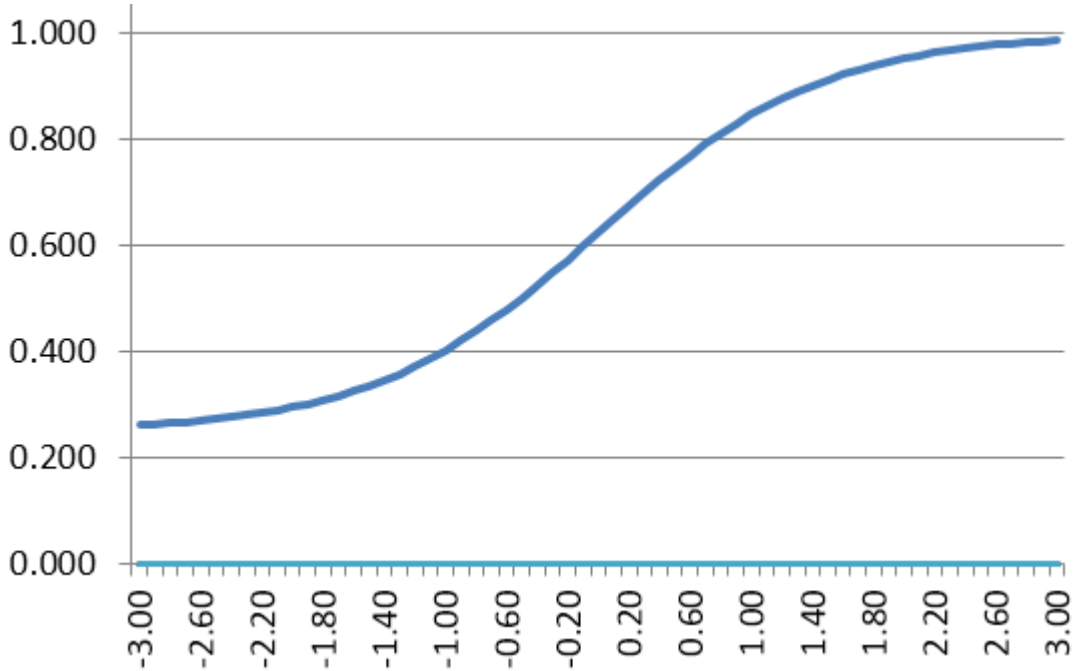
Then also placing items on the same x-axis



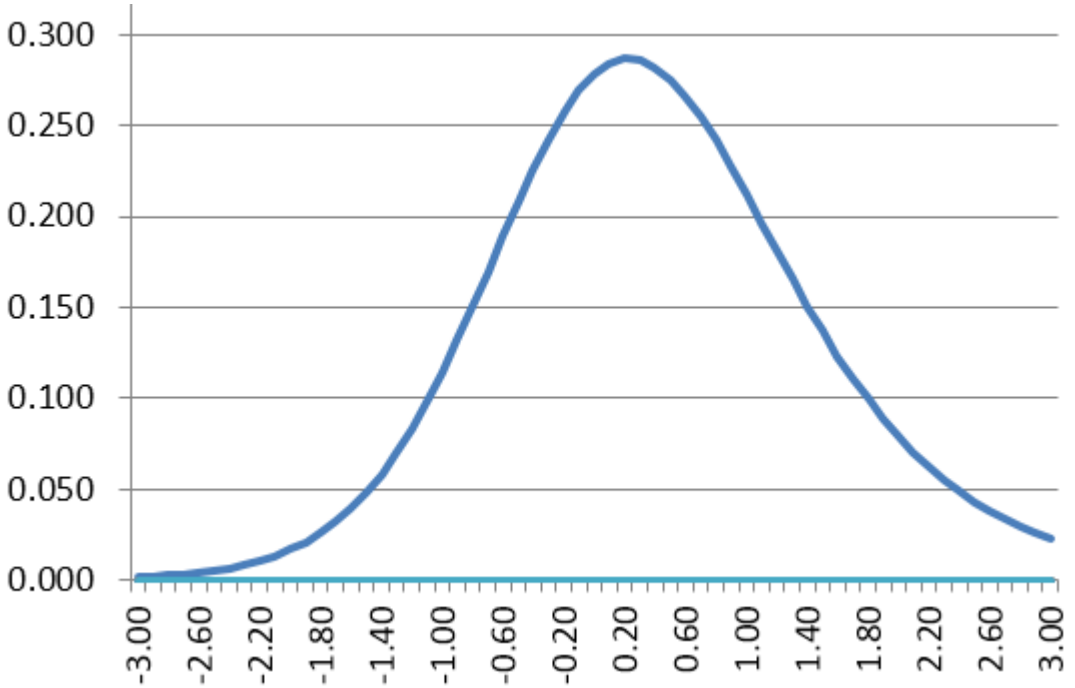
ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

1. Calibrated item bank

Items are placed on the scale by estimating an *item response function* (IRF)



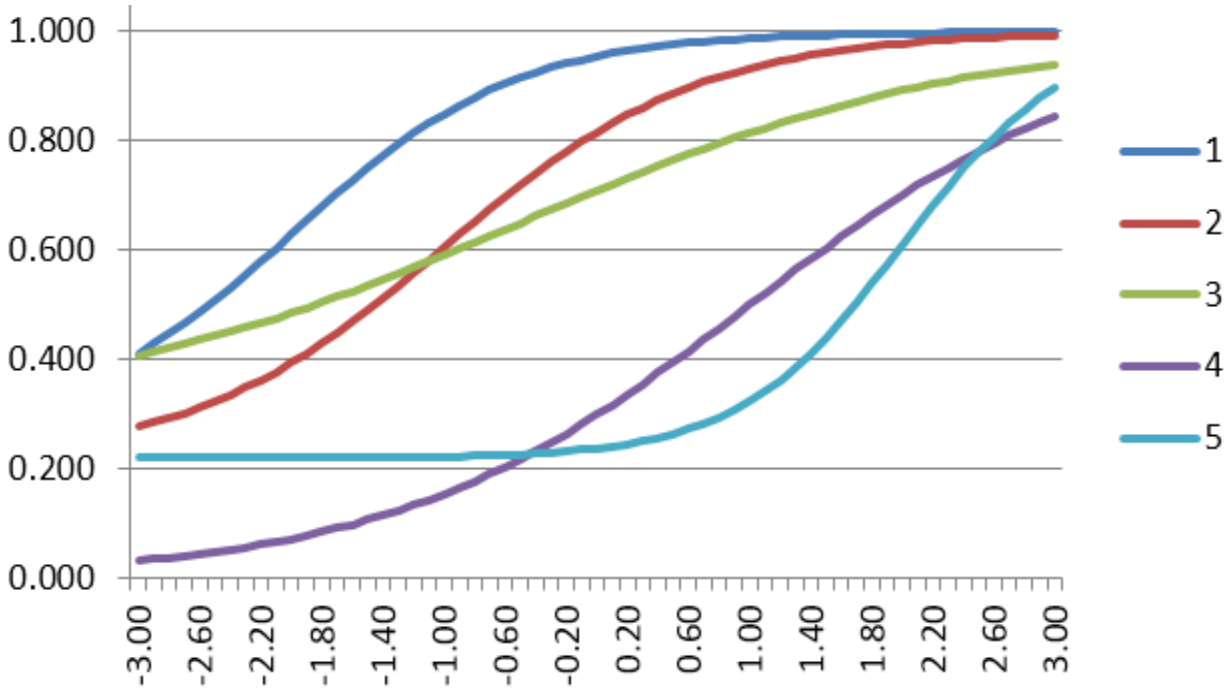
The IRF can be converted to an *item information function* (IIF)

