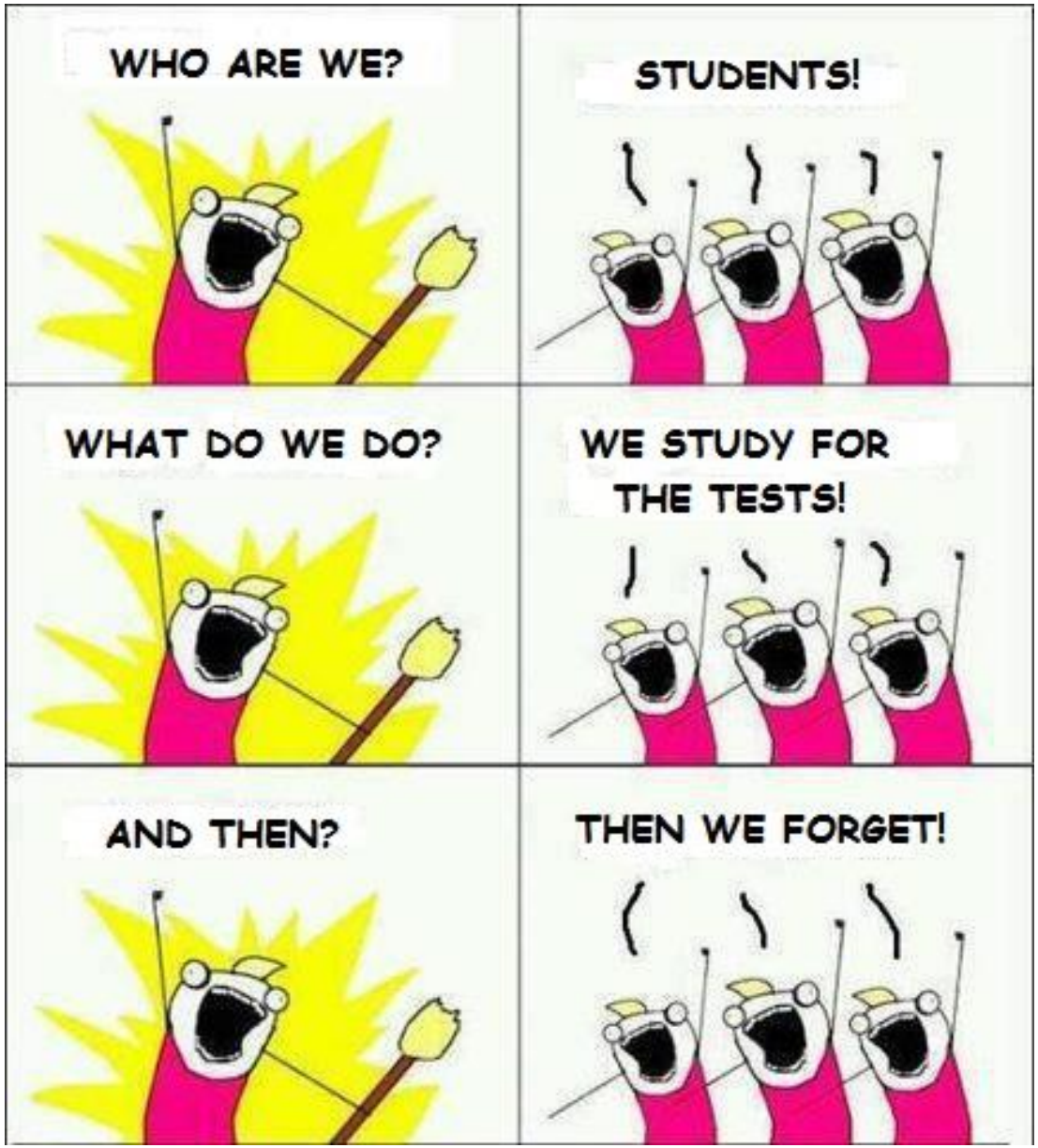
Adaptive assessment

Carlos F. Collares, M.D., M.Sc., Ph.D., FACMT

EDLAB Teach-Meet

April 13th, 2018

Maastricht University


WHO ARE WE?
STUDENTS!
WHAT DO WE DO?
WE STUDY FOR THE TESTS!
AND THEN?
THEN WE FORGET!

Progress testing as an assessment strategy

- Longitudinal assessment strategy.
- Systematic, repeated testing of all students of a school using the same test.
- Comprehensively covers all medical knowledge domains
- Can be administered from 2 to 4 times per year.
- Used by traditional, PBL and TBL schools.
- Used in undergraduate and postgraduate settings
- End-of-course level.
- MCQs
- Variable length and duration.

How is it with traditional, paper-based tests?

- All students answer the same set of questions
- Tests do not take into account the students' knowledge level
- Mismatch between test difficulty and knowledge levels may cause student demotivation, lower reliability and higher measurement error.
- Paper-based tests present less realistic challenges, as they do not allow test items to have pictures, audio and video, limiting the professional authenticity of the assessment.
- Paper-based tests also have more risk of breaches to test safety such as ilegal collusion.
- Re-testing (e.g. re-sits) might be burdensome when a new paper-base d test has to be created.

