Test Design and Development: Psychometric Considerations

Enis Doğan

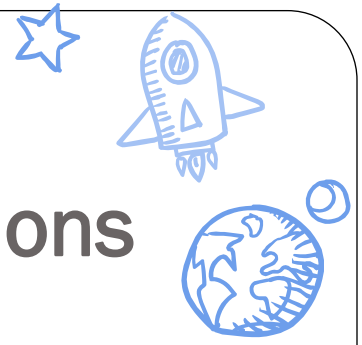
Research Fellow/Resident Expert at NCES

February, 19 2016

Overview

- Relevant *Standards*
- Where to start?
- Psychometric specifications
  - Individual items
  - Forms
  - The reporting scale
- Recommendations

The Standards for Educational and Psychological Testing: Test Specifications



Standard 4.2: In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections.

NCES STANDARD 2-6: EDUCATIONAL ASSESSMENT AND TESTING

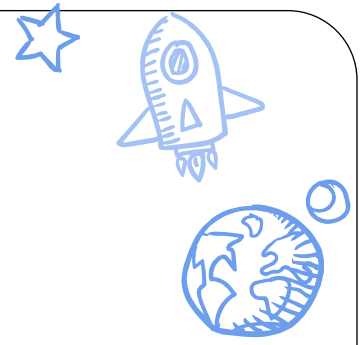
- **STANDARD 2-6-1: Instrument Development**—

  All test instruments used in NCES assessment surveys must be developed following an explicit set of specifications. …The instrument documentation must include the following:

  ….

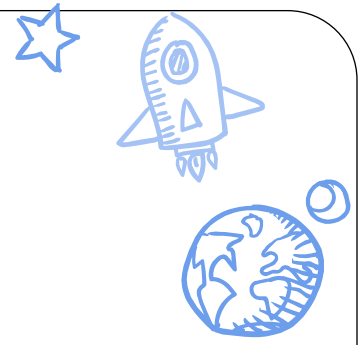
  8. Desired psychometric properties of the items, and the instrument as a whole

  ….



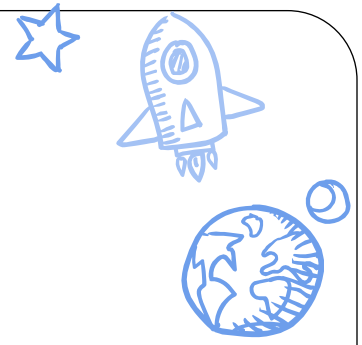
Start with the end in mind

- What will you be reporting out?
  - Scale scores
  - Subscales
  - Growth scores
  - Achievement level classification
  - Mastery probability



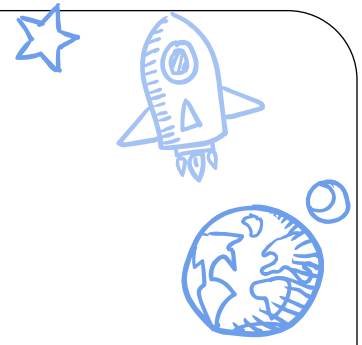
Psychometric specifications

1. Individual items
2. Forms
3. The reporting scale



Psychometric specifications

1. Individual items
   - Scoring rules, rubrics, and rater reliability
   - Difficulty and response distribution
   - Discrimination
   - DIF
2. Forms
3. The reporting scale



I. Individual items: Scoring

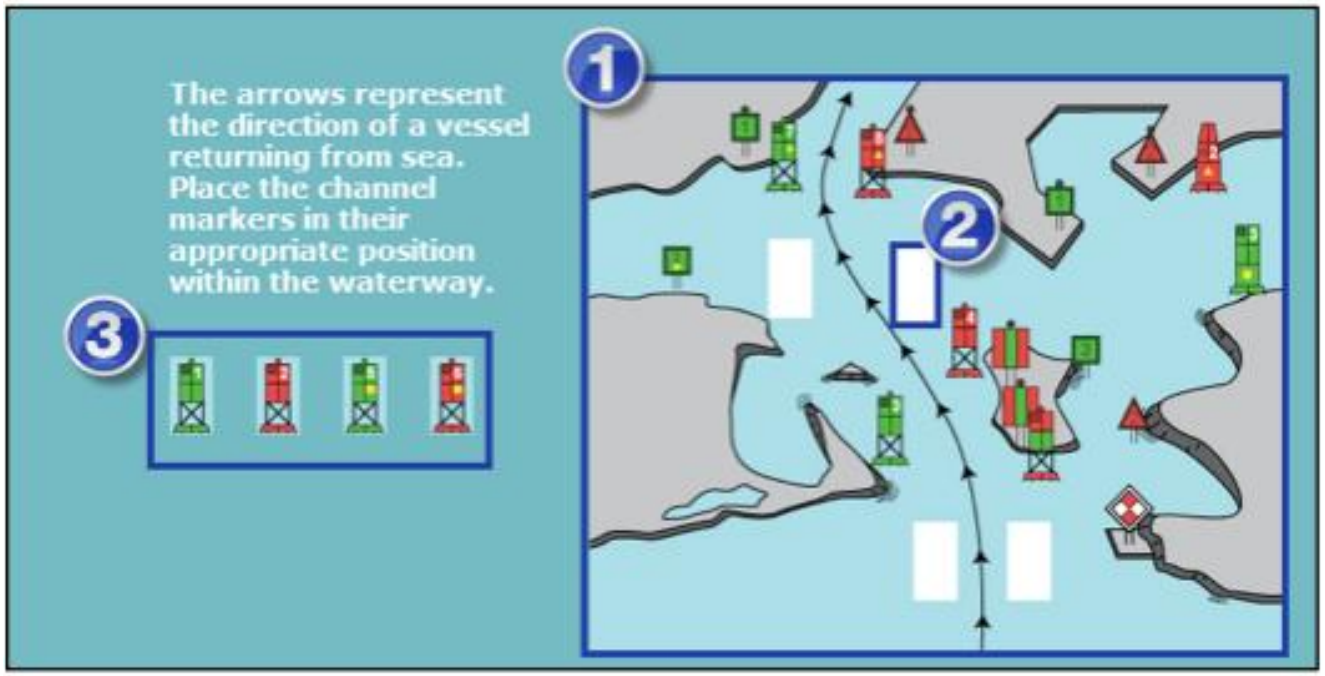
- Machine-scored items
  - Multiple choice, multiple select
  - Hot spot
  - Drag and drop

I. Individual items: Scoring

- Machine-scored items
  - Multiple choice, multiple select
  - Hot spot
  - Drag and drop

I. Individual items: Scoring

- Machine-scored items
  - Multiple choice, multiple select
  - Hot spot
  - Drag and drop



I. Individual items: Scoring

- Machine-scored items
  - Multiple choice, multiple select
  - Hot spot
  - Drag and drop

- Partial credit rules in machine-scored items
  - Give partial credit whenever you can
  - Scoring rules first, item development later

I. Individual items: Scoring

Read the story "Feathers," a traditional story about a rabbi who is a spiritual community leader. Then answer the questions.

Feathers

1 A sharp-tongued woman was accused of starting a rumor. When she was brought before the village rabbi, she said, "I was only joking. My words were spread by others, and so I am not to blame."

2 But the victim demanded justice, saying, "Your words soiled my good name!"

3 "I'll take back what I said," replied the sharp-tongued woman, "and that will take away my guilt." When the rabbi heard this, he knew that this woman truly did not understand her crime.

Part A

What is the meaning of **soiled** as it is used in paragraph 2?

A. Involved
B. Damaged
C. Emphasized
D. Identified

Part B

Which **two phrases** help the reader understand the meaning of soiled?

A. " … starting a rumor." (paragraph 1)
B. " … I was only joking." (paragraph 1)
C. " … my good name!" (paragraph 2)
D. I'll take back …" (paragraph 3)
E. " … take away my guilt" (paragraph 1)
F. " … understand her crime" (paragraph 3)

I. Individual items: Scoring

Part A	Part B	Final score?
Incorrect	Incorrect	0
Incorrect	Partially correct	?
Incorrect	Correct	?
Correct	Incorrect	?
Correct	Partially correct	?
Correct	Correct	2

I. Individual items: Rubric choices

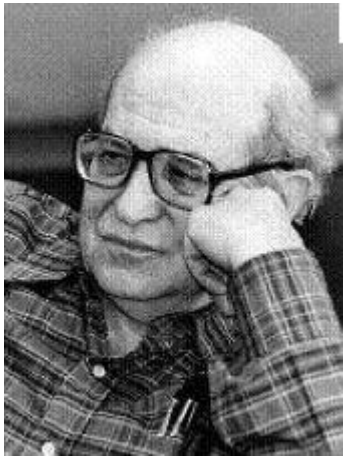
- Generic vs. task-specific
- **Analytic/Trait Rubrics**: Individual characteristics of a response judged separately
  - Advantage: Provides more differentiated evaluation
  - Disadvantage: Time consuming and susceptible to Halo effect
- **Holistic Rubrics** : Overall judgment of the quality of response
  - Advantage: Less time consuming, no Halo effect
  - Disadvantage: When student work is at varying levels spanning the criteria points it can be difficult to select the single best description.