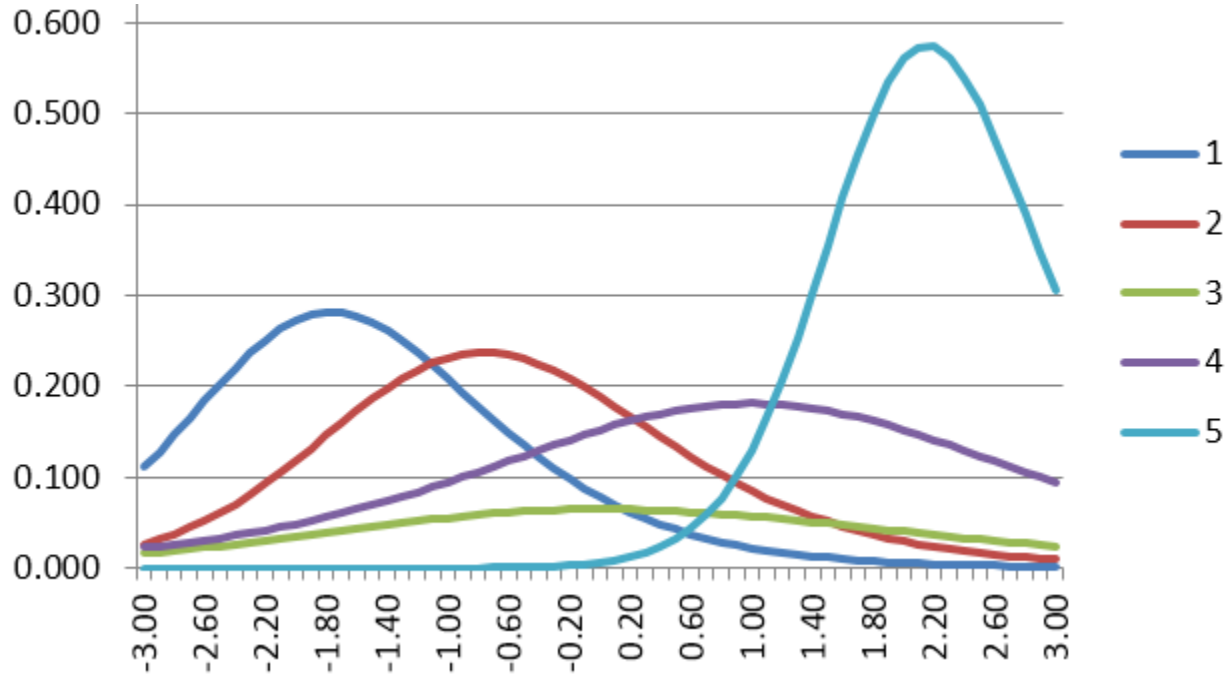


ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

1. Calibrated item bank

You'll then have a bank of IRFs and IIFs





ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

2. Starting rule

θ = 0.0 (average)

- Security flaw: everyone gets the same first item

Random within a range

- -0.5 t 0.5; improves security

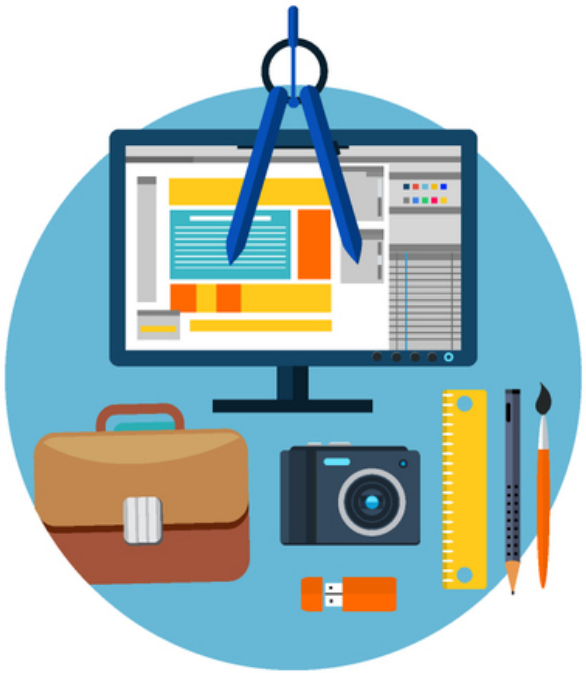
Prediction from external data

- For example, college grades

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

3. Item selection rule

- Select the item that tells us the most about the person: *maximum information*
- Think of it as finding the item with the closest difficulty and highest discrimination
- Really it is the derivative of the IRF



ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

3. Item selection

Also, there are usually practical constraints in item selection

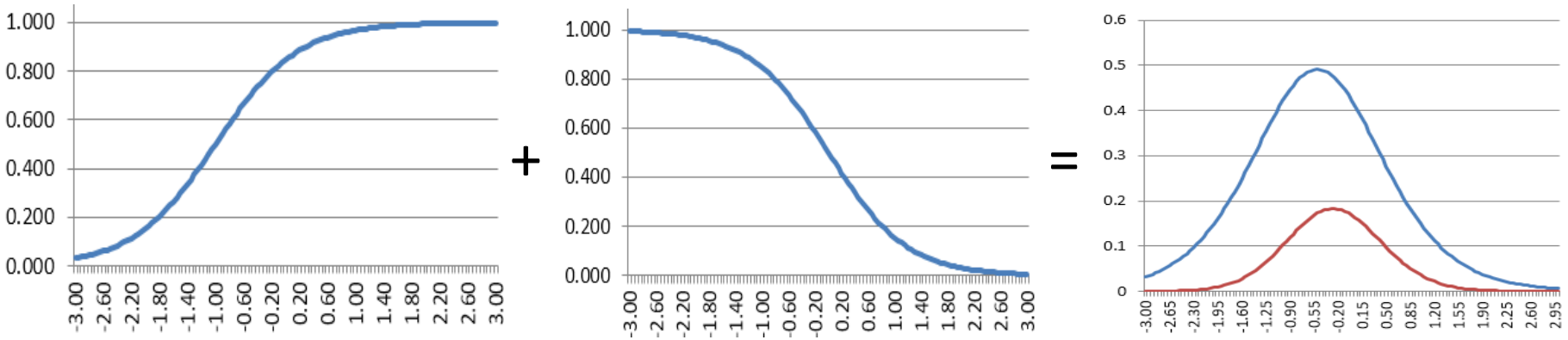
- Item exposure
- Content area (domain)
- Cognitive level
- Etc.

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

4. Scoring rule

Score with items so far to find theta (z)

Typically, MLE is used to score examinees after each item (often with temp Bayesian)



ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

5. Stopping rule

Depends primarily on purpose of the test: *point estimation* or *classification?*

Point Estimation	• Stop when accurate score for examinee (low standard error)
Classification	• We do NOT need an accurate score, just a classification into pass/fail etc.

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

5. Stopping rule

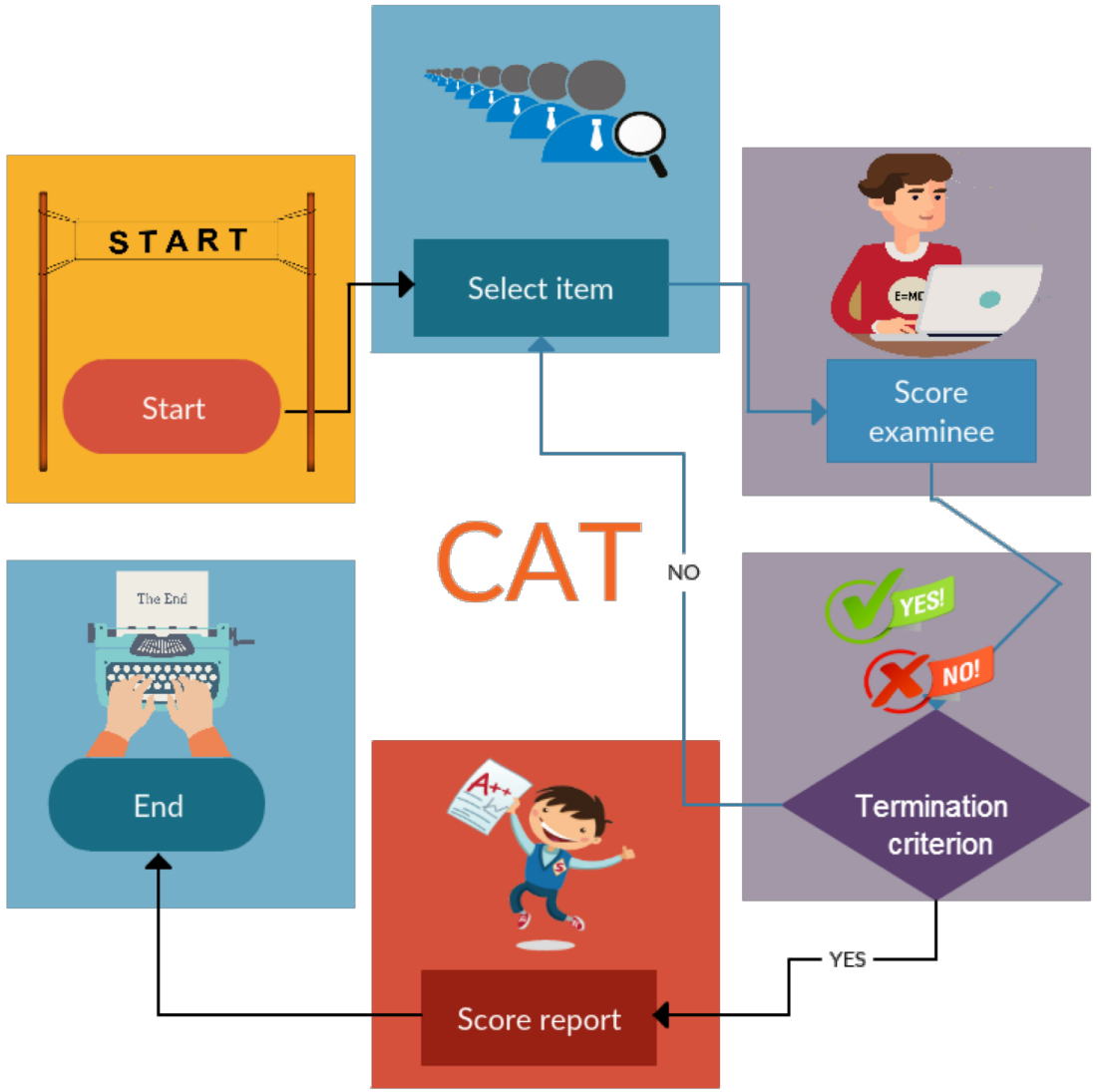
CAT can be *fixed-length* or *variable-length*

Fixed-length is a bad idea from a psychometric perspective but can greatly enhance *perceived fairness*