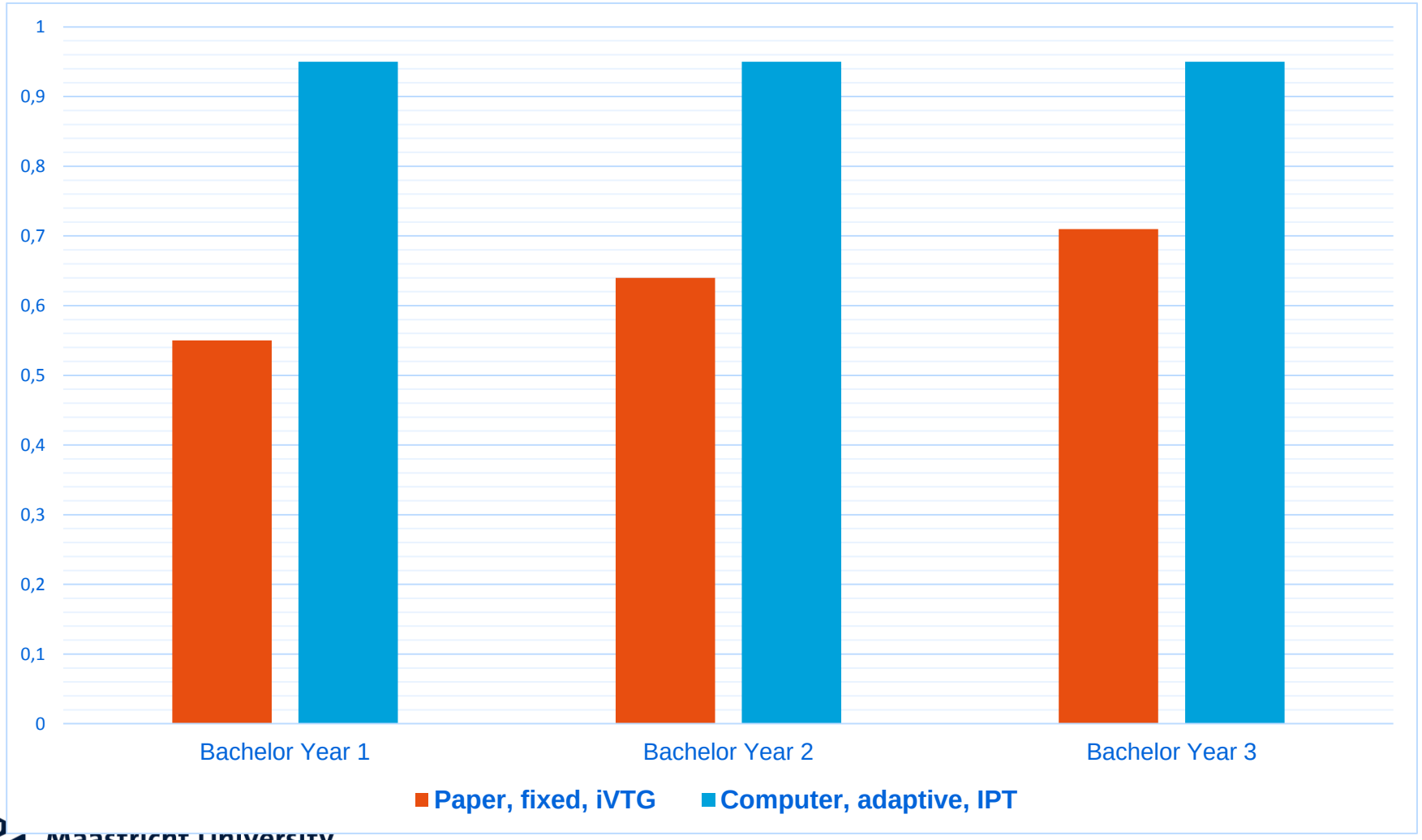
Computerized adaptive testing is an alternative

- CAT matches items' difficulty to students' ability
- An algorithm dynamically selects the difficulty of the next items based on students' performance in the previous answers.
- Instead of answering the same set of questions, each one of the test takers will receive an individually customized test, tailored to their level of knowledge
- CAT can reduce the length of the test by roughly 50%
- Potentially decreases student fatigue, while keeping or even enhancing reliability

What is reliability?

