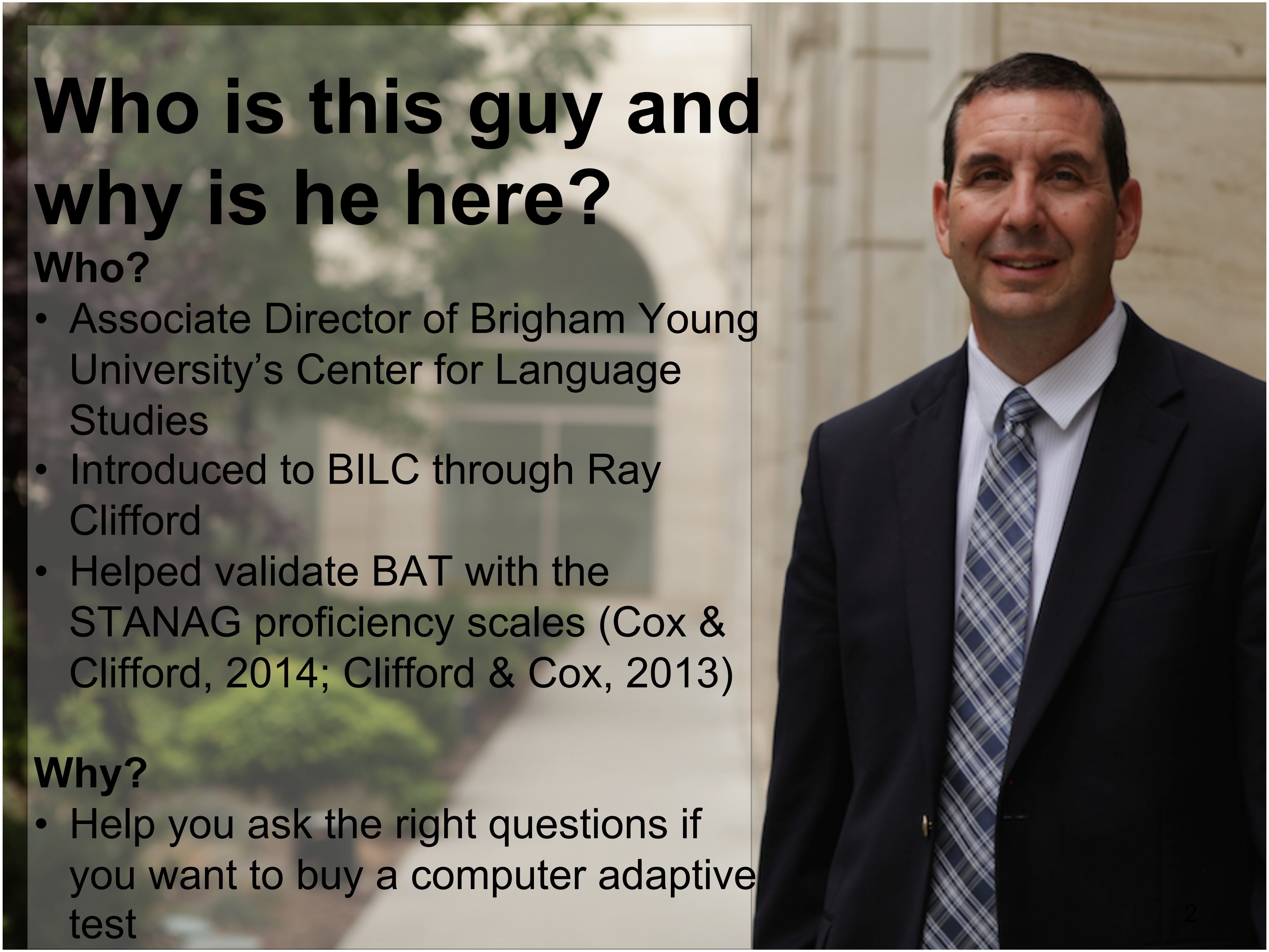
Making a Computer Adaptive Test: What Shirts, Tunnels and Hurdles (and Bikes) Teach Us

Troy L. Cox, PhD
Ray T. Clifford, PhD
Brigham Young University
Provo, UT

Who is this guy and why is he here?



Who?

- Associate Director of Brigham Young University's Center for Language Studies
- Introduced to BILC through Ray Clifford
- Helped validate BAT with the STANAG proficiency scales (Cox & Clifford, 2014; Clifford & Cox, 2013)

Why?

- Help you ask the right questions if you want to buy a computer adaptive test

CAT: Formal Definition

A computer-assisted, sequential form of testing in which successive items in the test are chosen based on the responses to previous items.

(Source: Concise Encyclopedia of Psychology, 2nd Ed.)

Why use a computer adaptive test (CAT)?