Variable-length is much more efficient and what enables the 50% reduction

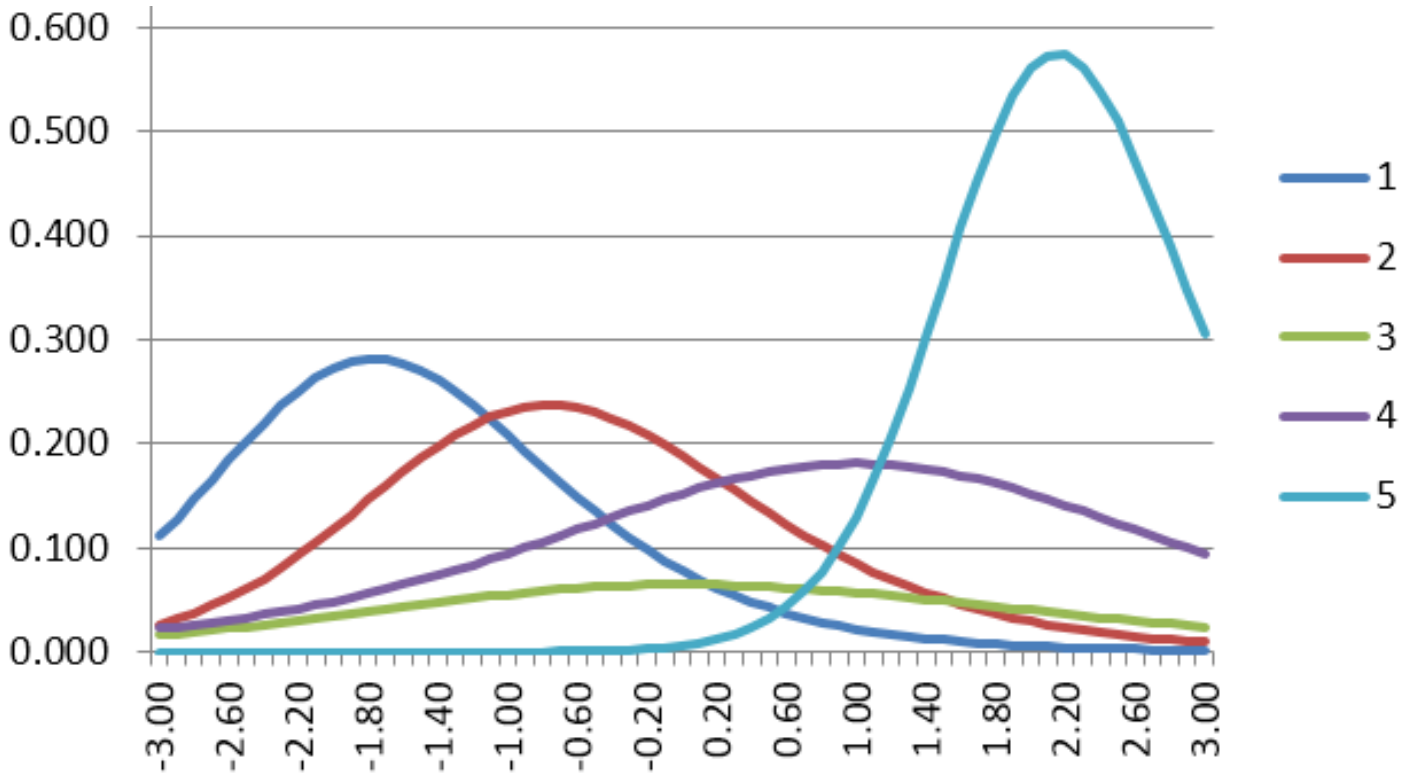
ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