Narrative Task (NT)

Construct Measured	Score Point 3	Score Point 2	Score Point 1	Score Point 0
Written Expression	The student response • is ***effectively*** developed with narrative elements and is ***consistently appropriate*** to the task; • is ***effectively*** organized with ***clear and coherent*** writing • uses language **effectively** to clarify ideas.	The student response • is developed with ***some*** narrative elements and is ***generally appropriate*** to the task; • is organized with ***mostly coherent*** writing; • uses language that is ***mostly effective*** to clarify ideas.	The student response • is ***minimally*** developed with ***few*** narrative elements and is ***limited in its appropriateness*** to the task; • demonstrates ***limited*** organization and coherence; • uses language to express ideas with ***limited*** clarity.	The student response • is ***undeveloped*** and/or ***inappropriate*** to the task; • **lacks** organization and coherence; • ***does not*** use language to express ideas with clarity.
Knowledge of Language and Conventions	The student response to the prompt demonstrates **full command** of the conventions of standard English at an appropriate level of complexity. There may be a **few minor errors** in mechanics, grammar, and usage, but **meaning is clear**.	The student response to the prompt demonstrates **some command** of the conventions of standard English at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **occasionally impede understanding**, but the **meaning is generally clear**.	The student response to the prompt demonstrates **limited command** of the conventions of standard English at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **often impede understanding.**	The student response to the prompt **does not demonstrate command** of the conventions of standard English at the appropriate level of complexity. **Frequent and varied errors** in mechanics, grammar, and usage **impede understanding**.

I. Individual items: Rater reliability

- Exact and adjacent agreement
- Kappa: adjusts for chance agreement
- Common mistakes:
  - Using same criteria for all items regardless of score range
  - Not using Kappa
  - Not specifying criteria in advance



Jacob Cohen1923 – 1998

I. Individual items: Rater reliability

npr WAMU 88.5 news arts & life music programs shop

TECHNOLOGY

Robot Eyes As Good As Humans When Grading Essays

Updated April 24, 2012 · 5:37 PM ET
Published April 24, 2012 · 3:00 PM ET

Listen to the Story
All Things Considered
4:12

+ Playlist Download Embed Transcript

A new study has determined that some automated essay graders can do as good of a job as humans. Melissa Block talks with *New York Times* education columnist Michael Winerip about the study and the weaknesses of automatic essay readers.

I. Individual items: Difficulty

- Desired range of difficulty for individual items
  - p-values between .05 and .95

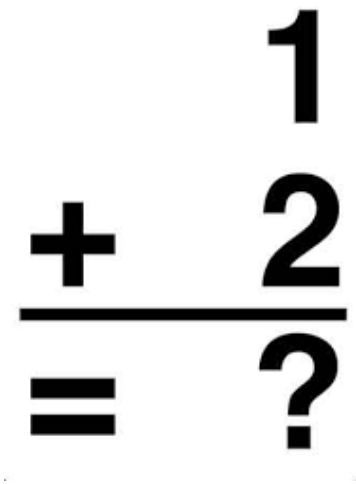
Which of the functions below is the inverse of $f(x) = 6x + 4$?

A) $y = 4x - 6$

B) $y = \frac{x-4}{6}$

C) $y = 6x - 4$

D) $y = \frac{x-6}{4}$



I. Individual items: Difficulty

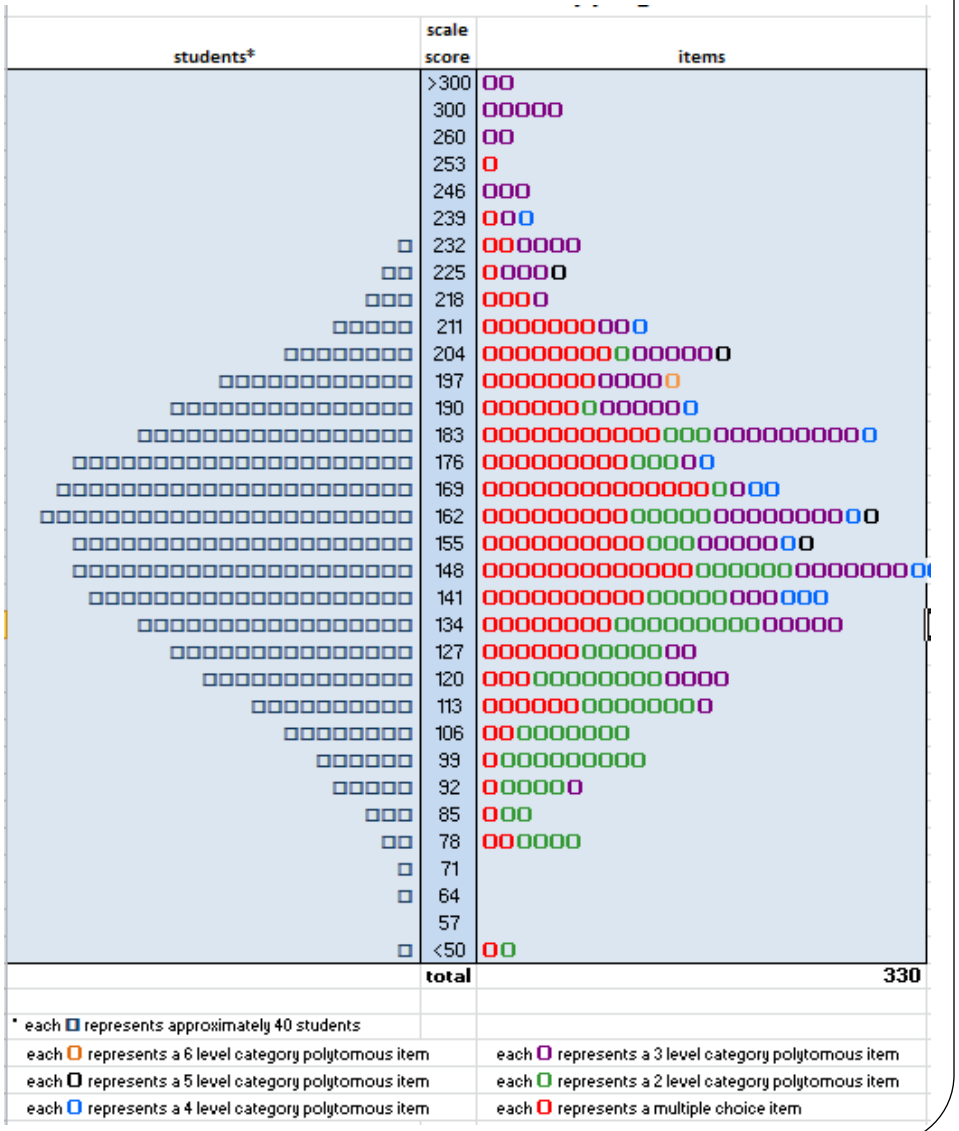
- How to write items with a certain level of difficulty?
  - Need a framework for item difficulty
    - Cognitive complexity
    - Response format
    - Stimulus type and load
    - Passage difficulty (Reading and Writing)
  - Requires ongoing research as we create in types of digital items and as students' familiarity with them changes over time

I. Individual items: Difficulty

- Target distribution of item difficulty
  - Approximately uniform
  - Non-uniform:
    - At where the students are
    - Below where the students are
    - Above where the students are
    - More dense at selected points

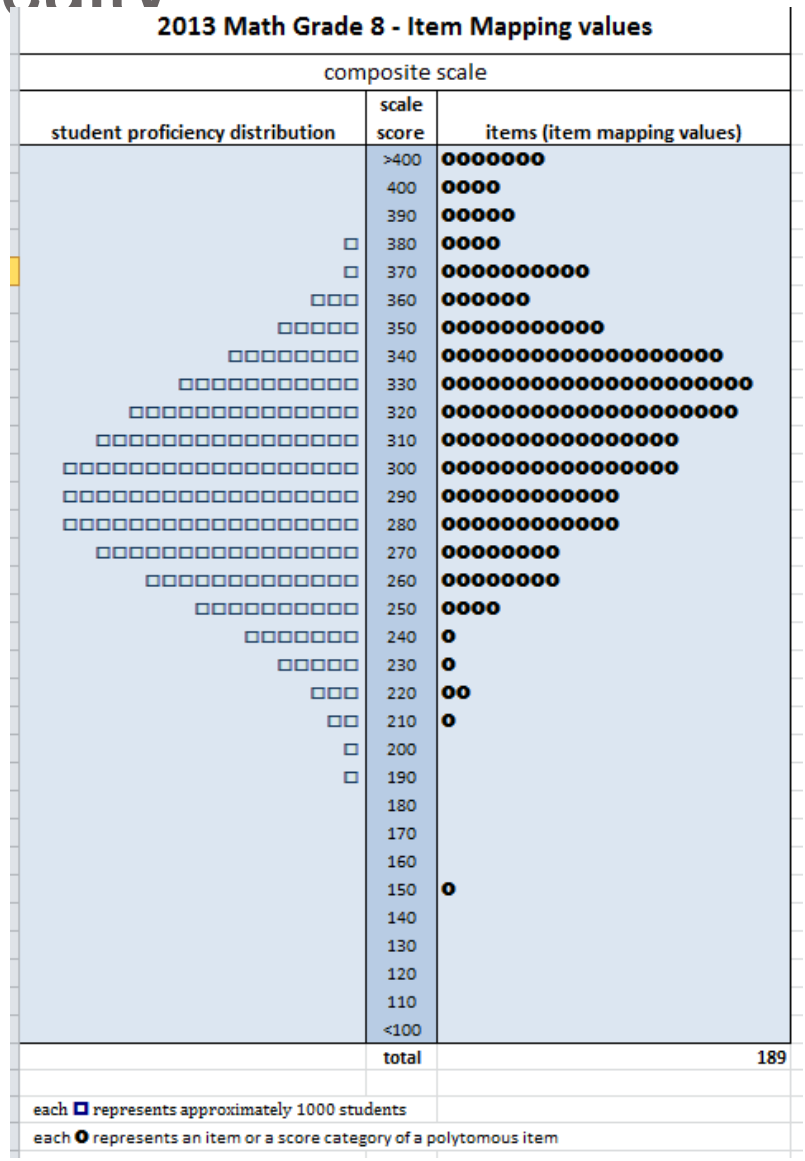
I. Individual items: Difficulty

- Target distribution of item difficulty
  - Approximately uniform
  - Non-uniform:
    - At where the students are
    - Below where the students are
    - Above where the students are
    - More dense at selected points