- The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group" (AERA, APA, NCME, 2014)
- Total score variance = "true" variance + error variance
- $Reliability = \frac{"true"\ variance}{total\ variance}$
- In other words, it is a signal-to-noise ratio
- Values below 0,5 suggest that your test scores have more noise than signal.
- Values close to 1,0 indicate low levels of measurement error

Comparison of reliability estimates between a CAPT and a paper-based progress test



Reliability of paper-based progress tests

2012; 34: 683–697 MEDICAL TEACHER

AMEE GUIDE

A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71

WILLIAM WRIGLEY, CEES PM VAN DER VLEUTEN, ADRIAN FREEMAN & ARNO MUIJTJENS

Department of Educational Development and Research, The Netherlands

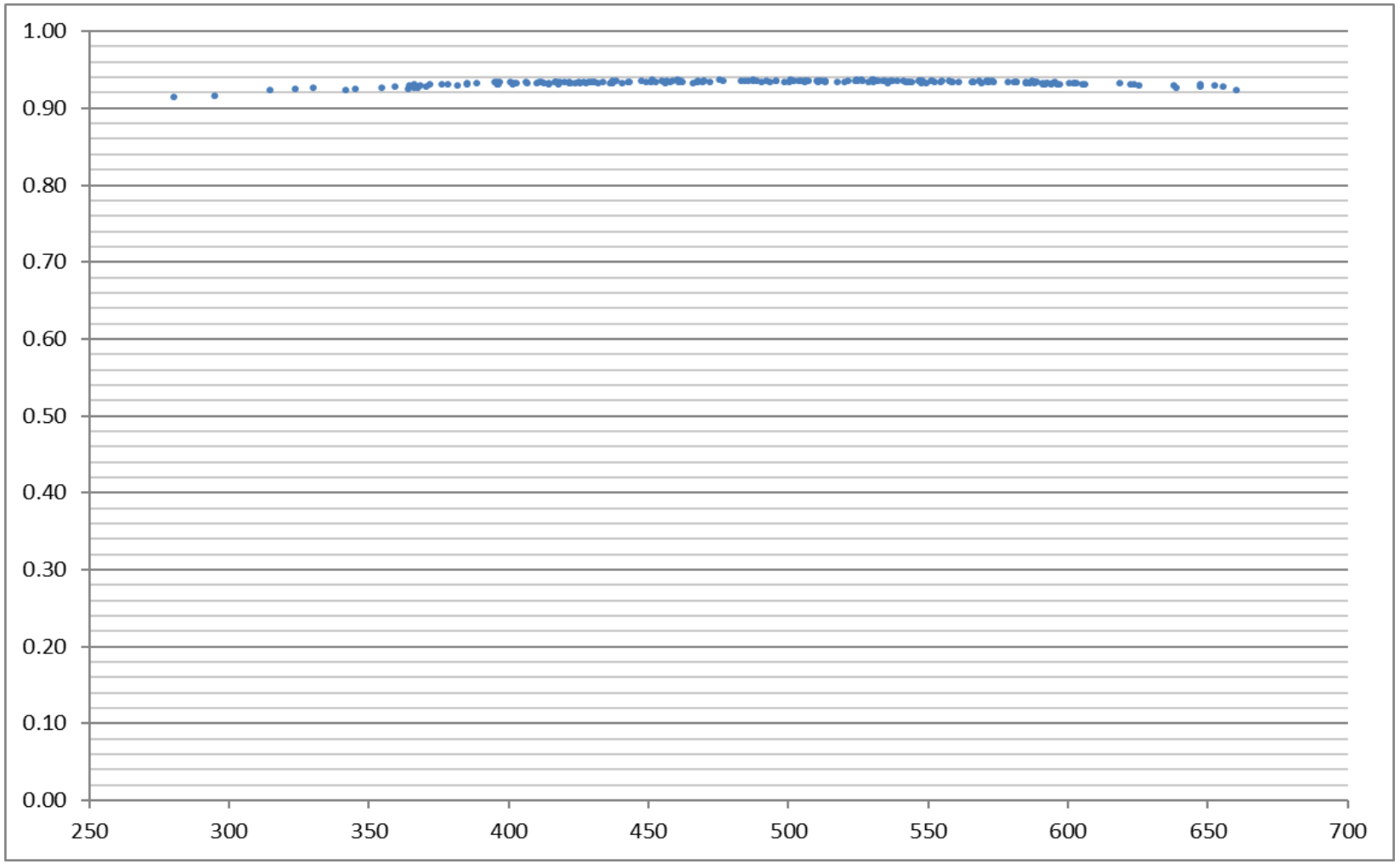
Table 2. G coefficients for test size (number of items) by test frequency for Maastricht University students Years 1-6 in the academic year 2010/11.