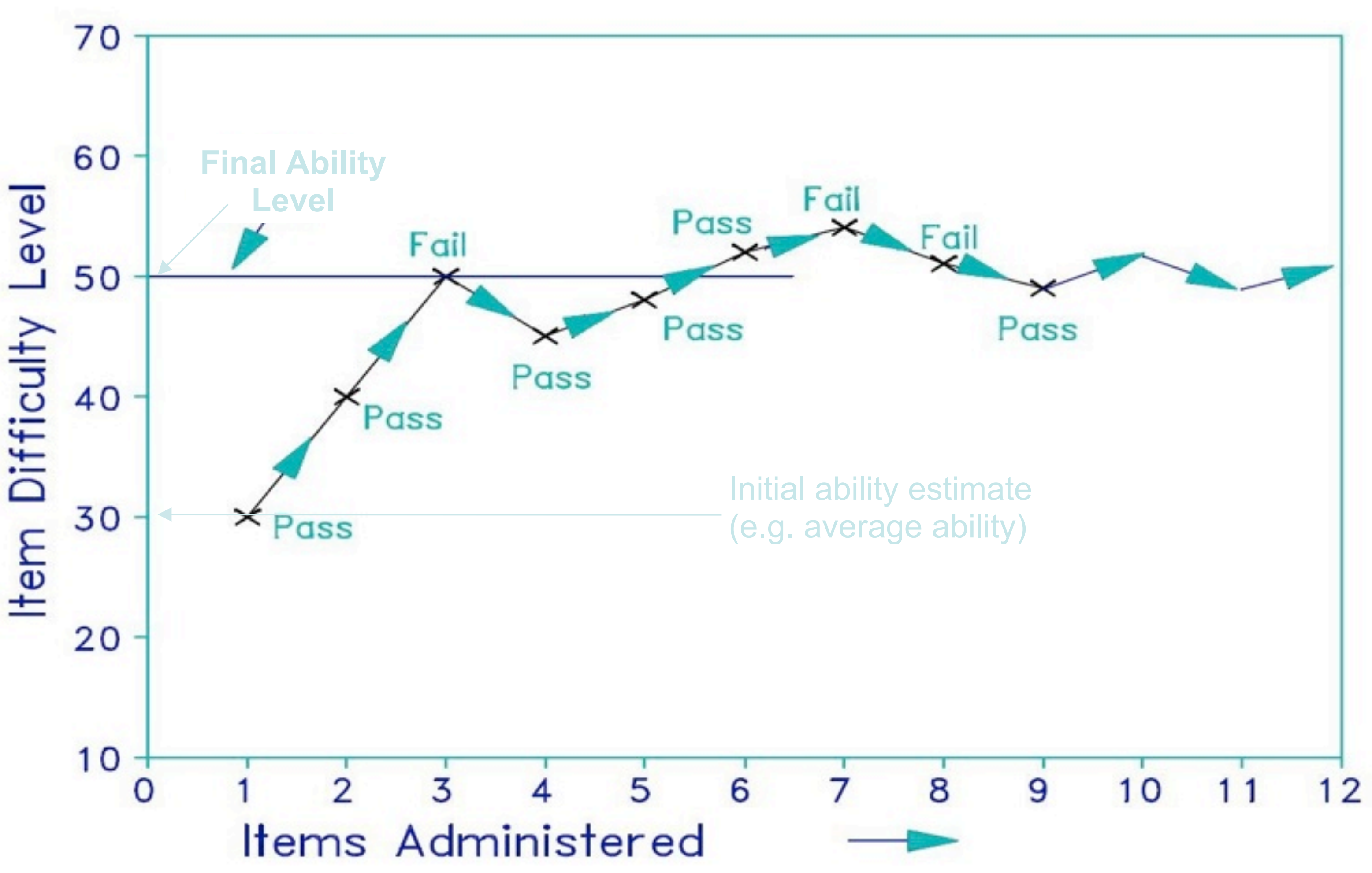
- Test can be shorter (with smaller Standard Errors than CTT)
- Avoids the use of too easy/difficult items
- Test security can be increased
  - Item Exposure
  - Cheating
- Tests are individually paced
- Can provide accurate measures over a wide range of abilities.
- Test experience is enjoyable and can improve individual performance

Stages of CAT

- Select Initial Item(s)
  - Item with midrange difficulty
  - Small range of items with varying difficulties
- Calculate examinee ability estimate
- Present item with difficulty level near examinee ability level
  - Item Bank Needed
- Stop Test
  - Standard error reaches predetermined level
  - Time

CAT Ability Scoring Example

Figure 1. Dichotomous CAT Test Administration.



Source: Linacre, John Michael (2000)

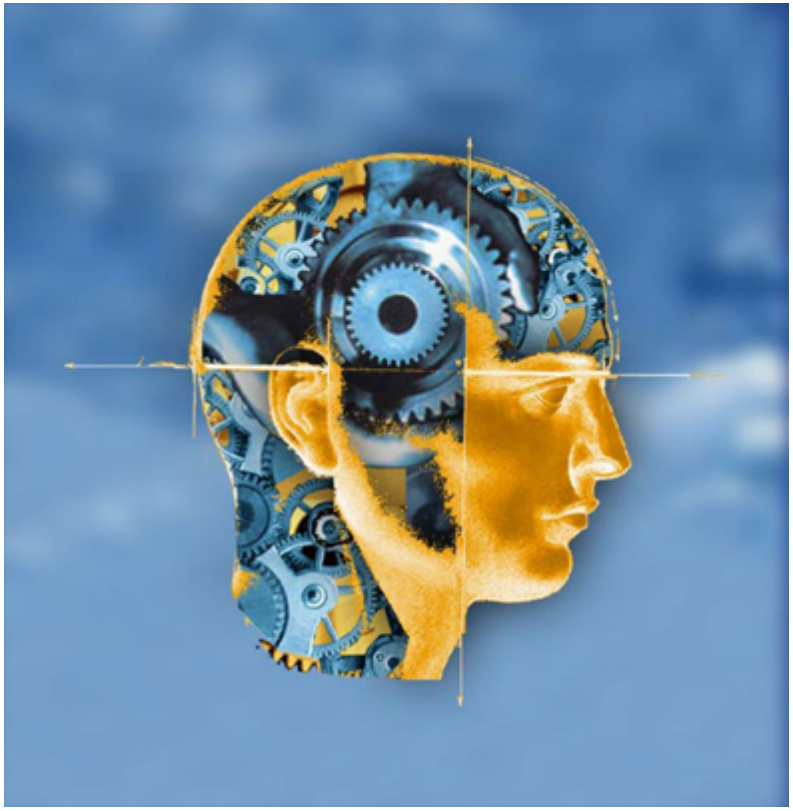
© Cox, 2014

Making a CAT: The Recipe

- Computer
  - Programming
  - Equipment
- Adaptive
  - Algorithm
- Test
  - Psychometrics