I. Individual items: Difficulty

- Target distribution of item difficulty
  - Approximately uniform
  - Non-uniform:
    - At where the students are
    - Below where the students are
    - Above where the students are
    - More dense at selected points



2013 Math Grade 8 - Item Mapping values

composite scale

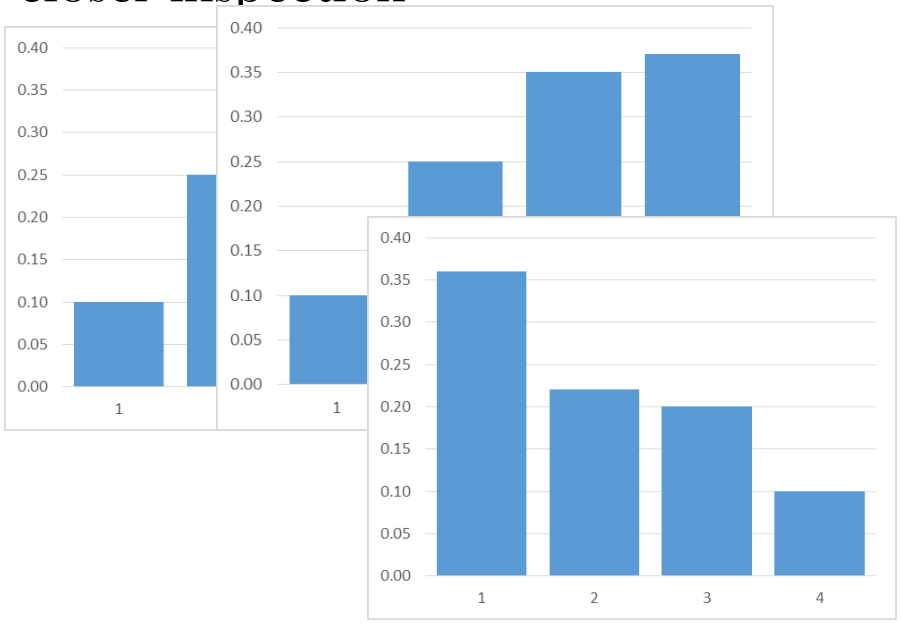
student proficiency distribution	scale score	items (item mapping values)
	>400	oooooooo
	400	oooo
	390	ooooo
□	380	oooo
□	370	ooooooooooo
□□□	360	oooooo
□□□□□	350	oooooooooooo
□□□□□□□□	340	oooooooooooooooooo
□□□□□□□□□□□	330	ooooooooooooooooooooo
□□□□□□□□□□□□□	320	oooooooooooooooooo
□□□□□□□□□□□□□□□	310	ooooooooooooooo
□□□□□□□□□□□□□□□□□	300	ooooooooooooooo
□□□□□□□□□□□□□□□□□	290	oooooooooooo
□□□□□□□□□□□□□□□□□	280	oooooooooooo
□□□□□□□□□□□□□□□□	270	oooooooo
□□□□□□□□□□□□□□	260	oooooooo
□□□□□□□□□□	250	oooo
□□□□□□□	240	o
□□□□□	230	o
□□□	220	oo
□□	210	o
□	200	
□	190	
	180	
	170	
	160	
	150	o
	140	
	130	
	120	
	110	
	<100	
	total	189

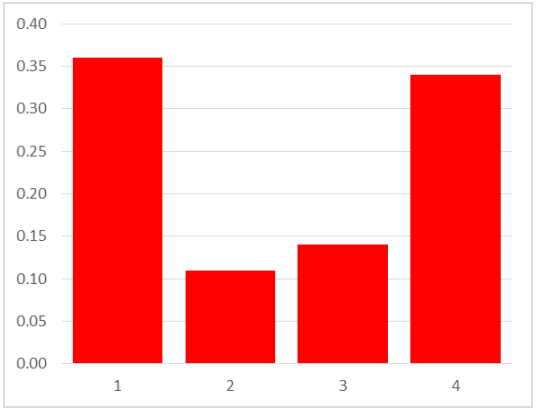
each □ represents approximately 1000 students

each o represents an item or a score category of a polytomous item

I. Individual items: Response distribution

- A "U-shaped" distribution may indicate an issue and warrants a closer inspection





I. Individual items: Response distribution

- A "U-shaped" distribution may indicate an issue and warrants a closer inspection

- For items with many possible responses (e.g., multiple select, drag and drop, etc.), most frequent responses should be looked at.

I. Individual items: Response distribution

- For items with many possible constructed responses (e.g., multiple select, drag and drop, etc.), frequency distributions presenting most common responses should be looked at.

Part A

What is the meaning of **soiled** as it is used in paragraph 2?

A. Involved
B. Damaged
C. Emphasized
D. Identified

Part B

Which two phrases help the reader understand the meaning of soiled?

A. " … starting a rumor." (paragraph 1)
B. " … I was only joking." (paragraph 1)
C. " … my good name!" (paragraph 2)
D. I'll take back …" (paragraph 3)
E. " … take away my guilt" (paragraph 1)
F. " … understand her crime" (paragraph 3)

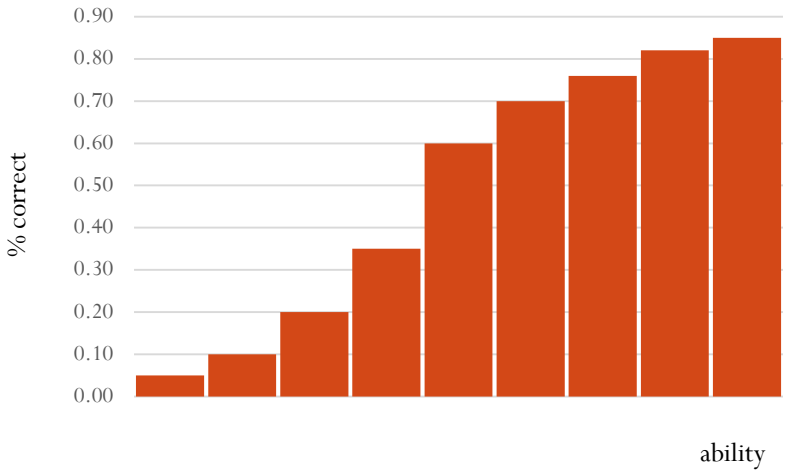
Part A	Part B	%
B	A, E	.18
B	A	.09
B	C	.08
B	C, E	.09
C	C, F	.07
C	A, B	.06
D	C, F	.06
D	F, F	.04
D	E, F	.03

I. Individual items: Discrimination

- Discrimination: how well the item separates low and high ability students.
- Between 0 and 1 - the higher the better
- Rule of thumb: Item to total score correlation >. 10

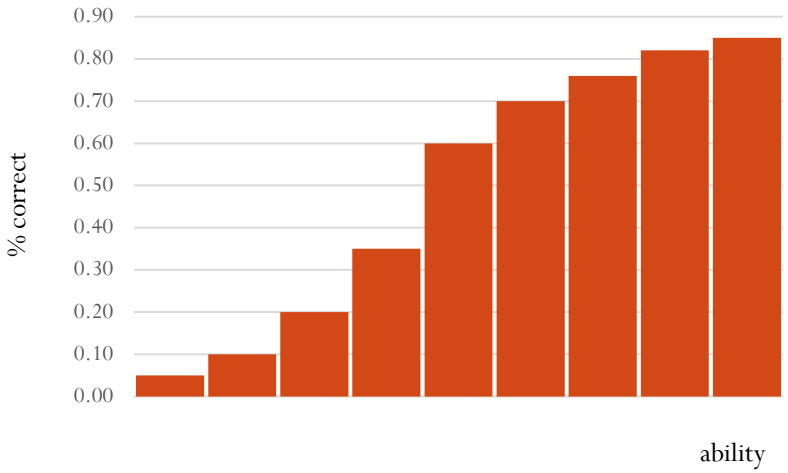
I. Individual items: Discrimination

- Discrimination: how well the item separates low and high ability students.
- Between 0 and 1 - the higher the better
- Rule of thumb: Item to total score correlation >. 10