The big picture



ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

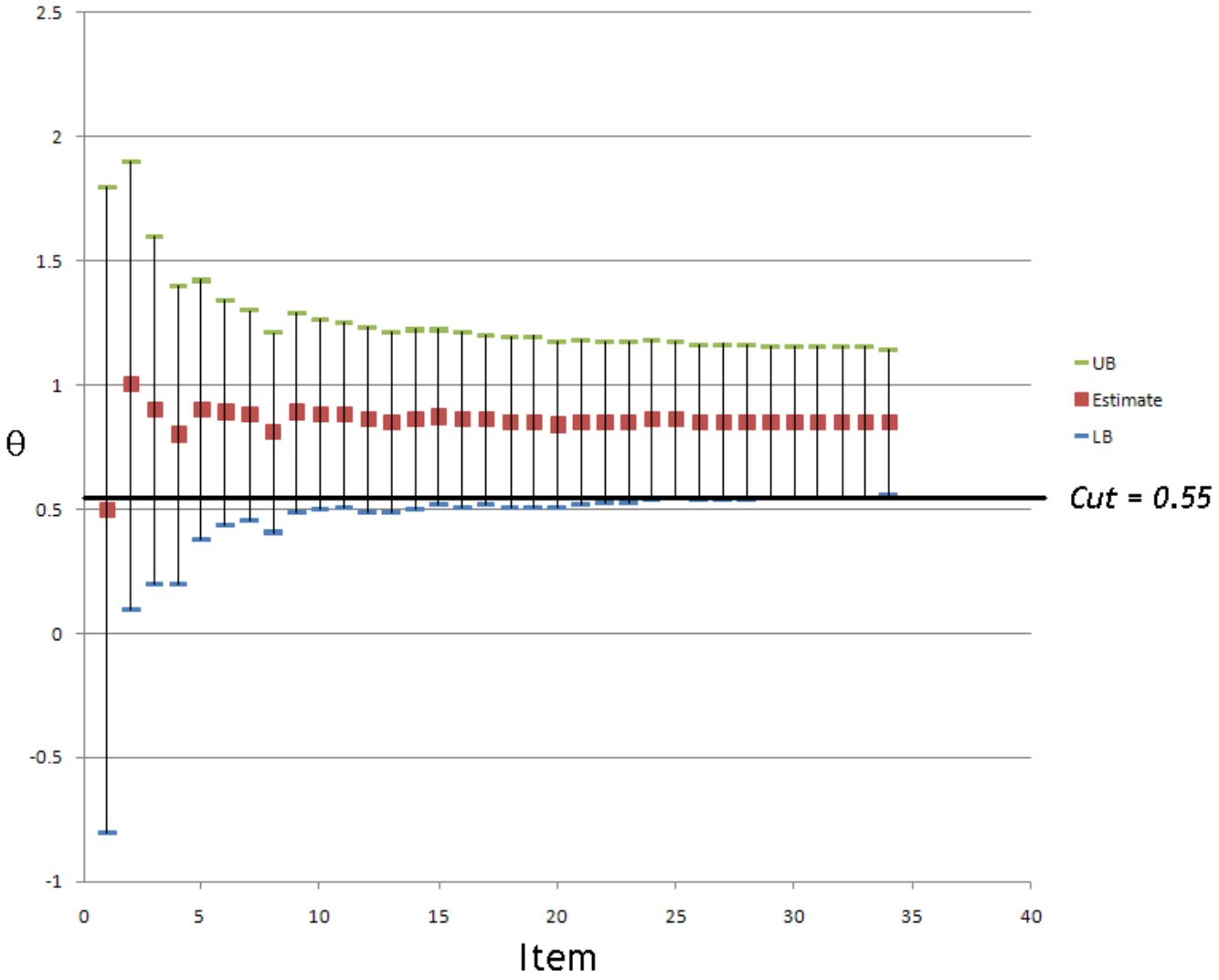
Example: item selection



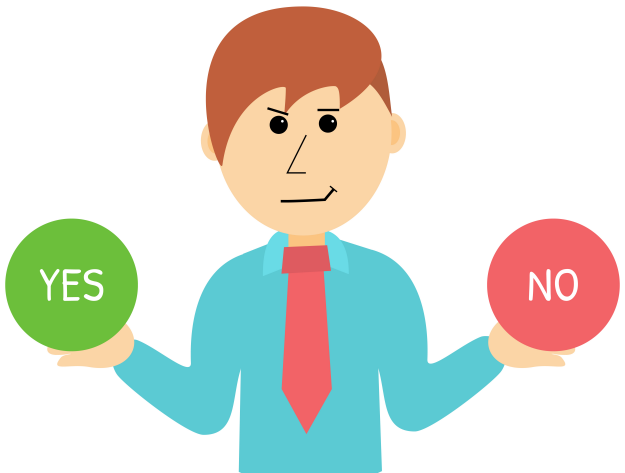
What would a CAT do?

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Example: classification



Pass/Fail CAT based on confidence interval



ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Part 2

A 5 step model to develop a CAT

(and provide validity documentation!)

Thompson & Weiss (2011) 5 step model

Seq.	Stage	Primary work
1	Feasibility, applicability, and planning studies	Monte carlo simulation; business case evaluation
2	Develop item bank content or utilize existing bank	Item writing and review
3	Pretest and calibrate item bank	Pretesting; item analysis
4	Determine specifications for final CAT	Post-hoc or hybrid simulations
5	Publish live CAT	Publishing and distribution; software development

1. Feasibility, applicability, planning

Big question: is CAT worth the investment?

If so, how can we develop a project plan and timeline?

1. Feasibility, applicability, planning

Answer: simulations

Simulate how a CAT would operate under specific, plausible conditions

Think of the results table you want to see

Bank size	Target SEM	Mean test length	Mean SEM
(current test)	-	100	.32
200	0.30	?	?
200	0.40	?	?
300	0.30	?	?
300	0.40	?	?

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

1. Feasibility, applicability, planning

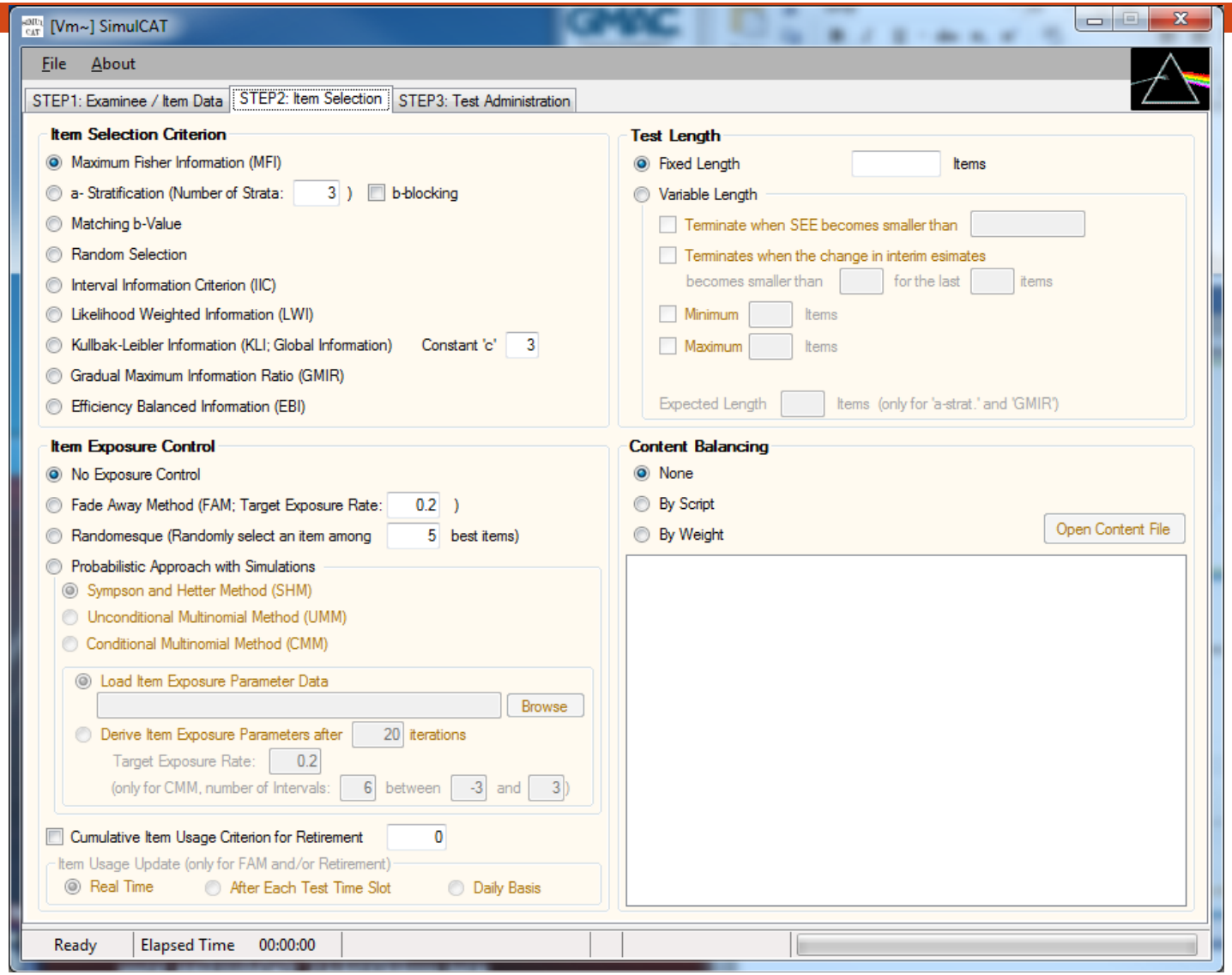
Software will do this for you, allowing you to simulate CATs for thousands of examinees in seconds

- CATSim (ASC)
- SimulCAT (Han)
- FireStar (Choi)

Easy to set up broad experiment

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

1. Feasibility, applicability, planning



ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

1. Feasibility, applicability, planning

Example takeaway:

CAT

- SEM=0.25
- average of 53 items

Current fixed test

- 100 items
- SEM=0.23 in middle and 0.35+ beyond θ of ±1.5

CAT will make test more accurate for extreme examinees, about same accuracy for middle, but with 50% reduction

1. Feasibility, applicability, planning

Another question: Business Case Evaluation

- Remember this table? Make one yourself at this point.