Psychometrics

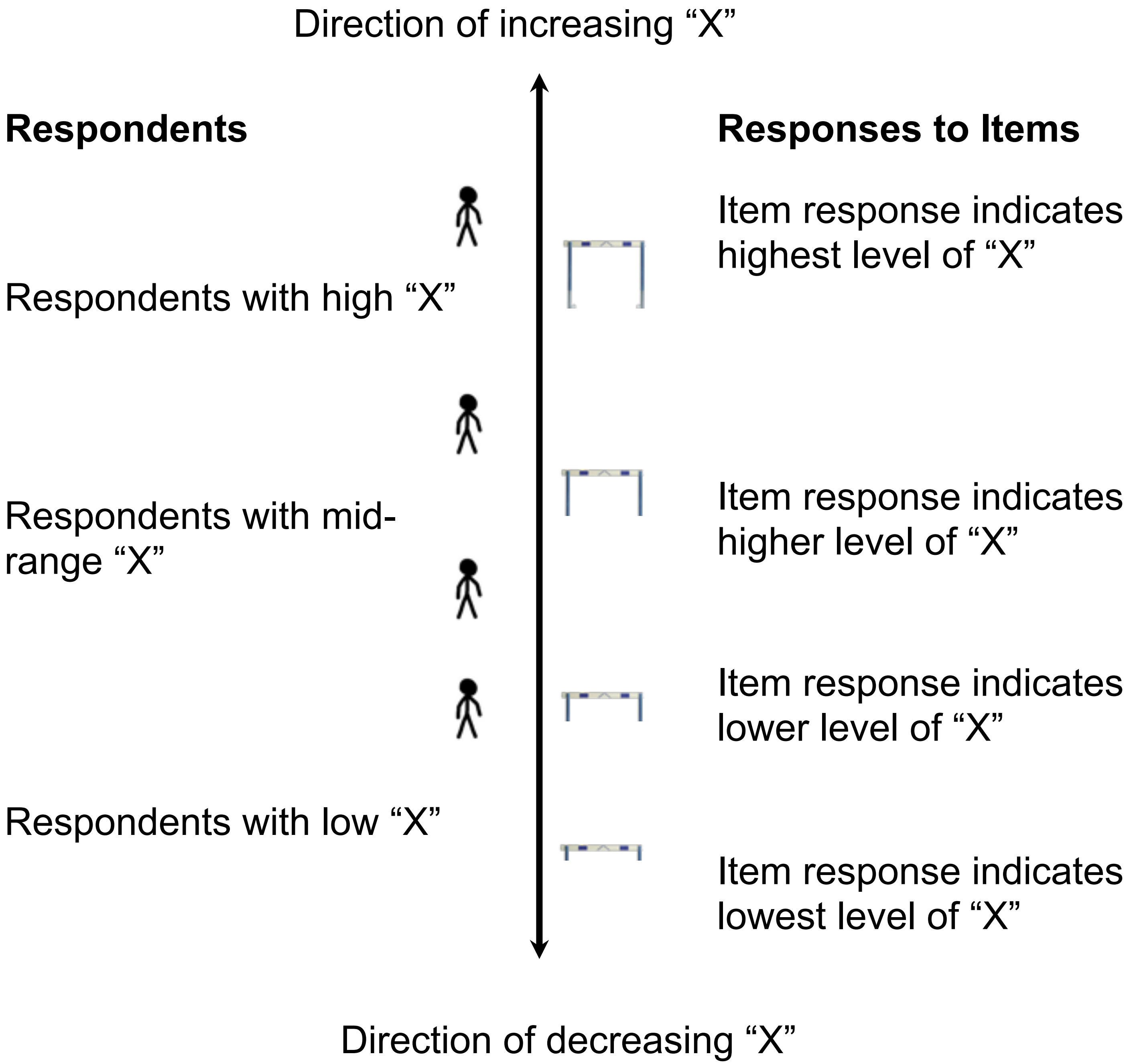
- Psych—Mind
- Metric—Measurement



So, what are we measuring?

- Construct
  - Our theoretical object of interest

- The instrument is always secondary.
  - What is the purpose?
  - What is the context?



Direction of increasing “X”

Respondents

Responses to Items

Item response indicates highest level of “X”

Respondents with high “X”

Respondents with mid-range “X”

Item response indicates higher level of “X”

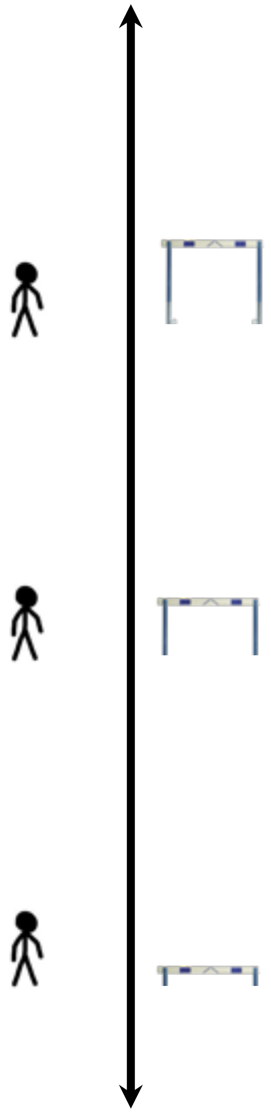
Item response indicates lower level of “X”

Respondents with low “X”

Item response indicates lowest level of “X”

Direction of decreasing “X”

Direction of Increasing Proficiency



Respondents

Responses to Items

Students with STANAG 3 Proficiency

Correct item response indicates .5 probability of having STANAG 3 Proficiency

Students with STANAG 2 Proficiency

Correct item response indicates .5 probability of having STANAG 2 Proficiency

Students with STANAG 1 Proficiency

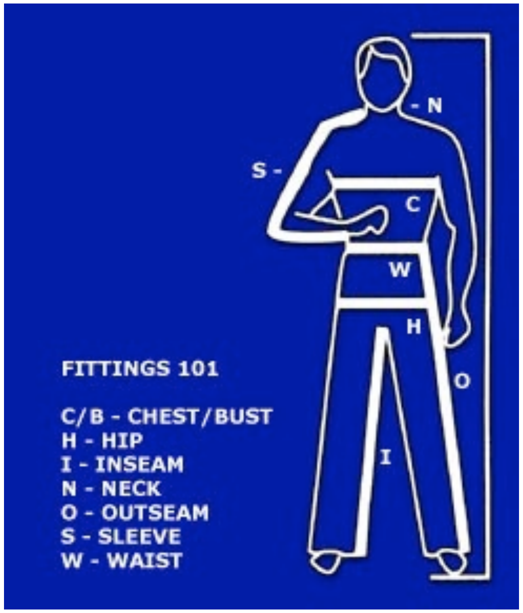
Correct item response indicates .5 probability of having STANAG 1 Proficiency

Direction of Decreasing Proficiency

How many dimensions are we measuring?