I. Individual items: Discrimination

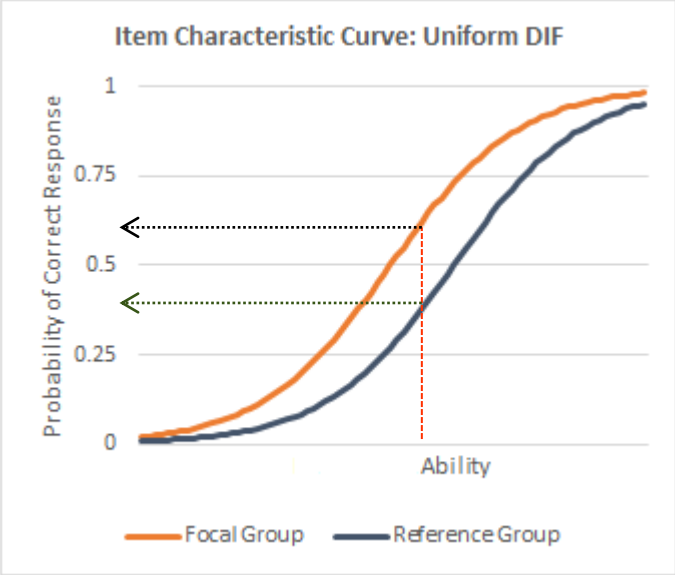
- Discrimination: how well the item separates low and high ability students.
- Between -1 and 1, the higher the better
- Rule of thumb: Item to total score correlation >. 10



- Distractor correlations must be negative or zero

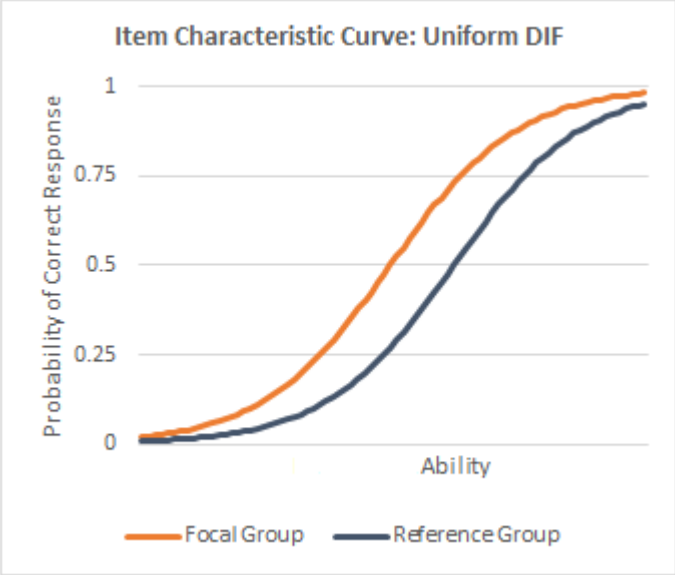
I. Individual items: DIF

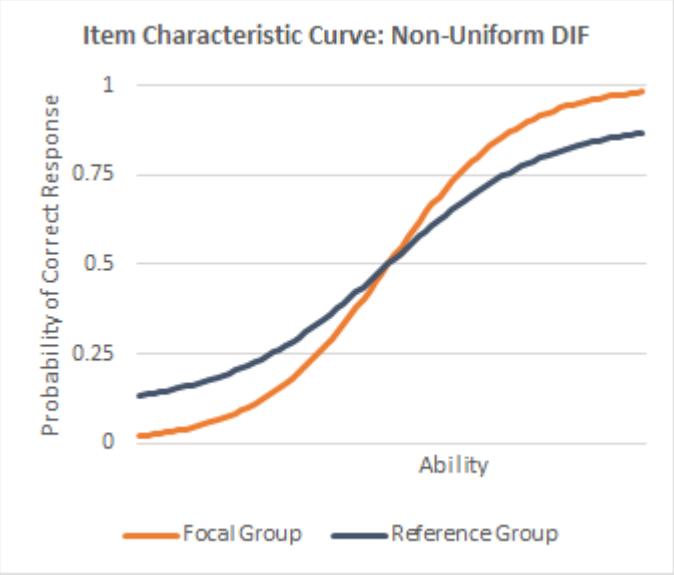
- DIF: Do students of same ability from two different groups have the same chance of correctly answering the given item?



I. Individual items: DIF

- DIF: Do students of same ability from two different groups have the same chance of correctly answering the given item?





I. Individual items: DIF

- DIF
  - Define groups (Reference and Focal) in advance
  - Specify method and criteria
  - Have a process to decide what to do with DIF items

I. Individual items: Data review cards

Item ID	Grade	Subject	Paper or Electronic	Form ID	Position	Item type	Standard	Calculator	Flag
XX001	4	M	E	Y123	3	CR	AB.c1	Y	Red

Item statistics			
N	Max score	p-value	Biserial
120	2	0.4	0.59

Score distribution				
	0	1	2	Omit
Percent	45%	30%	25%	3%
Mean Raw score	7.19	13.4	21.2	8

DIF analysis					
	Male/ Female	White/ Black	White/ Hispanic	Without/ With Disability	Non-ELL/ ELL
Focal Group N	600	800	800	900	850
Reference Group N	600	400	300	200	250
DIF flag?	No	No	No	Yes	No
Favored Group	-	-	-	Without Disability	-

Mode analysis			
	N	p-value	DIF?
Paper	900	0.38	No
Electronic	1200	0.40	

I. Individual items: Data review cards

Administration

Form Name	Use Function	Rptg Flag	Seq	Period	Year	Session	Calc	Model/Ext	Grade
[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	1	Yes		HS

Traditional Statistics

N	P-Val	Mean	Item Total Corr
[illegible]	0.34		0.10

Fit Statistics

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit
9.9	9.9	1.28	1.18				

IRT Statistics

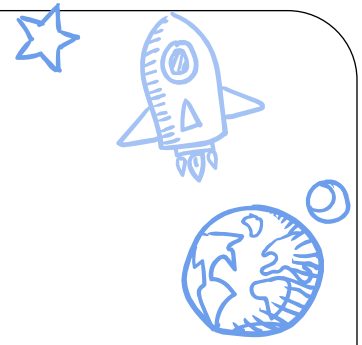
Label	Final	Final S.E.	Preliminary	Preliminary S.E.
Location	1.39	0.02		

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Threshold
A*	0.34	0.10		
B	0.24	0.11		
C	0.25	-0.22		
D	0.17	0.01		
MULTS	0.00			
OMITS	0.00			

DIF Analysis

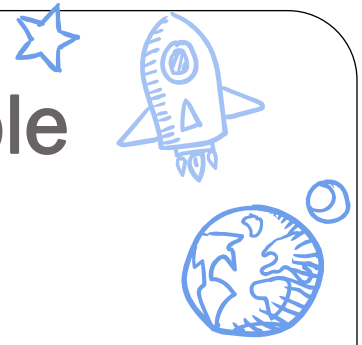
Category	Bias Code	Num Value	N - Ref	N - Focal
MALEFEMALE	A-	-0.13	4709	4550
PAPERONLINE	A+	0.15	8242	1029
WHITEBLACK	A-	-0.23	6812	1245
WHITEHISPANIC	A-	-0.16	6812	726



Psychometric specifications

1. Properties of individual items
2. Properties of forms
   - Linking multiple forms
   - Target measurement precision
3. Properties of the reporting scale

II. Properties of Forms: Linking multiple forms



- Anchor items: common across forms, stitches them together
- Example
  - Form A: Items 1, 2, 3
  - Form B: Items 3, 4, 5

II. Properties of Forms: Linking multiple forms

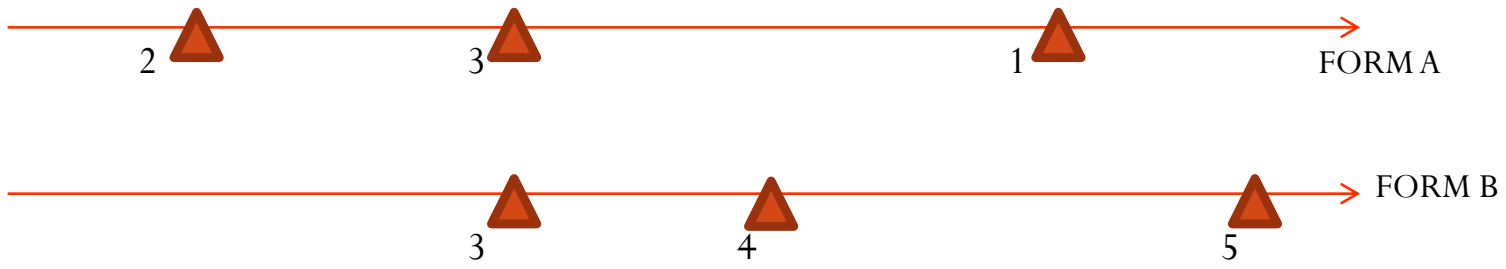
- Anchor items: common across forms, stitches them together
- Example
  - Form A: Items 1, 2, 3
  - Form B: Items 3, 4, 5

II. Properties of Forms: Linking multiple forms

- Anchor items: common across forms, stitches them together
- Example
  - Form A: Items 1, 2, 3
  - Form B: Items 3, 4, 5

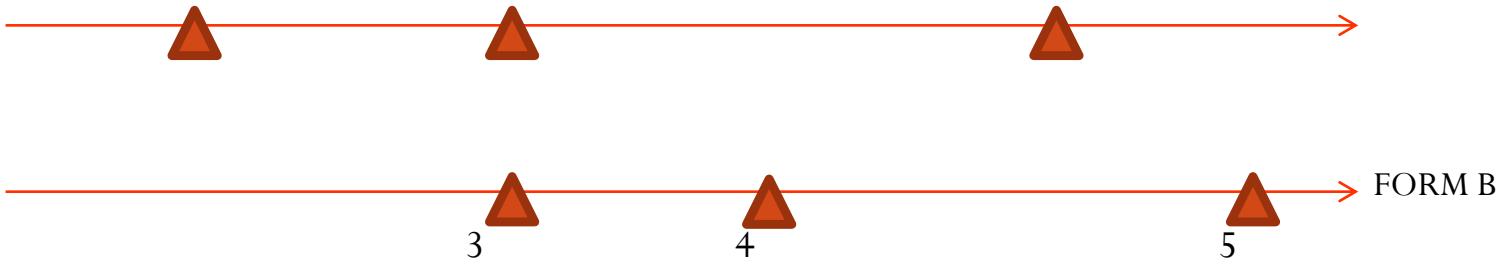
II. Properties of Forms: Linking multiple forms