		Year 1						Year 2						
		Test size							Test size					
		25	50	75	100	150	200		25	50	75	100	150	200
Test Frequency	1	0.18	0.29	0.36	0.42	0.49	0.54	Test Frequency	1	0.23	0.37	0.45	0.51	0.59	0.63
	2	0.30	0.45	0.53	0.59	0.66	0.70		2	0.38	0.54	0.62	0.67	0.74	0.78
	3	0.40	0.55	0.63	0.68	0.74	0.78		3	0.48	0.63	0.71	0.76	0.81	0.84
	4	0.47	0.62	0.70	0.74	0.79	0.82		4	0.55	0.70	0.77	0.81	0.85	0.87
		Year 3								Year 4					
		Test size							Test size						
		25	50	75	100	150	200		25	50	75	100	150	200
Test Frequency	1	0.23	0.36	0.45	0.51	0.59	0.64	Test Frequency	1	0.32	0.48	0.57	0.63	0.71	0.76
	2	0.37	0.53	0.62	0.68	0.74	0.78		2	0.49	0.65	0.73	0.78	0.83	0.86
	3	0.47	0.63	0.71	0.76	0.81	0.84		3	0.59	0.74	0.80	0.84	0.88	0.90
	4	0.54	0.69	0.77	0.81	0.85	0.88		4	0.66	0.79	0.84	0.87	0.91	0.93
		Year 5								Year 6					
		Test size							Test size						
		25	50	75	100	150	200		25	50	75	100	150	200
Test Frequency	1	0.30	0.46	0.55	0.62	0.70	0.74	Test Frequency	1	0.30	0.45	0.55	0.61	0.69	0.74
	2	0.47	0.63	0.71	0.76	0.82	0.85		2	0.46	0.62	0.71	0.76	0.82	0.85
	3	0.57	0.72	0.79	0.83	0.87	0.90		3	0.56	0.71	0.78	0.82	0.87	0.89
	4	0.64	0.77	0.83	0.87	0.90	0.92		4	0.63	0.77	0.83	0.86	0.90	0.92

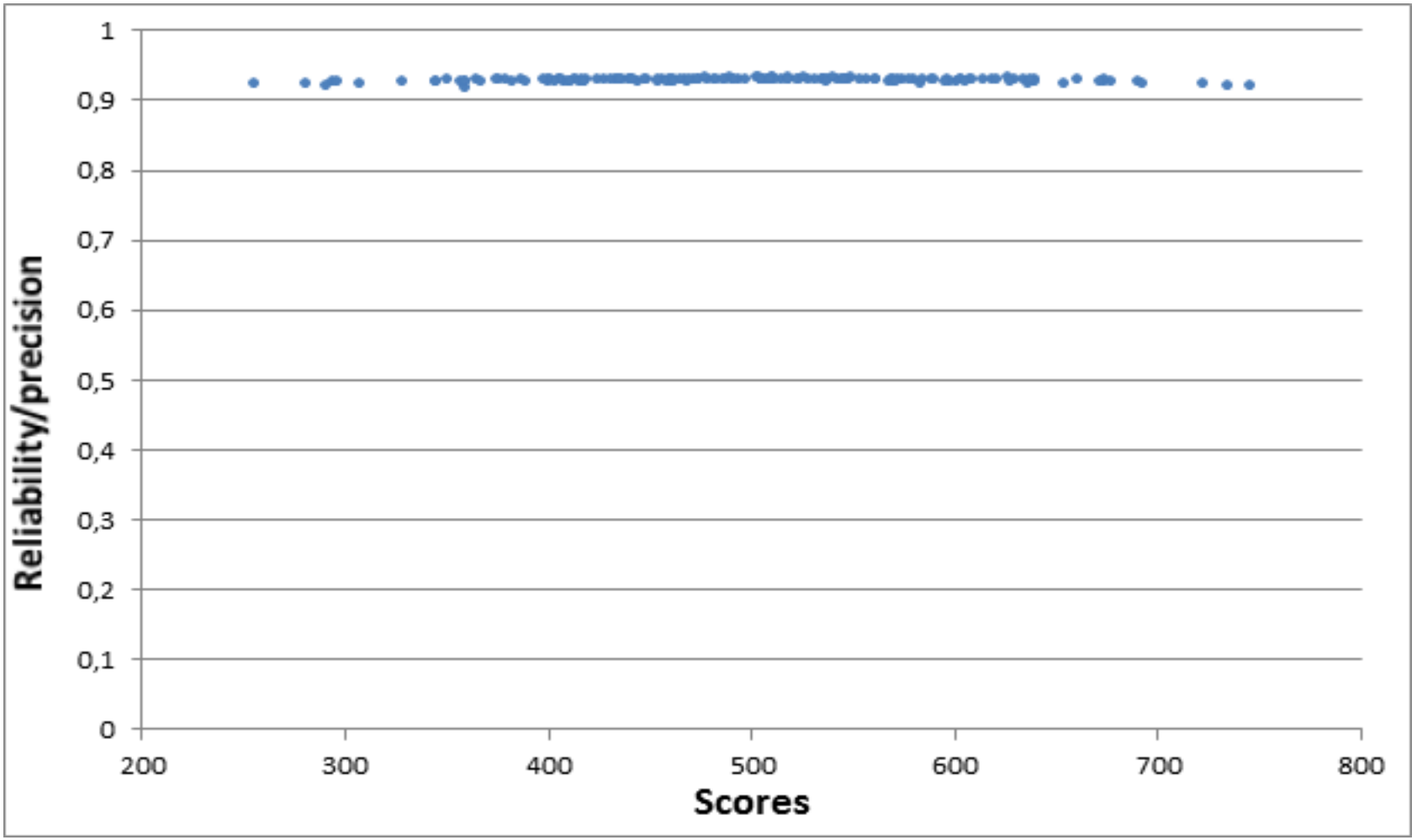
Computerized adaptive testing is reliable