Annual examinees	Seat cost savings
1,000	$25,000
10,000	$250,000
100,000	$2,500,000

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

2. Develop item bank

Now that we have an idea what we need, we need to build it

CAT-based considerations:

- Difficulty spread
- Anticipated exposure/security issues
- TIF adequacy

Normal considerations

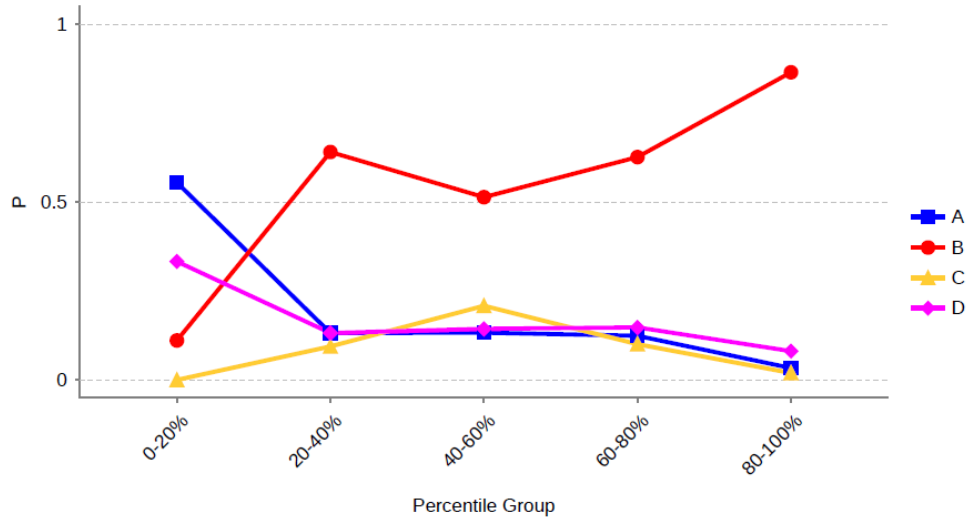
- Content blueprints
- Cognitive level
- Item types

3. Pretesting and analysis

Then calibrate - usually IRT

Also perform other due diligence

- Dimensionality
- DIF
- Model fit
- Distractor analysis
- Remove/revise items based on stats?
- Etc.



Item Statistics

Key	Scored	Options	Subscores	N	P	Total Rpbis	Mean	Flags
B	Yes	4	Initial Bank	1234	0.323	0.637	0.323	Low P

Option Statistics

Jason's mother gives him $300 to start a savings account at the bank. Jason earns $250 per month from his part-time job, and he puts half of this into the savings account. How much money does the account contain after 6 months, if there is no interest earned?

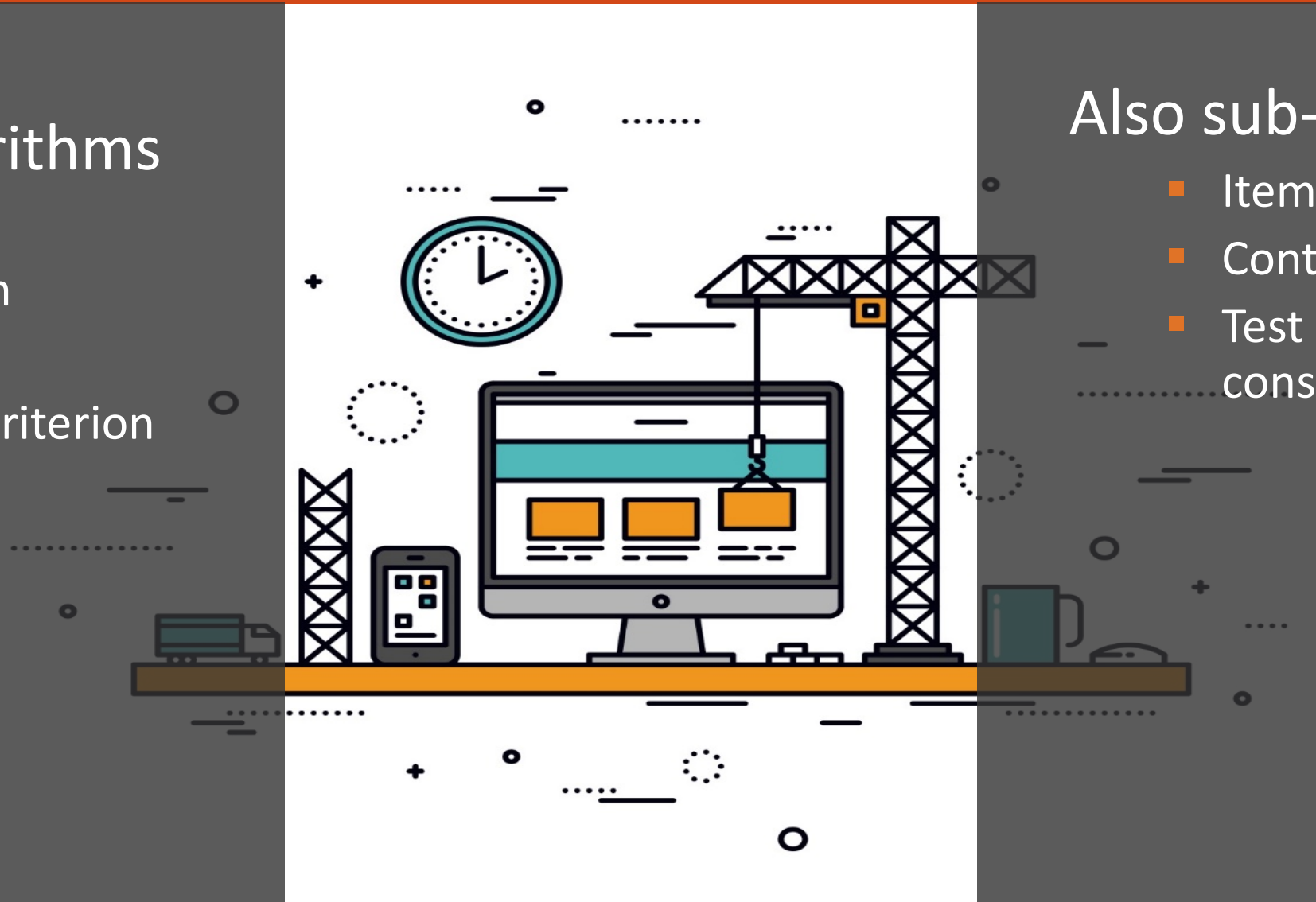
Option	Text	N	Prop	Rpbis	Weight
A	$750	68	0.055	0.103	0
B	**$1,050**	**398**	**0.323**	**0.637**	**1**
C	$1,300	67	0.054	0.099	0
D	$1,800	84	0.068	0.160	0
Omit		617	0.500	0.000	