- Anchor items: common across forms, stitches them together
- Example
  - Form A: Items 1, 2, 3
  - Form B: Items 3, 4, 5



II. Properties of Forms: Linking multiple forms

- Anchor items: common across forms, stitches them together
- Example
  - Form A: Items 1, 2, 3
  - Form B: Items 3, 4, 5



II. Properties of Forms: Linking multiple forms

- Anchor items: common across forms, stitches them together
- Example
  - Form A: Items 1, 2, 3
  - Form B: Items 3, 4, 5

II. Properties of Forms: Linking multiple forms

- External vs. internal anchor items
  - External: Items do not count towards student performance
  - Internal: Items DO count towards student performance

- Example: 9 blue-print sets of items, where Set 9 is divided into eight pieces (9a, …, 9h) and used as external anchor items

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
Unique set	1	2	3	4	5	6	7	8
Linking set I	9a	9a	9b	9b	9c	9c	9d	9d
Linking set II	9e	9f	9f	9g	9g	9h	9h	9e

II. Properties of Forms: Linking multiple forms

- Example: 8 blue-print sets of items, where each is divided into unique and common items

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
Unique set	1u	2u	3u	4u	5u	6u	7u	8u
Linking set I	1c	1c	3c	3c	5c	5c	7c	7c

II. Properties of Forms: Linking multiple forms

- Example: 8 blue-print sets of items, where each is divided into unique and common items

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
Unique set	1u	2u	3u	4u	5u	6u	7u	8u
Linking set I	1c	1c	3c	3c	5c	5c	7c	7c

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
Unique set	1u	2u	3u	4u	5u	6u	7u	8u
Linking set II	8c	2c	2c	4c	4c	6c	6c	8c

II. Properties of Forms: Linking multiple forms

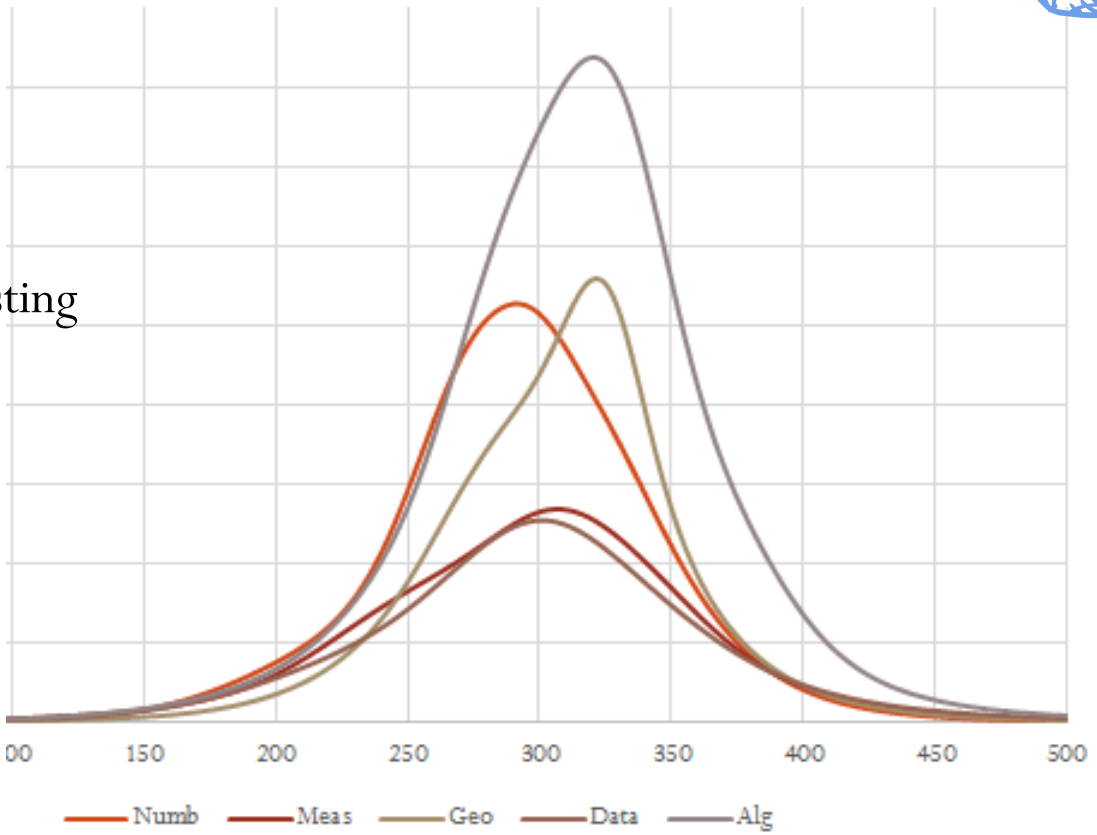
- Example: 8 blue-print sets of items, where each is divided into unique and common items

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
Unique set	1u	2u	3u	4u	5u	6u	7u	8u
Linking set I	1c	1c	3c	3c	5c	5c	7c	7c
Linking set II	8c	2c	2c	4c	4c	6c	6c	8c

- 1c is internal to Form 1, external to Form 2
- 1u+ 1c = full blue print

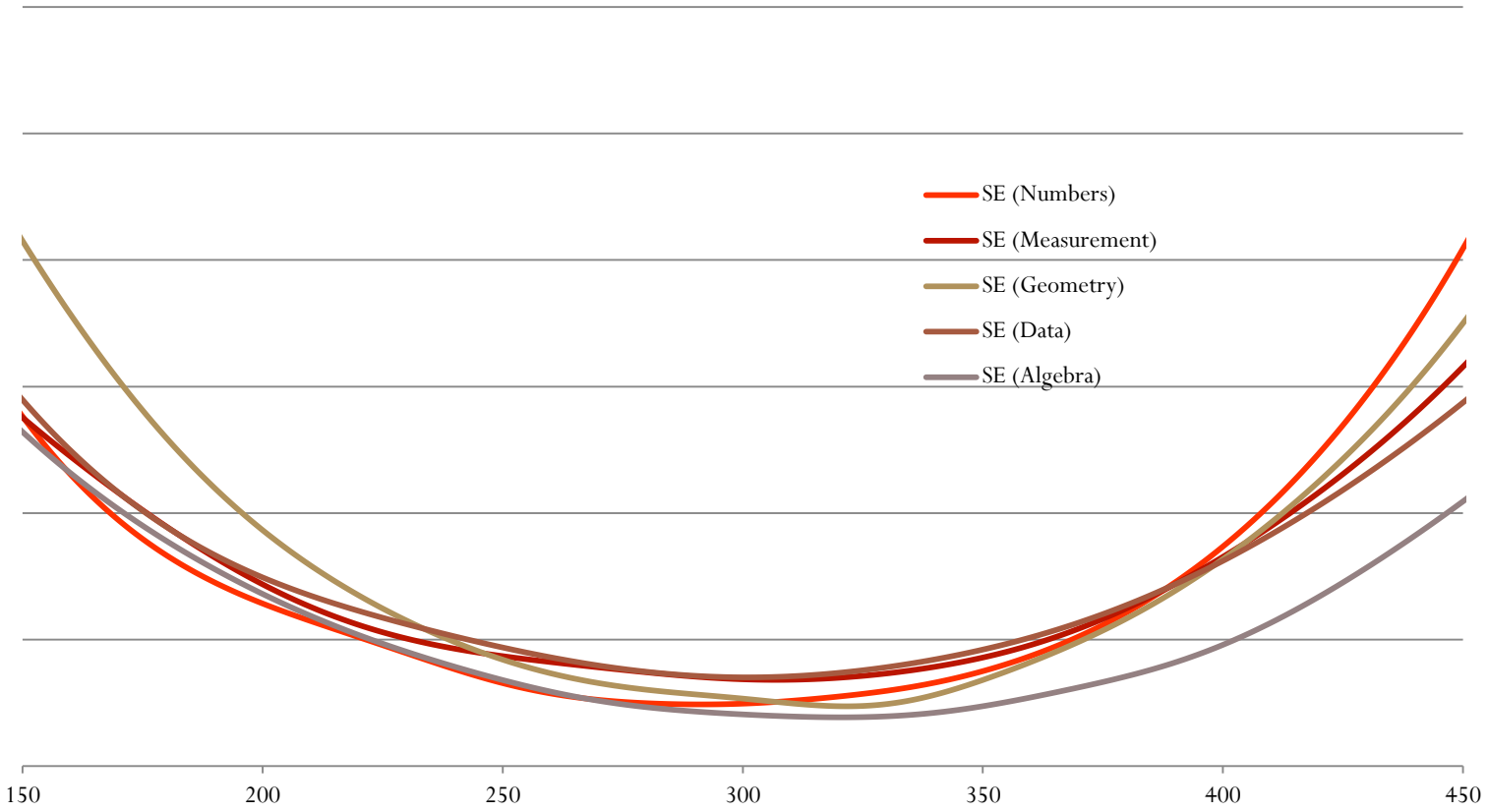
II. Properties of Forms: Precision

- Three factors:
  - Quality of items
  - Test length
  - Linear vs. adaptive testing