- Computerized adaptive progress tests (CAPTs) have been used in Brazil, Mexico, Finland and Georgia
- In all instances, CAPTs had formative purposes only.
- Maastricht University is a pioneer in the use of CAPTs for summative purposes
- Reliability > 0,90
- Test-retest reliability > 0,70
- Disattenuated correlation > 0,80

Individual reliability estimates of a CAPT in Helsinki, 2017



Individual reliability estimates of a CAPT in Maastricht, 2017

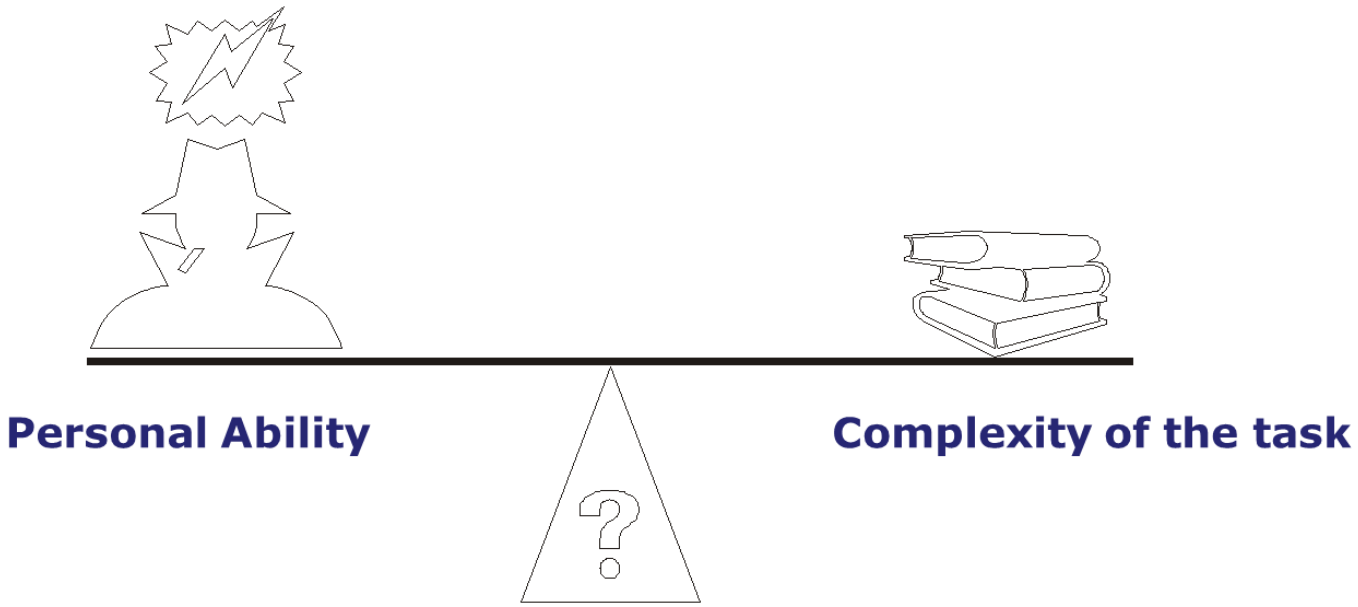


Proposition 1

- Computerized adaptive testing is an adequate tool for the assessment ***of*** learning.

- But how does adaptive testing work?
- Time for a technical intermezzo!

Each test item is a kind of “battle”



What is the probability that the person is “better” than the complexity of the task?

What is the probability that the person will win the “see-saw battle”?

There are mathematical models able to estimate the probability of who will win this battle

- The most robust model is was created by the Danish mathematician Georg Rasch in the 1960s.
- It establishes a formal relationship between the probability of success in the item, the difficulty of the item and the ability of the test taker.

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}}$$

Translating the formula in plain words…

- The Rasch model takes the difficulty of the items into account to provide more accurate estimates of the ability levels of the test takers.
- This is not accomplished by classical scoring approaches.

So what is the secret?

Will I be able to understand this Rasch model?

How does it work precisely?