- Think of a physical analog

- Measuring for a man’s shirt

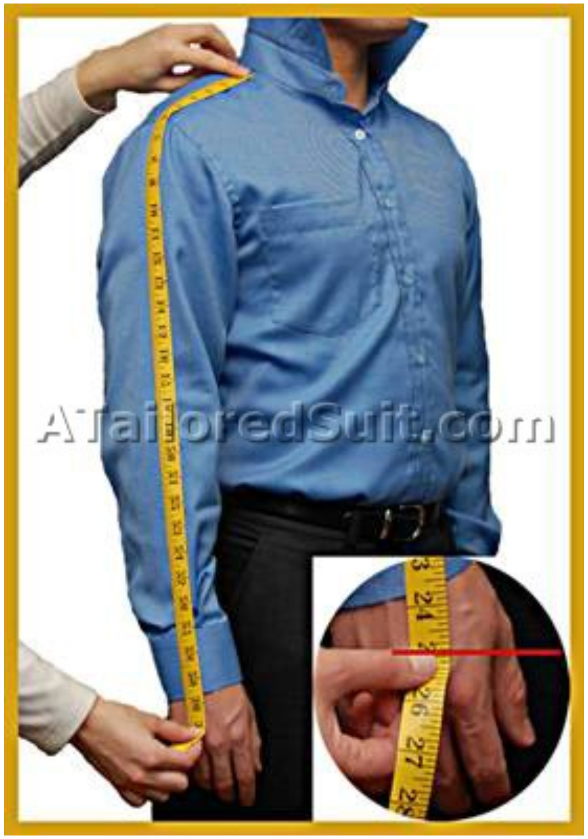


What will we measure?

- Neck?

What will we measure?

- Arm Length?

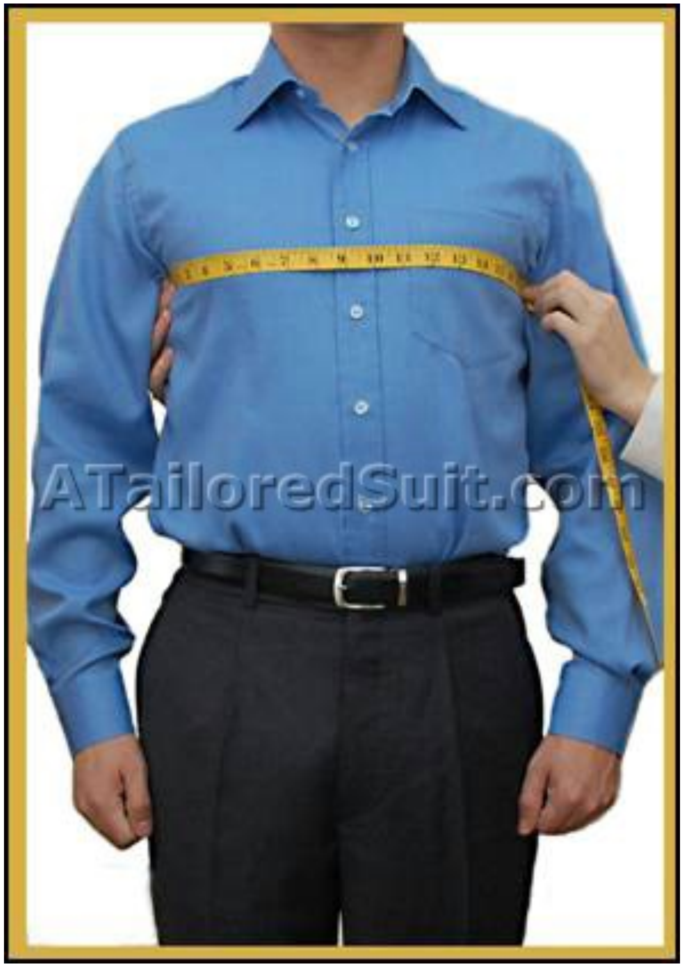


What will we measure?

- Waist/Stomach?

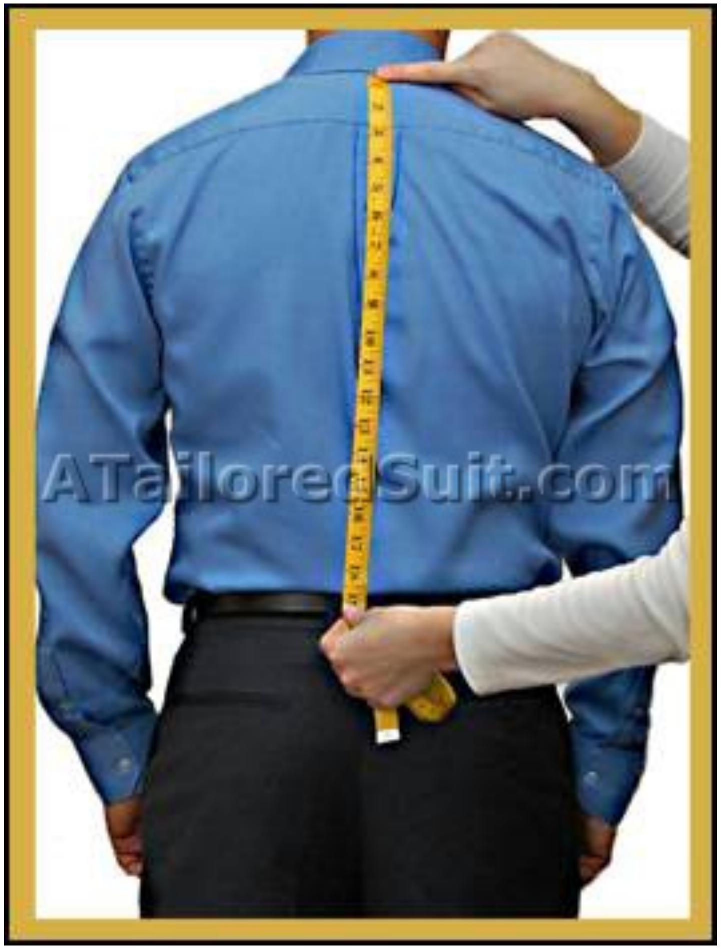
What will we measure?

- Chest?