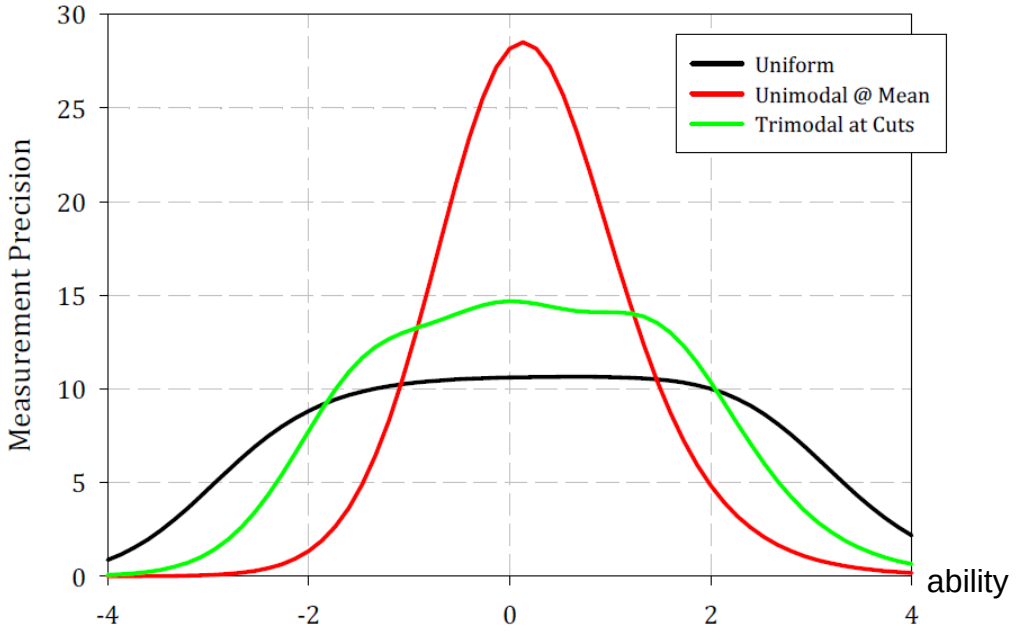
II. Properties of Forms: Precision

- CSEM



How to decide on a target TIF?

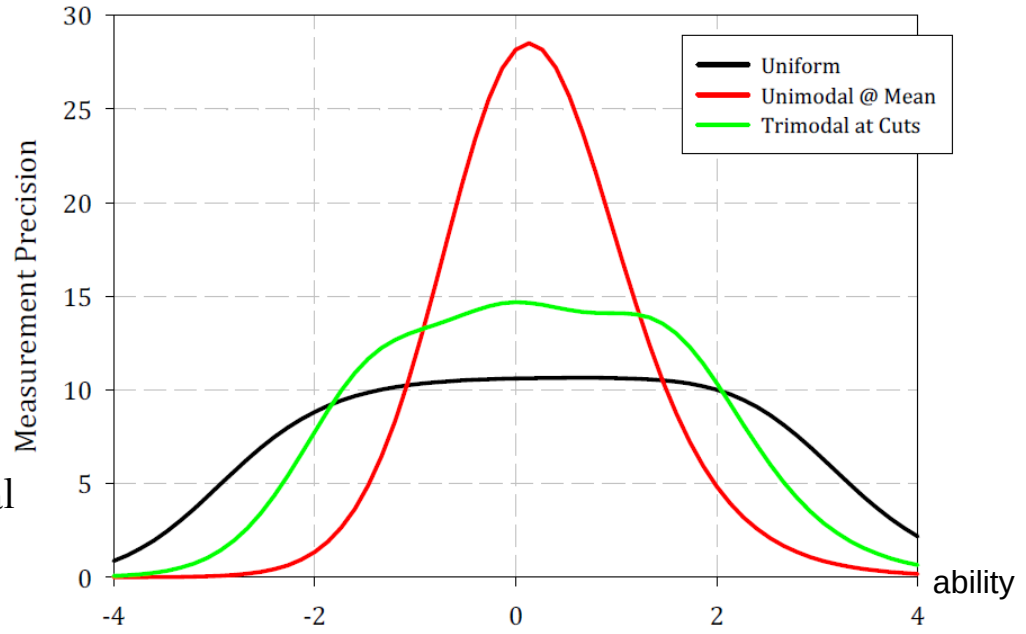
- Choices:
  - Relatively uniform
  - Peaks at cut points
  - Mirrors ability distribution



SOURCE: Luecht, R. M. (2011)

How to decide on a target TIF?

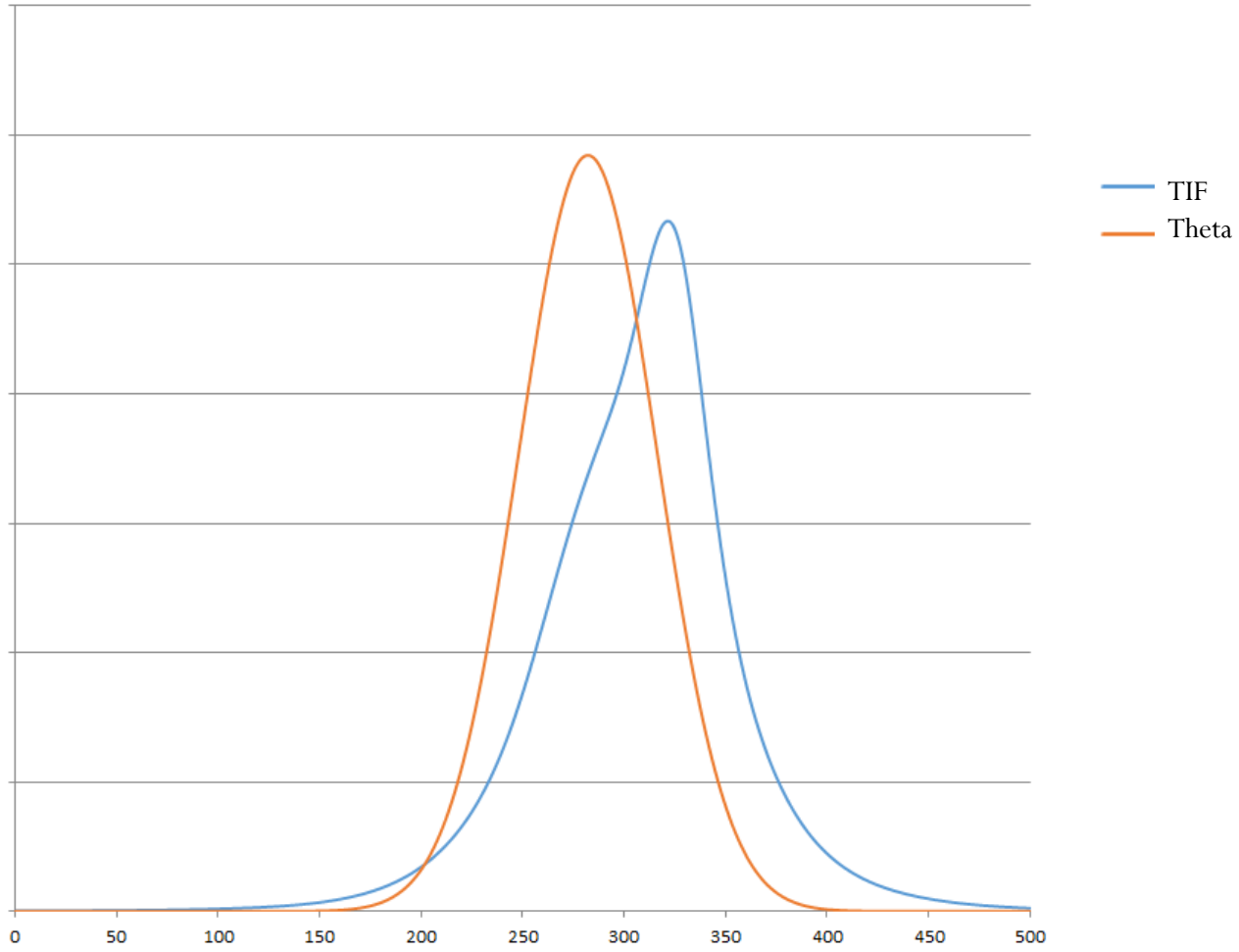
- Choices:
  - Relatively uniform
  - Peaks at cut points
  - Mirrors ability distribution
- Depends on priorities:
  - Classification
  - Measuring all ability continuum with relatively equal precision



SOURCE: Luecht, R. M. (2011)