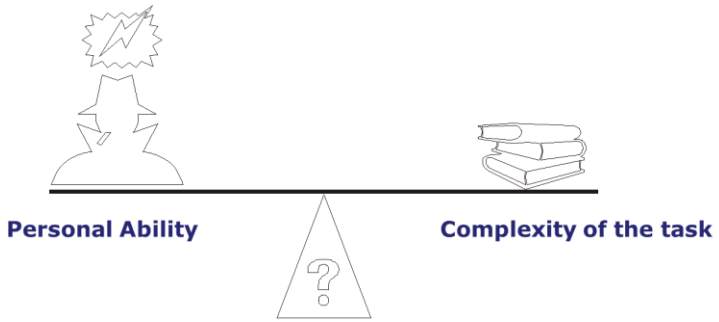
ord		Item 1	Item 2	Item 3	Item 4	Item 5	Total		
1	Suj 2	1	1	1	1	1	5	Soma	36
2	Suj 5	1	1	1	1	0	4	Média	2,57
3	Suj 1	1	1	1	1	1	5	Var	3,10
4	Suj 11	1	0	0	0	0	1	DP	1,76
5	Suj 3	1	1	1	1	0	4	Alfa	
6	Suj 13	0	0	0	0	0	0	N/N-1	1,08
7	Suj 10	1	0	0	0	0	1		
8	Suj 12	1	0	0	0	0	1	Alfa/KR	0,71
9	Suj 9	1	0	0	0	0	1		
10	Suj 7	1	1	1	0	0	3		
11	Suj 8	1	1	1	0	0	3		
12	Suj 4	1	1	1	1	0	4		
12	Suj 14	0	0	0	0	0	0		
13	Suj 6	1	1	1	1	0	4		
	ID	0,86	0,57	0,57	0,43	0,14	2,57		
	Corr It-Tot	0,60	0,94	0,94	0,87	0,56	1,00		
	Desv. Padr	0,35	0,49	0,49	0,49	0,35	Soma:		2,18
	Var	0,13	0,26	0,26	0,26	0,13	Soma:		1,05

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1,00				
Item 2	0,47	1,00			
Item 3	0,47	1,00	1,00		
Item 4	0,35	0,75	0,75	1,00	
Item 5	0,17	0,35	0,35	0,47	1,00

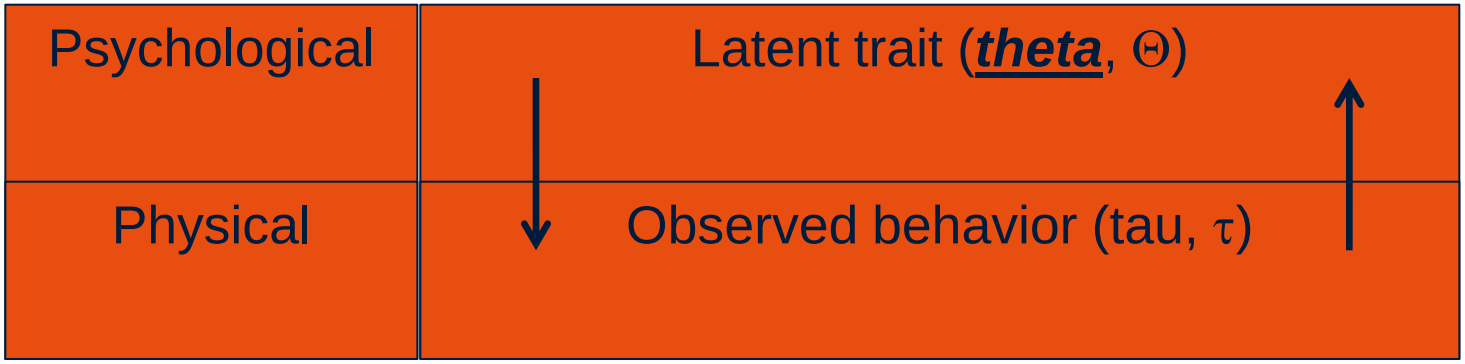
	Item 1	Item 2	Item 3	Item 4	Item 5	Total		
Suj 1	1	1	1	1	1	5	Soma	36
Suj 2	1	1	1	1	1	5	Média	2,57
Suj 3	1	1	1	1	0	4	Var	3,10
Suj 4	1	1	1	1	0	4	DP	1,76
Suj 5	1	1	1	1	0	4	Alfa	
Suj 6	1	1	1	1	0	4	N/N-1	1,08
Suj 7	1	1	1	0	0	3		
Suj 8	1	1	1	0	0	3	Alfa/KR	0,71
Suj 9	1	0	0	0	0	1		
Suj 10	1	0	0	0	0	1		
Suj 11	1	0	0	0	0	1		
Suj 12	1	0	0	0	0	1		
Suj 13	0	0	0	0	0	0		
Suj 14	0	0	0	0	0	0		
ID	0,86	0,57	0,57	0,43	0,14	2,57		
Corr It-Tot	0,60	0,94	0,94	0,87	0,56	1,00		
Desv. Padr	0,35	0,49	0,49	0,49	0,35	Soma:		2,18
Var	0,13	0,26	0,26	0,26	0,13	Soma:		1,05