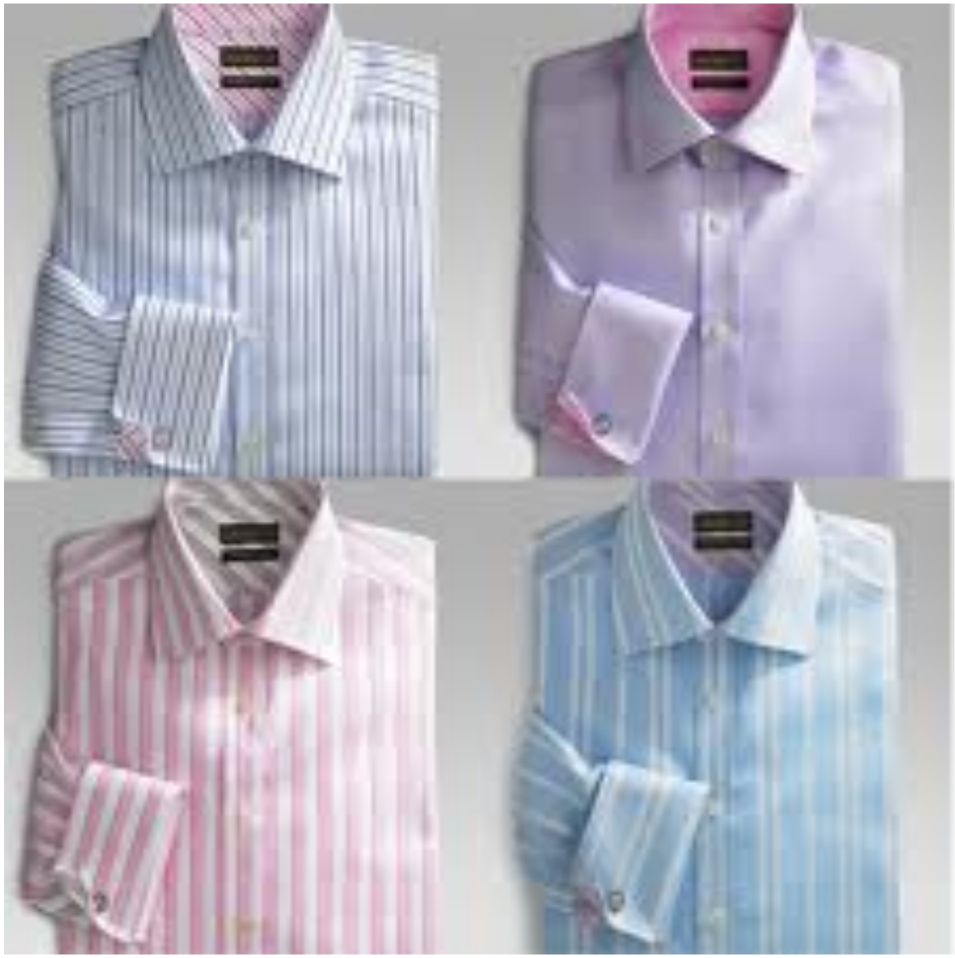


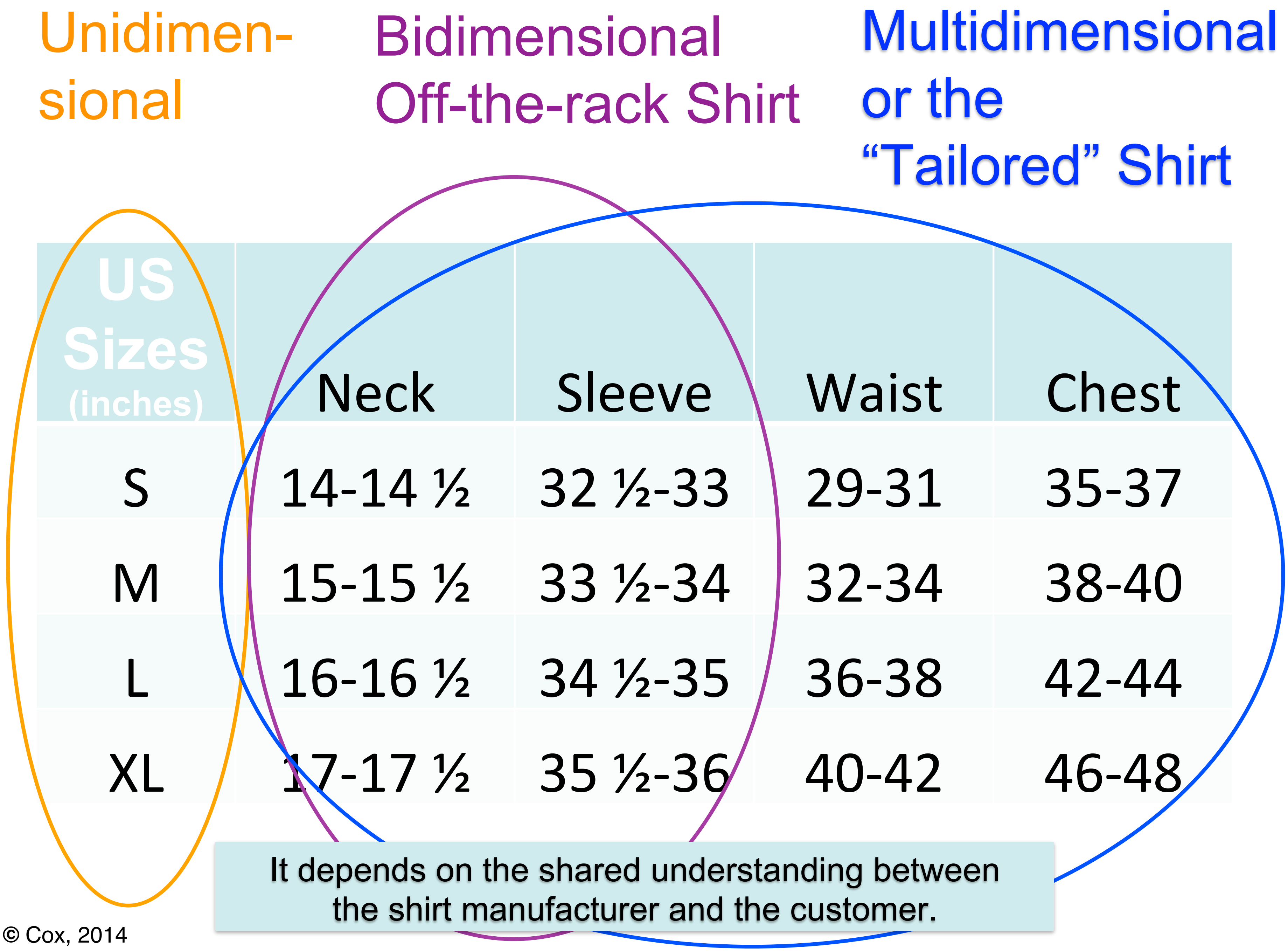
What will we measure?