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

4. Determine final specifications

Specify algorithms

- Starting point
- Item selection
- Scoring
- Termination criterion



Also sub-algorithms

- Item exposure
- Content distribution
- Test length constraints

4. Determine final specifications

But we must have a reason for selecting specifications

- Validity documentation
- Defensibility

Again, we turn to simulation studies

- Define competing conditions
- Create an experiment
- Run simulations
- Collate and analyze

Example simulation study

Bank type	Items	SEM	Max length	Bank-CAT θ correlation	ATL	90% less than __ items
Real	72	0.25	72	0.995	47.135	72
Real	72	0.3	72	0.981	28.486	72
Real	72	0.25	50	0.994	40.015	50
Real	72	0.3	50	0.981	24.987	50
Sim	100	0.25	50	0.988	34.544	50
Sim	150	0.25	50	0.980	28.663	40
Sim	200	0.25	50	0.976	27.188	34
Sim with b SD = 1.5	200	0.25	50	0.974	23.097	27
Sim	100	0.3	50	0.973	21.389	28
Sim	150	0.3	50	0.964	18.341	23
Sim	200	0.3	50	0.962	17.385	21
Sim with b SD = 1.5	200	0.3	50	0.956	15.268	18

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

4. Determine final specifications

Next: evaluate more practical issues

For example, time limits

- Can’t really set until you know how many items
- CAT-ASVAB approach: 90-95%

5. Publish live CAT

Once you have finalized your item bank and CAT design, time to publish

Need to put everything into item banker and CAT engine

- If developing your own, this can be the biggest step
- If purchasing, this is the easiest step

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Epilogue: Maintaining CAT

Like fixed form testing, maintenance is usually necessary

Check that performing as expected

- Is termination criterion being satisfied?
- Examinees hitting test length or other constraints?
- Average test length what you expected?

Exposure or security issues?

Part 3

Software and other hurdles

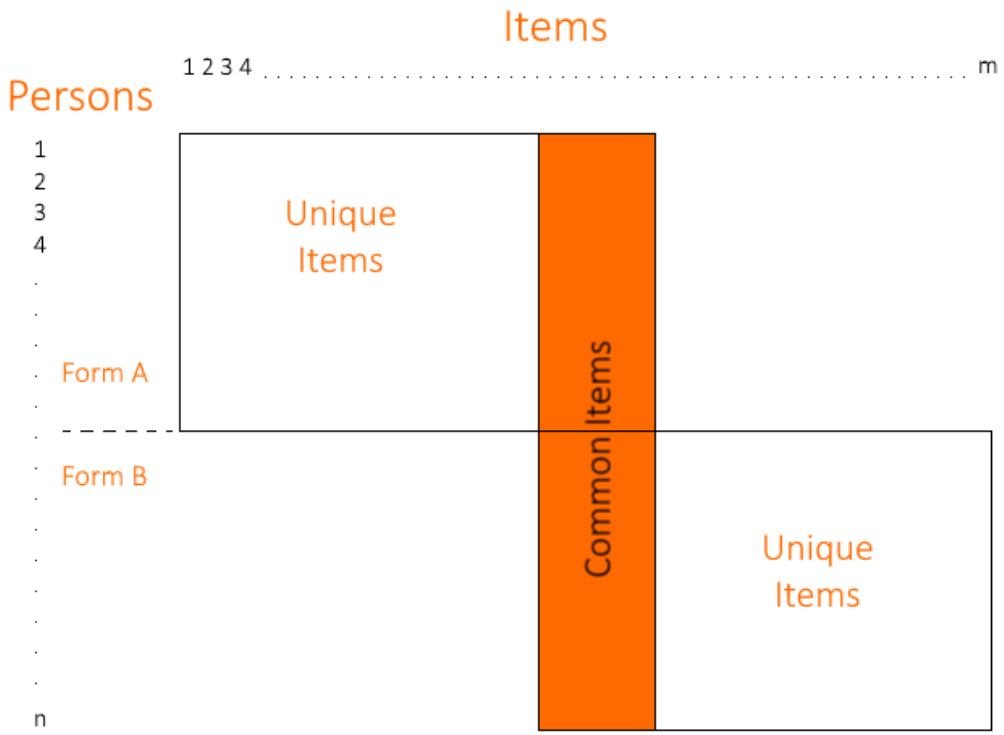
Software

IRT calibration

- R (free)
- jMetrik (free)
- Xcalibre
- FlexMIRT
- IRTPRO

IRT equating

- IRTEQ (Han, 2010)



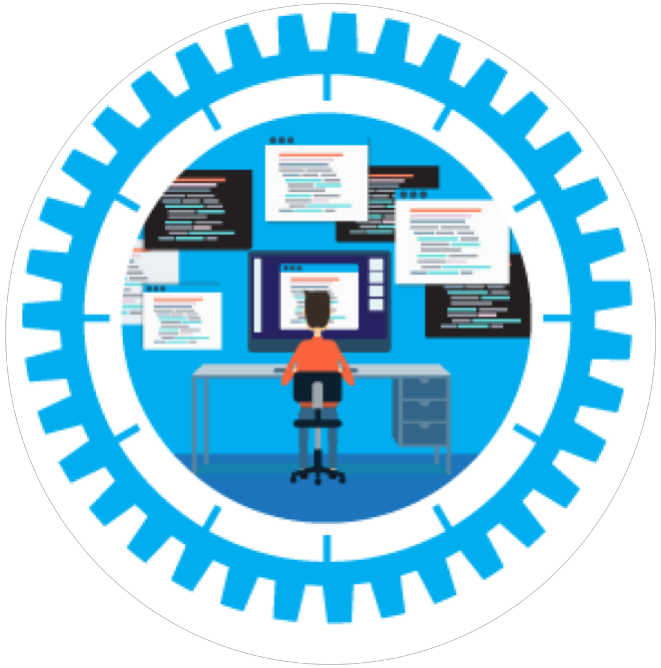
Item	a	b	c
Item.001	0.67	-2.95	0.26
Item.002	0.71	-1.74	0.26
Item.003	0.96	0.11	0.36
Item.004	1.19	1.87	0.24
Item.005	0.97	-1.24	0.25
Item.006	0.84	0.47	0.25
Item.007	1.01	-1.64	0.25
Item.008	1.08	0.71	0.24
Item.009	0.96	0.82	0.24
Item.010	0.96	1.04	0.25