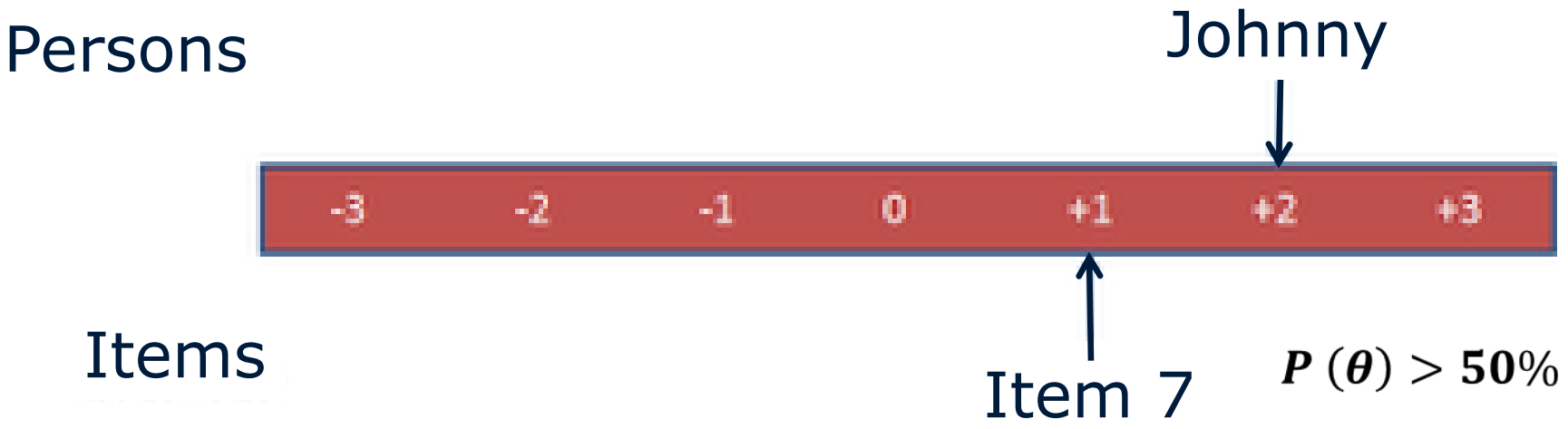
	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1,00				
Item 2	0,47	1,00			
Item 3	0,47	1,00	1,00		
Item 4	0,35	0,75	0,75	1,00	
Item 5	0,17	0,35	0,35	0,47	1,00

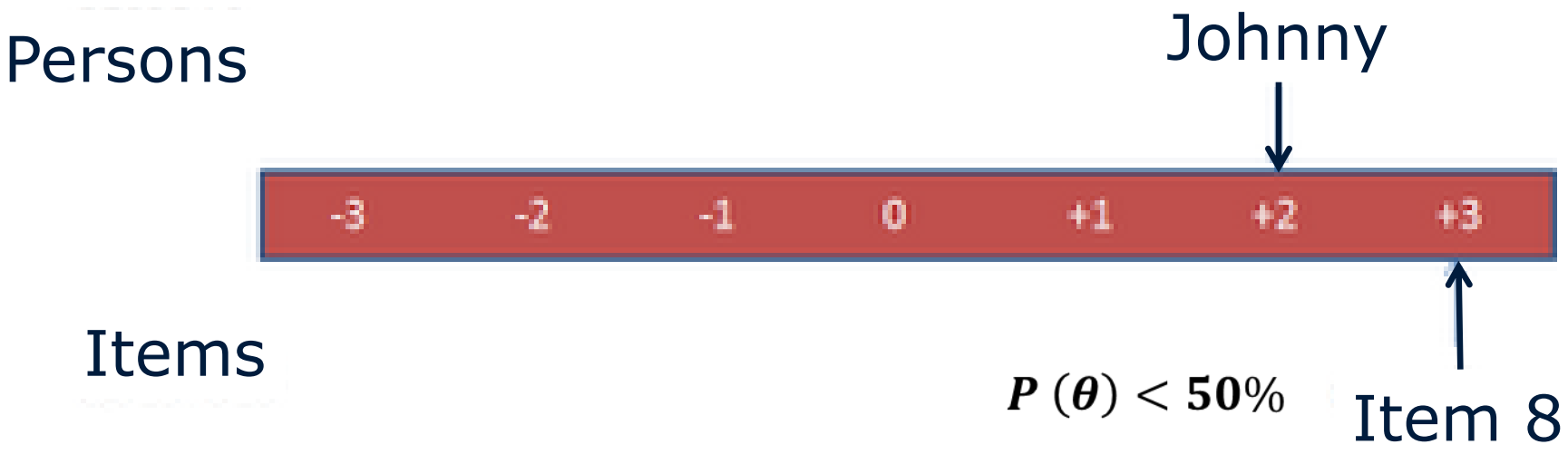
The secret of the Rasch model is that, through a series of successive attempts, it puts items' difficulties and students' knowledge levels in the same scale: the theta scale



Item response theory models such as the Rasch model use the notion of a latent variable (not observed) from the observed behaviors (raw scores)