- Torso Length?



Is this a uni-, bi- or multi-dimensional measurement?





Unidimensional

Bidimensional Off-the-rack Shirt

Multidimensional or the "Tailored" Shirt

US Sizes (inches)	Neck	Sleeve	Waist	Chest
S	14-14 ½	32 ½-33	29-31	35-37
M	15-15 ½	33 ½-34	32-34	38-40
L	16-16 ½	34 ½-35	36-38	42-44
XL	17-17 ½	35 ½-36	40-42	46-48

It depends on the shared understanding between the shirt manufacturer and the customer.

Construct Map

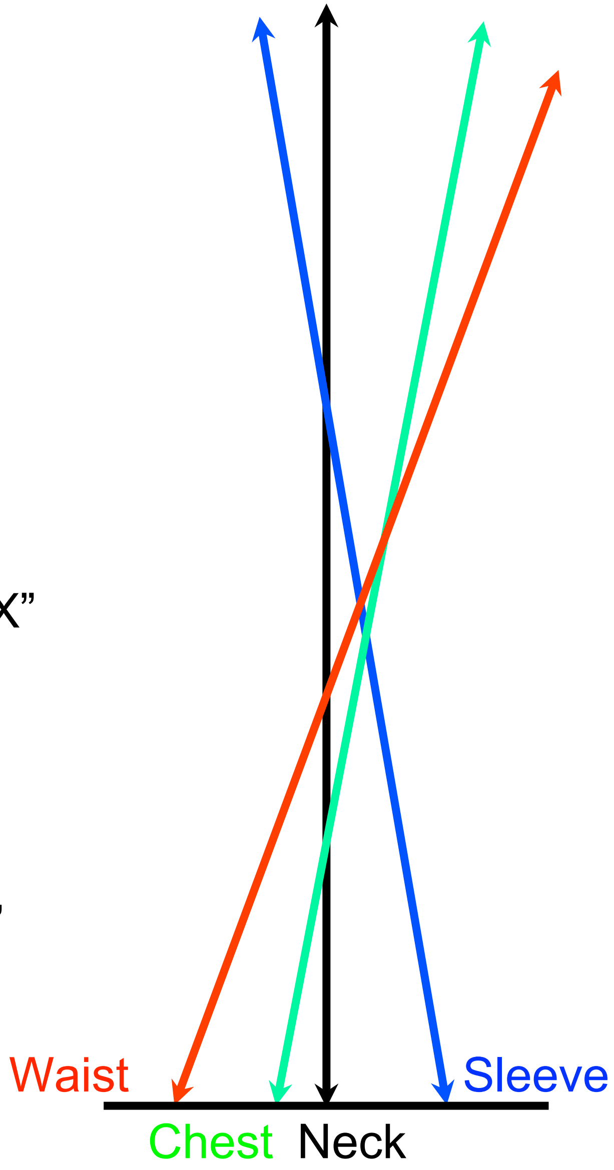
Direction of increasing “X”

Respondents

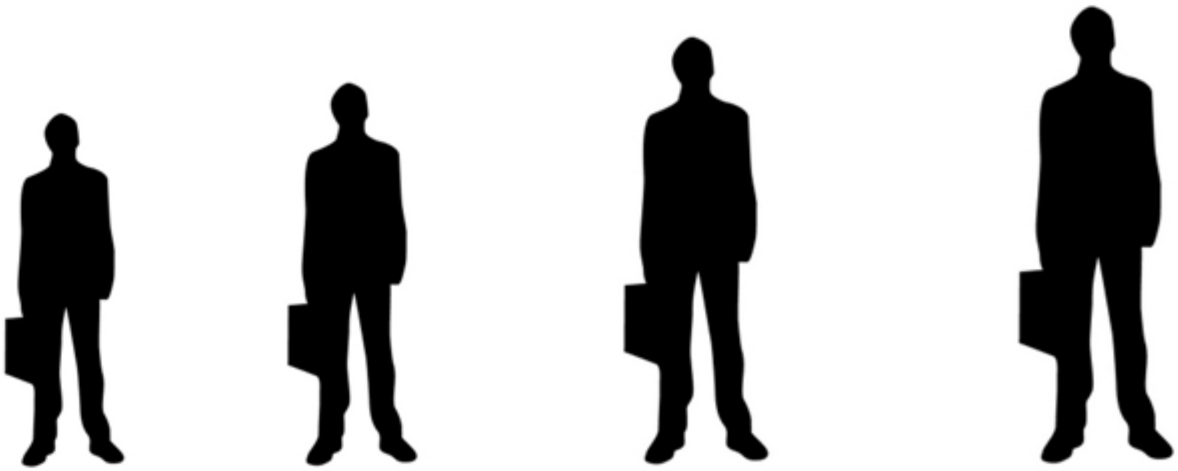
Respondents with large “X”

Respondents with medium “X”

Respondents with small “X”



Direction of decreasing “X”



Neck, waist, sleeve and chest *tend* to co-occur.

US Sizes (inches)	Neck	Sleeve	Waist	Chest
S	14-14 ½	32 ½-33	29-31	35-37
M	15-15 ½	33 ½-34	32-34	38-40
L	16-16 ½	34 ½-35	36-38	42-44
XL	17-17 ½	35 ½-36	40-42	46-48

US Sizes (inches)	Neck	Sleeve	Waist	Chest
S	14-14 ½	32 ½-33	29-31	35-37
M	15-15 ½	33 ½-34	32-34	38-40
L	16-16 ½	34 ½-35	36-38	42-44
XL	17-17 ½	35 ½-36	40-42	46-48

Compensatory
=
Large

US Sizes (inches)	Neck	Sleeve	Waist	Chest
S	14-14 ½	32 ½-33	29-31	35-37
M	15-15 ½	33 ½-34	32-34	38-40
L	16-16 ½	34 ½-35	36-38	42-44
XL	17-17 ½	35 ½-36	40-42	46-48

Problem with Compensatory Measurement

	Bilbo	**Gandalf**	**Orin**
Neck	14 (S)	15 (M)	17 (XL)
Arm Length	32 (S)	33 (M)	36 (XL)
Chest	42 (L)	40 (M)	35 (S)
Total	88	88	88
	S	M	S

Same score, but three very different profiles.