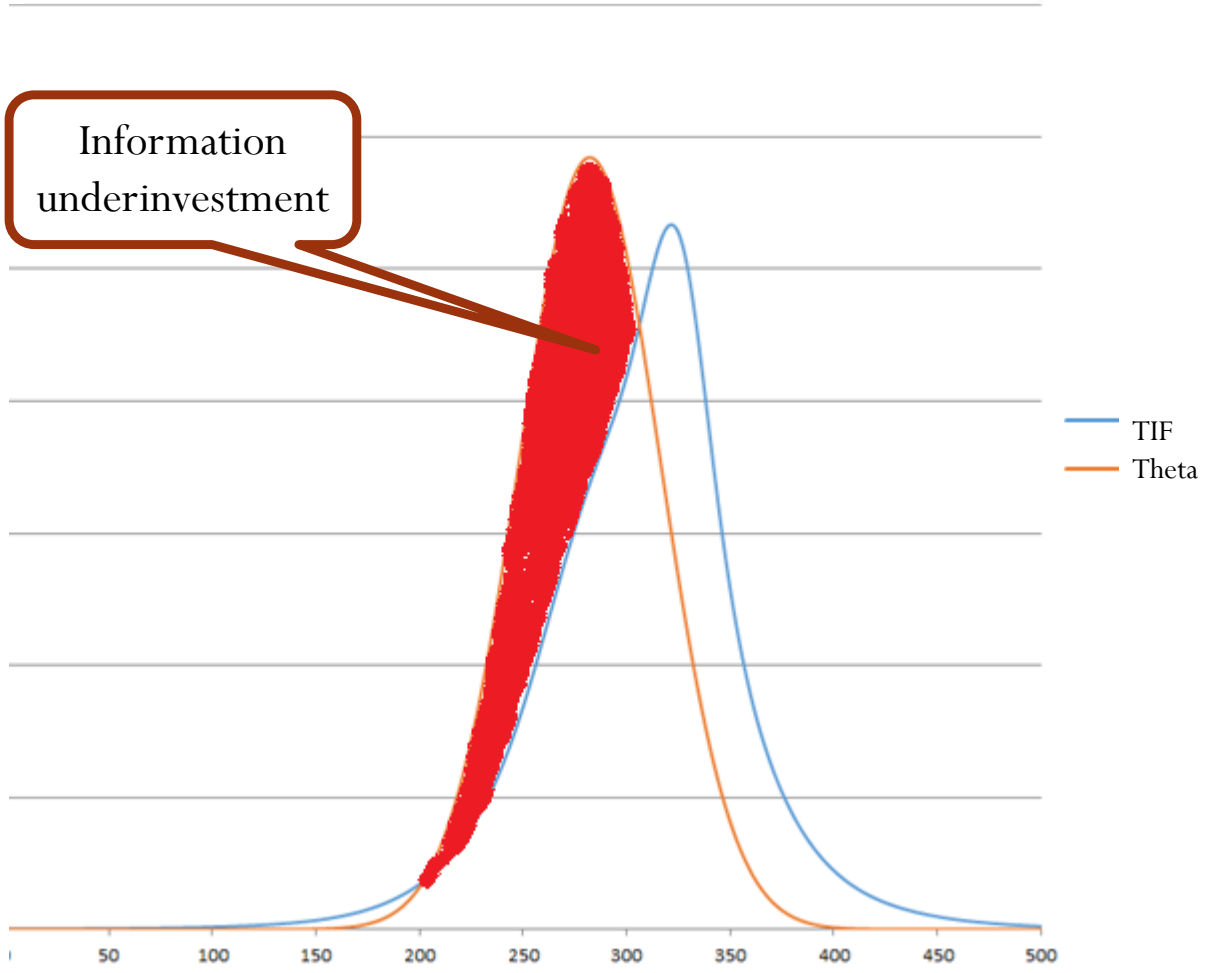
II. Properties of Forms: Precision

- Gap between actual and target TIFs



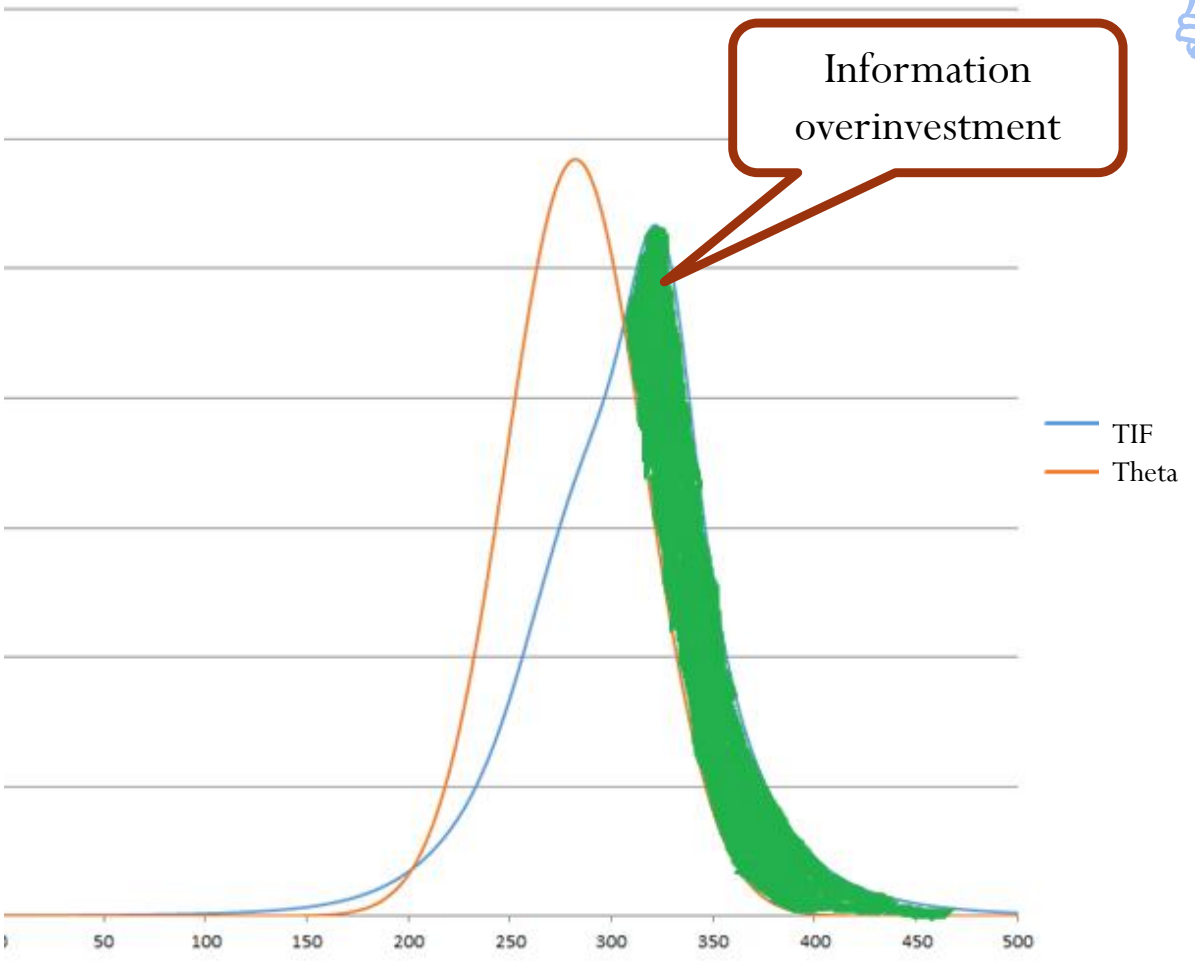
II. Properties of Forms: Precision

- Gap between actual and target TIFs



II. Properties of Forms: Precision

- Gap between actual and target TIFs



III. Scale properties

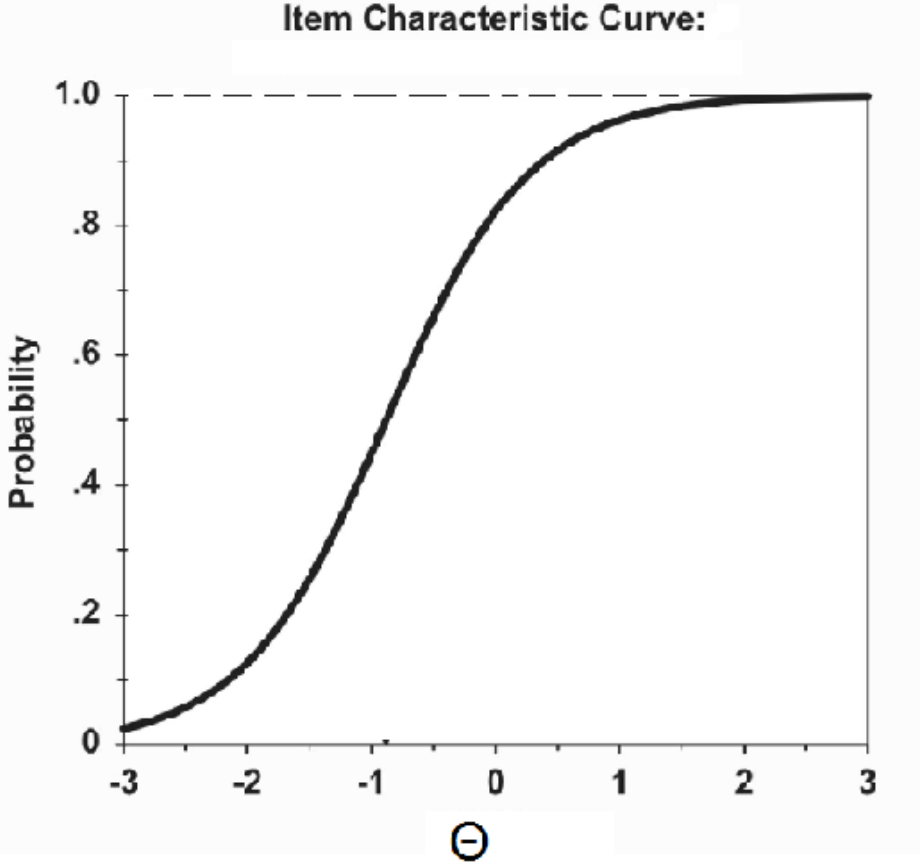
- Dimensional structure and subscores
  - Calibrate subscale items to same metric or not?
  - Overall scale: weighted composite or not?
    - Weighting options:
      - Number of score points
      - Proportional to reliability
      - Policy/content based weights

III. Scale properties

- Range of scale score points
  - Decide if subscales and overall scale will have the same range
  - Decide if a cut score will be kept the same across grades/subjects
  - Avoid scales that might be confused with other scales
  - Avoid scales that might suggest the scores are more precise than they actually are
  - Avoid scales with negative numbers and decimals

III. Scale properties: IRT or not?

- IRT offers powerful tools to create an item bank with all items calibrated to the same scale.