Persons
Johnny
-3
-2
-1
0
+1
+2
+3
Items
Item 7
$P(\theta) > 50\%$


Persons
Johnny
-3
-2
-1
0
+1
+2
+3
Items
$P(\theta) < 50\%$
Item 8

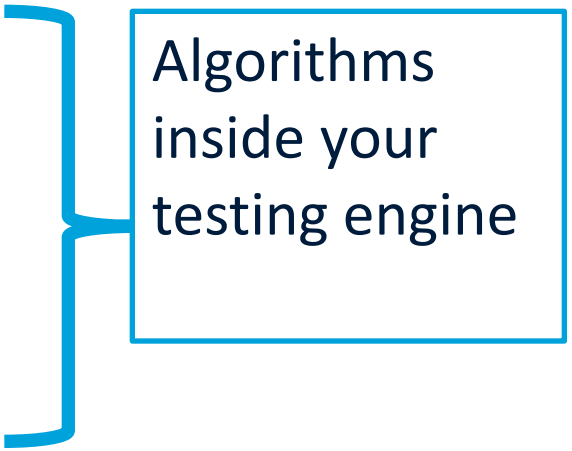
CAT Components

- 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

Calibrated item bank

- Size of the item bank: commensurate to the test size (min: 1:7; ideal > 1:13)
- Test equating/linking

- Anchor persons/anchor items
- Concurrent calibration
- Sequential linking

- Choice of the model:

- Rasch/1PLM; 2PLM; 3PLM
- More parameters, more overexposure → Rasch
- Non-IRT CAT (cognitive diagnostic modeling CAT) → future of adaptive testing

System options

- TestLife
- FastTest
- CONCERTO
- catR

End of technical intermezzo

Proposal 2

- Computerized adaptive testing is an excellent assessment tool ***for*** learning
- Alignment with learning theories
- Constructivism: “zone of proximal development”
- Cognitive load theory: the adaptive approach prevents cognitive under- and overload
- Social cognitive theory: better score accuracy leads to better self-regulation, self-efficacy and attainment

Proposal 3

- Computerized adaptive testing is an excellent assessment tool ***as*** learning
- Recent evidence demonstrates a positive impact of adaptive testing on students' achievement, motivation, engagement and subjective test experience

Journal of Educational Psychology
2018, Vol. 110, No. 1, 27–45

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000205

Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subjective Test Experience

Andrew J. Martin
University of New South Wales

Goran Lazendic
Australian Curriculum, Assessment, and Reporting Authority, Sydney, Australia

Lessons learned so far

- **1) Public relations**
  - Why certain things can happen, like failing after only a few questions
  - What are theta scores? What about the residues?
  - Educate staff, students, relatives, many times
  - *Communication is a key element for success*
- **2) Long-term Sustainability**
  - Requires specially designed software, good network and hardware infrastructure.
  - Home-made solutions are a protection against absurd pricing changes, but may limit access to features of commercial systems

Lessons learned so far

- **3) Item Exposure**
  - Some items will be used far more often than others, depending on the level of the test takers and your distribution of item difficulties.
  - 1PL has the same information for all items: less overexposure, but scores are also less predictive.
  - 3PL: more predictive, but yields much more overexposure.
  - **Start including easier items NOW**
  - **Start increasing scenario-based items NOW**

Lessons learned so far

- **4) “High maintenance”**
  - Requires experts for IRT calibration and CAT simulation research
  - Content expertise to systematically discard items that are no longer updated to current scientific standards to keep the item bank clean
  - Even though some items may leak, periodic surveillance of item parameter drift may quickly identify possibly items destined to retirement.

Lessons learned so far

- **5) Extra caution on content validity**
  - Requires much more refined blueprinting at the subscore level and subsequent algorithm specification to avoid exposing the student twice to a topic already covered in a previous item, or not exposing him at all to an important topic.
- **6) Use a representative sample to calibrate**
  - If your items are calibrated in a small sample, from just a few institutions, your scores will likely have improperly high or low values when compared to the whole population of interest,

Lessons learned so far

- **7) There is no such thing as a perfect world**
  - Students like the overall CAT experience and if given the opportunity to choose between CAT and paper-based test, most of them prefer CAT (>70-80%), especially due to less fatigue and immediate score reporting, BUT…
  - Feedback to students becomes limited to subscore level and/or feedback prompts (rubrics) due to test safety. Items cannot be disclosed anymore.
  - Students cannot go back to review the answers.
  - The higher the stakes, the higher is the probability of items leaking

The adaptive approach maximizes test utility

- **Reliability** = homogeneously high, including early years
- **Validity** = potential construct-irrelevant variance is no longer an issue; content validity ensured by blueprint
- **Educational impact** = aligned to modern learning theories, recent evidence suggests positive impact especially for females and older students
- **Acceptability** = usually high (some students compare it to a *video game*) but depends on local context
- **Costs** = decreasing as more schools participate of the item bank construction

E-mail: c.collares@maastrichtuniversity.nl
Twitter: @carlosfcollares