Conjunctive = Small or Extra Large

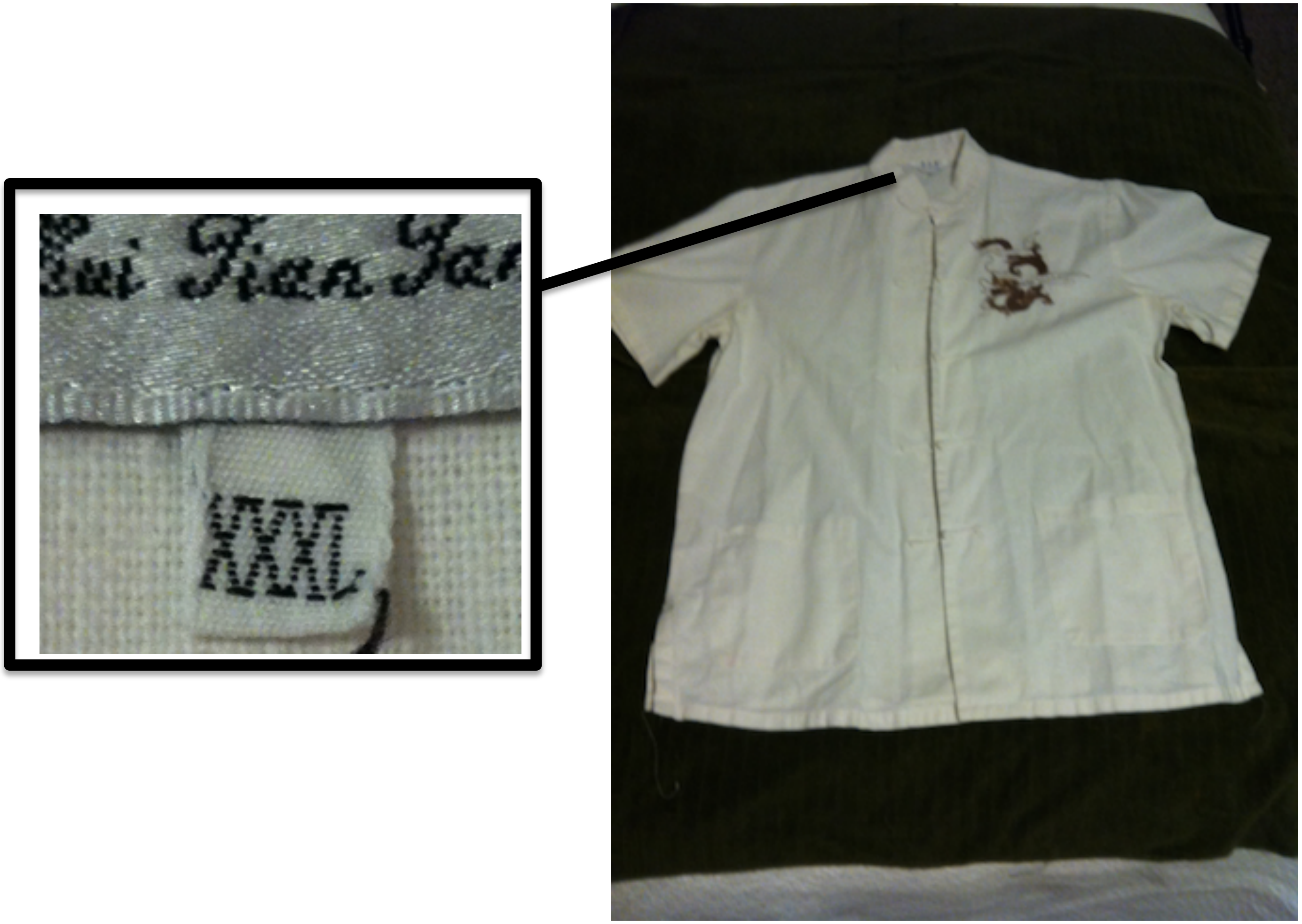
US Sizes (inches)	Neck	Sleeve	Waist	Chest
S	14-14 ½	32 ½-33	29-31	35-37
M	15-15 ½	33 ½-34	32-34	38-40
L	16-16 ½	34 ½-35	36-38	42-44
XL	17-17 ½	35 ½-36	40-42	46-48

Given that consumers will have measurement variation in neck, sleeve, waist and chest size, successful manufactures will ensure that all of their products construct their shirts based on their stated standards.

US Sizes (inches)	Neck	Sleeve	Waist	Chest
S ⟷	14-14 ½	32 ½-33	29-31	35-37
M ⟷	15-15 ½	33 ½-34	32-34	38-40
L ⟷	16-16 ½	34 ½-35	36-38	42-44
XL ⟷	17-17 ½	35 ½-36	40-42	46-48

Sizes are relative to population

Sizes are relative to population



So, what are we measuring?