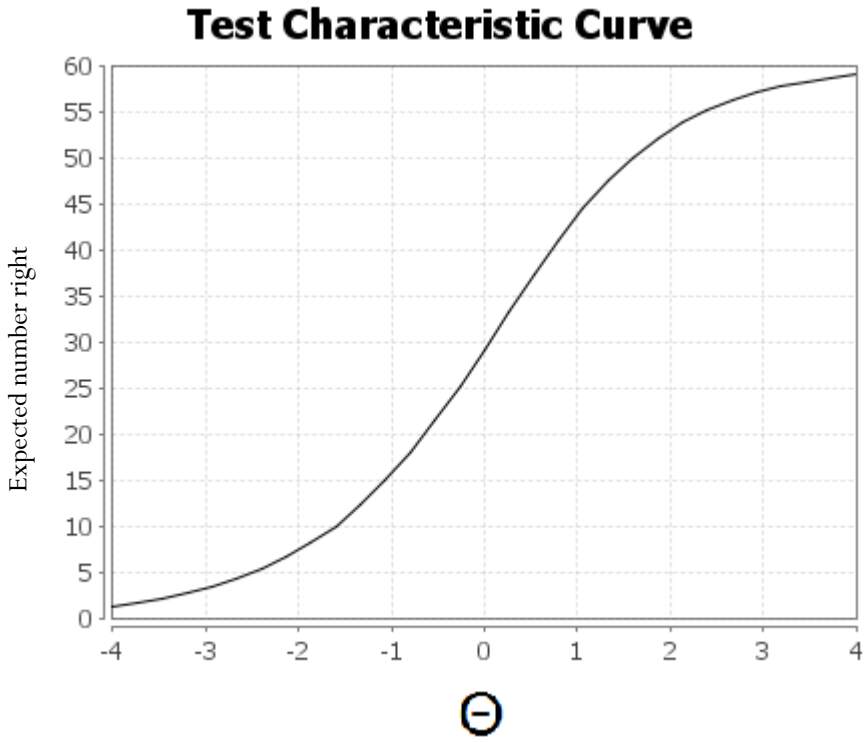


III. Scale properties: IRT or not?

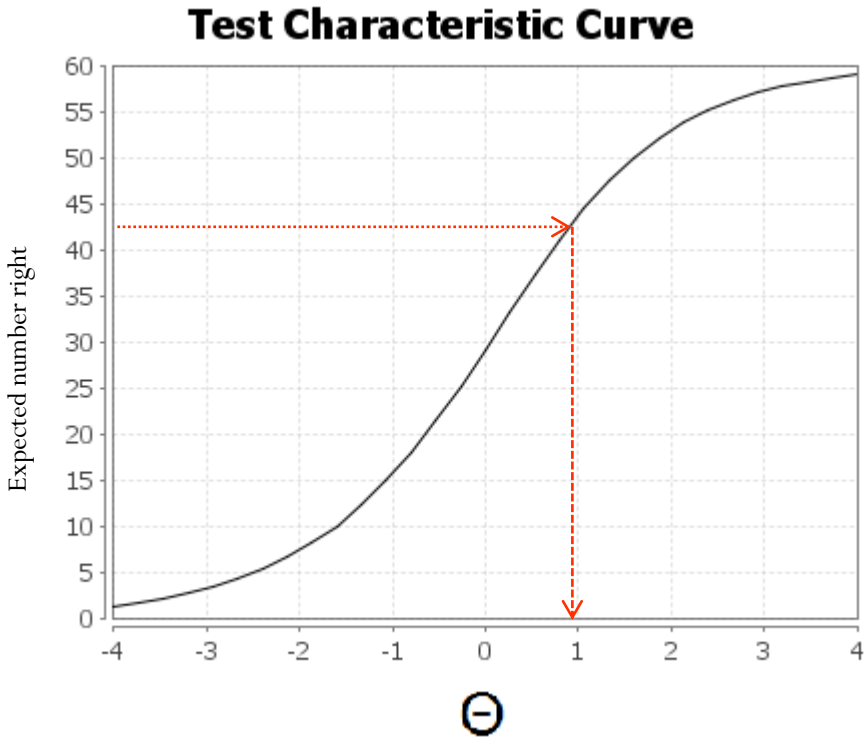
- IRT can also be used for scoring purposes ; known as pattern scoring
  - Not just how many items are answered correctly but also which items are answered correctly
- Model choice:
  - Dichotomously scored items: Rasch, 2PL, 3PL
  - Polytomously scored items: Partial Credit, GPC
- No need to specify model in advance; do specify/ask for methods to choose a model, and corresponding criteria.

III. Scale properties: IRT or not?

- Some programs use IRT for calibration purpose only and use TCC for scoring



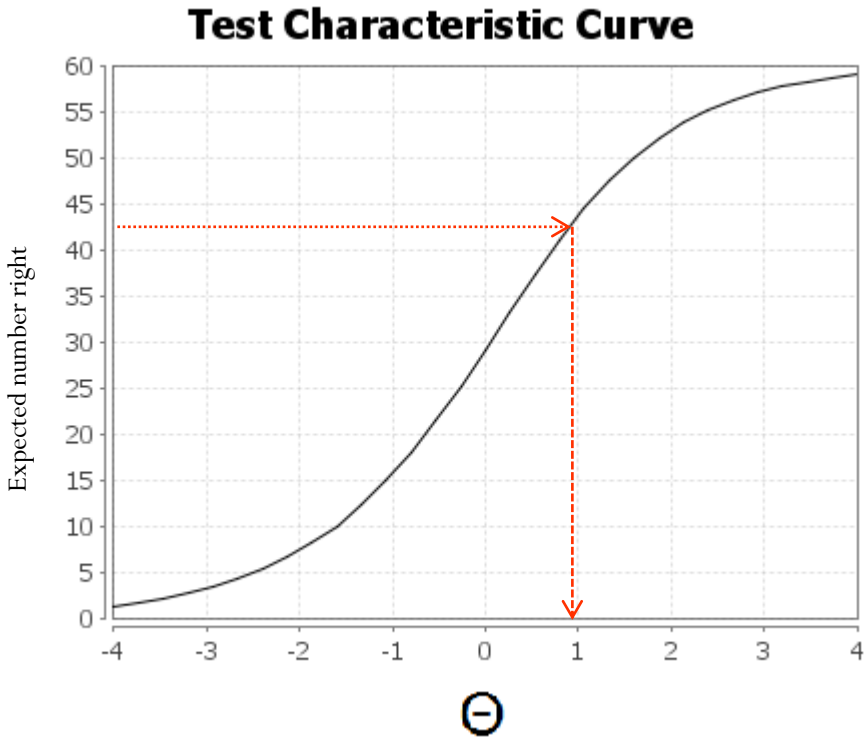
III. Scale properties: IRT or not?



- TCCs are form specific.
- Ensures every student with the same "total score" ends up with the same scale score.

Total score	Θ
…	…
42	0.95
43	**1.00**
44	1.06
45	1.07
…	…

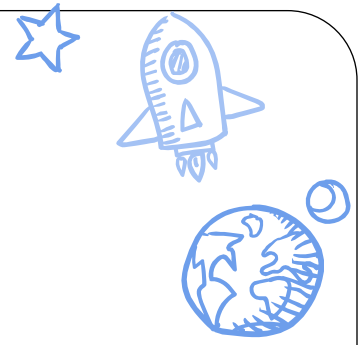
III. Scale properties: IRT or not?



- TCCs are form specific.
- Ensures every student with the same "total score" ends up with the same scale score.
- Still allows score comparability since Θ is common across forms.

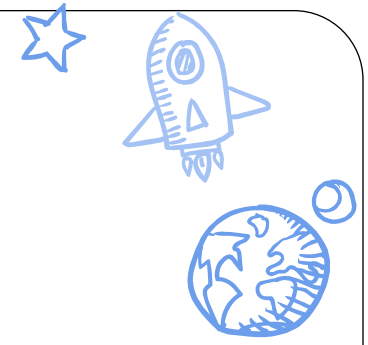
III. Scale properties

- Whether pattern scoring is used or not, scores on the Θ metric needs to be transformed to the reporting metric
- Many choices here too:
  - Linear transformation
    - Scale score = A + B * Θ (as in y = a + bx)
    - Shape of score distribution and Test Information remains the same
  - Nonlinear transformation
    - Will change shape of score distribution and Test Information



Recommendations

- Start with reporting specification
- Take your time in planning
- Have content and measurement people talk with each other (sooner than later)
- Develop and maintain a psychometric roadmap and a decisions log
- No need to specify every detail, but need to know when/how each decision will be made



References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Luecht, R. M. (2011). An investigation of statistical test design and delivery options relative to score precision for the PARCC assessments. Presentation at PARCC Technical Advisory Committee meeting. Washington DC.

THANK YOU

Any questions?

Enis.Dogan@ed.gov