Software

CAT Simulation

- CATSim
- SimulCAT (free)
- FireStar (free, R based)

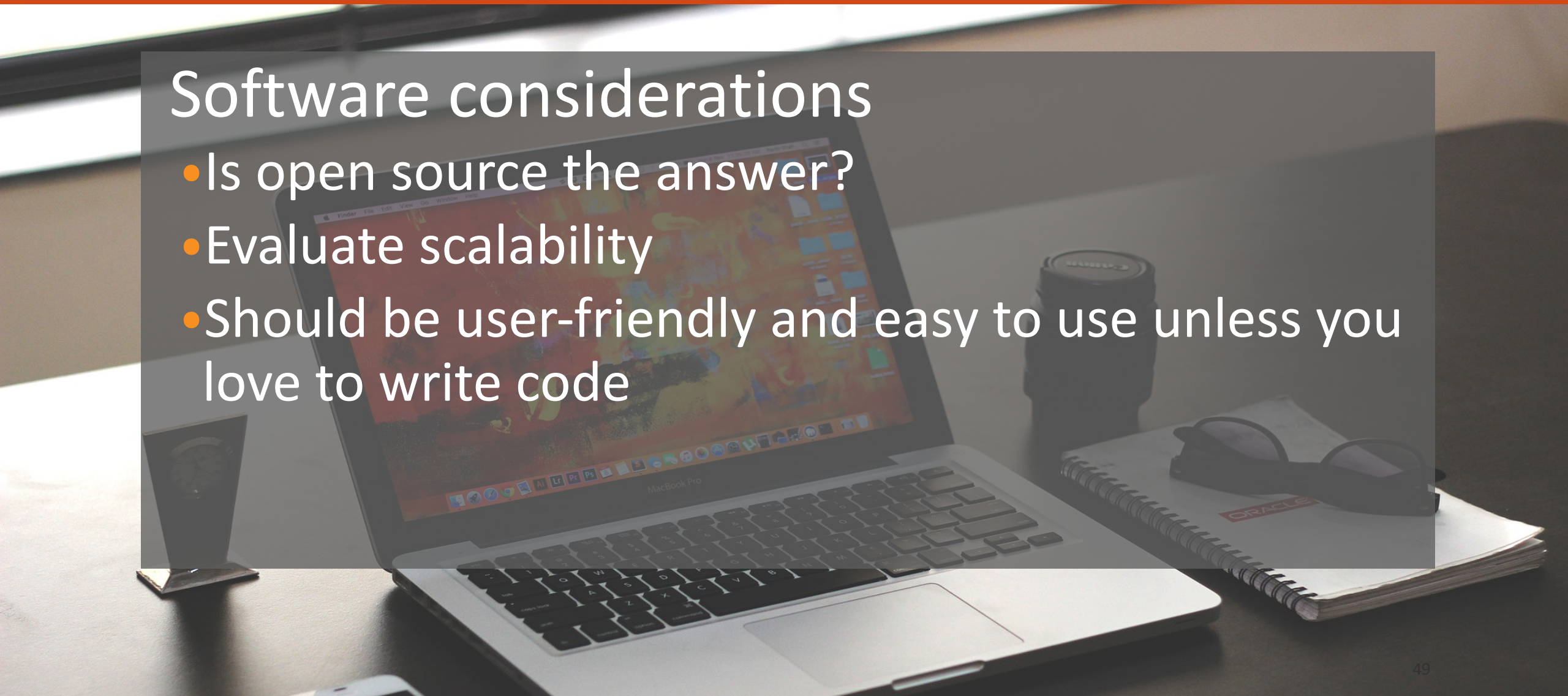
Let's run a simulation with SimulCAT!



CAT Delivery Platforms

- Buy
  - Prefab tests make it easy
- Rent
  - FastTest (Ada)
  - Concerto
  - TAO with new plugins
  - Can get consultants to help you
- Build
  - Always the option to build your own from scratch, but why reinvent the wheel?

Solutions



Software considerations

- Is open source the answer?
- Evaluate scalability
- Should be user-friendly and easy to use unless you love to write code

Part 4

CATs in the Wild

Hurdles to CAT (Just a Few)

Complexity

Need a platform and IRT software

Some orgs release of all items

Lack of data or access to it

Convincing stakeholders & PR

Positive ROI (Cost)

Sample size requirements

QTI does not support it

Few non-technical resources

Stories

Not really…

Preparedness

Let's go!

Examples of the hurdle effect:

- Biology professor
- Denmark health researcher
- Assessment coordinator in Mexico
- Grad student in Turkey

- Take-home message: it's not that hard if you have the right tools!

So, how do we breed more CATs?

- Invented in the 1970s
- Commercial and free platforms available
- HUNDREDS of publications, websites, resources
- A number of prefab CATs available

Yet still pretty much only used by large testing companies

Why?

Types of CATs

Pre-Employment Cognitive/Knowledge

- CAT-ASVAB
- LPCAT

Certification/Licensure

Types of CATs

Noncognitive – Likert scales (psych medical)

ADAPTIVE TESTING TECHNOLOGIES

MAYO CLINIC

NIH National Institutes of Health

PROMIS Patient Reported Outcomes Measurement Information System®

Noncognitive – Ideal point models (personality pre-employment)

UNITED STATES NAVY

CEB | SHL Talent Measurement

ASSESSMENT SYSTEMS FOR GOOD MEASURE™

Types of CATs

K12 assessments (formative and summative)

Cognitive Diagnostic Models

Types of CATs

University admissions

University placement

ACCUPLACER

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Types of CATs

Ready to take a real CAT?

Here we go!

Thank you!

Q&A?

assess.com/blog

nthompson@assess.com