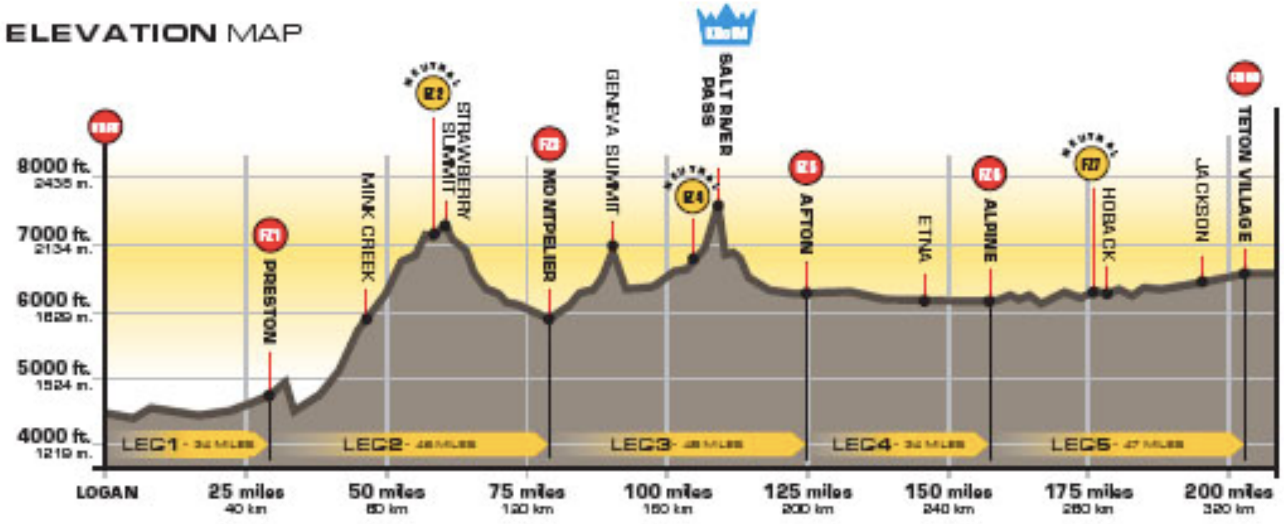
- Construct
  - Make sure you (e.g. the manufacturer) and your end-users (e.g. the customer) have a shared definition of your construct map.
- What's XXXL in one context may simply be L in another.

Criterion-Referenced vs. Norm-Referenced Tests

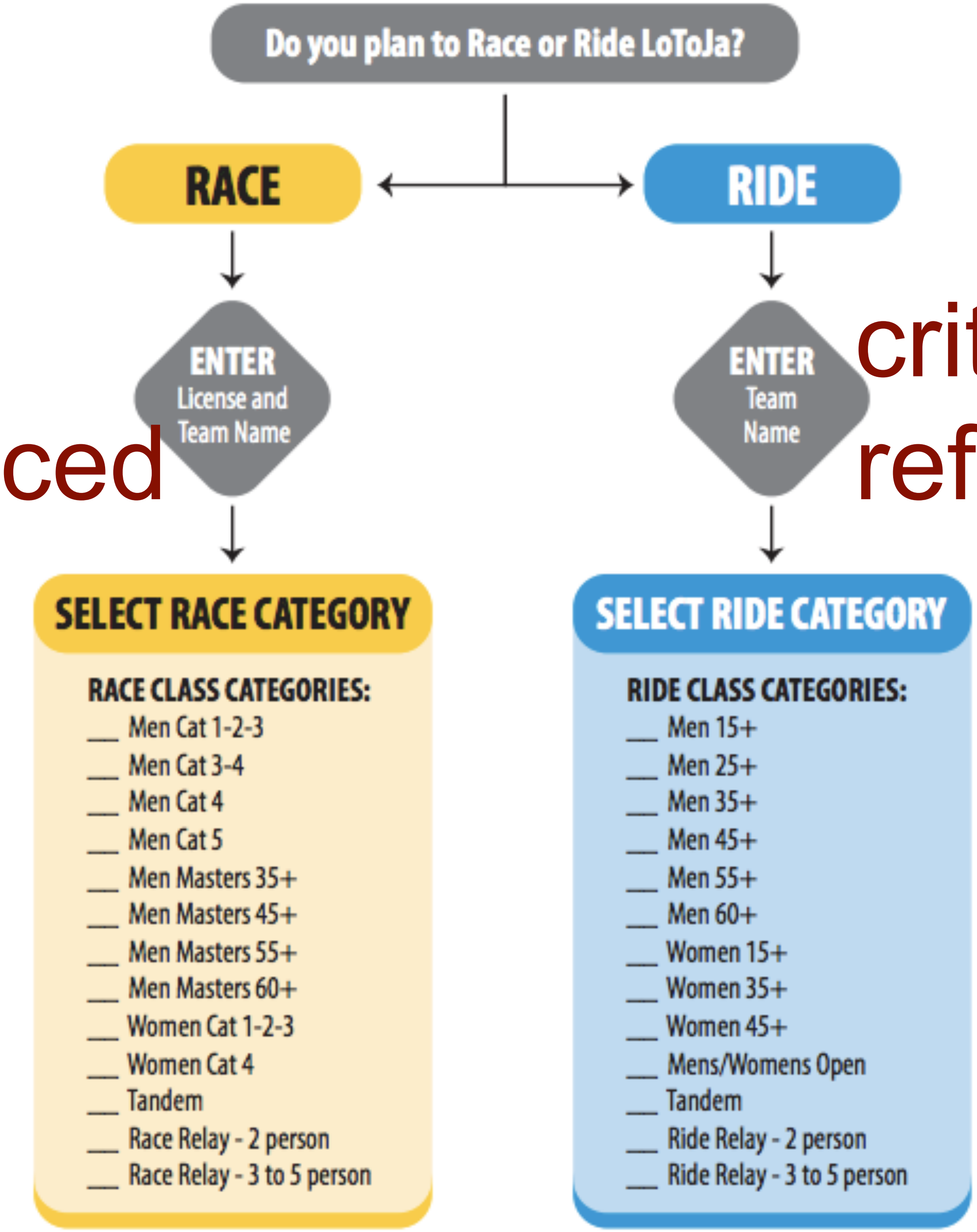
LOTOJA

CLASSIC

The LoToJa Bicycle Classic is a 206-mile (332 km), one-day amateur bicycle road race from **Lo**gan, UT **to Ja**ckson Hole, WY, USA.



Is LOTOJA a criterion or norm-referenced competition?



Do you plan to Race or Ride LoToJa?

RACE

ENTER License and Team Name

SELECT RACE CATEGORY

RACE CLASS CATEGORIES:

___ Men Cat 1-2-3
___ Men Cat 3-4
___ Men Cat 4
___ Men Cat 5
___ Men Masters 35+
___ Men Masters 45+
___ Men Masters 55+
___ Men Masters 60+
___ Women Cat 1-2-3
___ Women Cat 4
___ Tandem
___ Race Relay - 2 person
___ Race Relay - 3 to 5 person

RIDE

ENTER Team Name

SELECT RIDE CATEGORY

RIDE CLASS CATEGORIES:

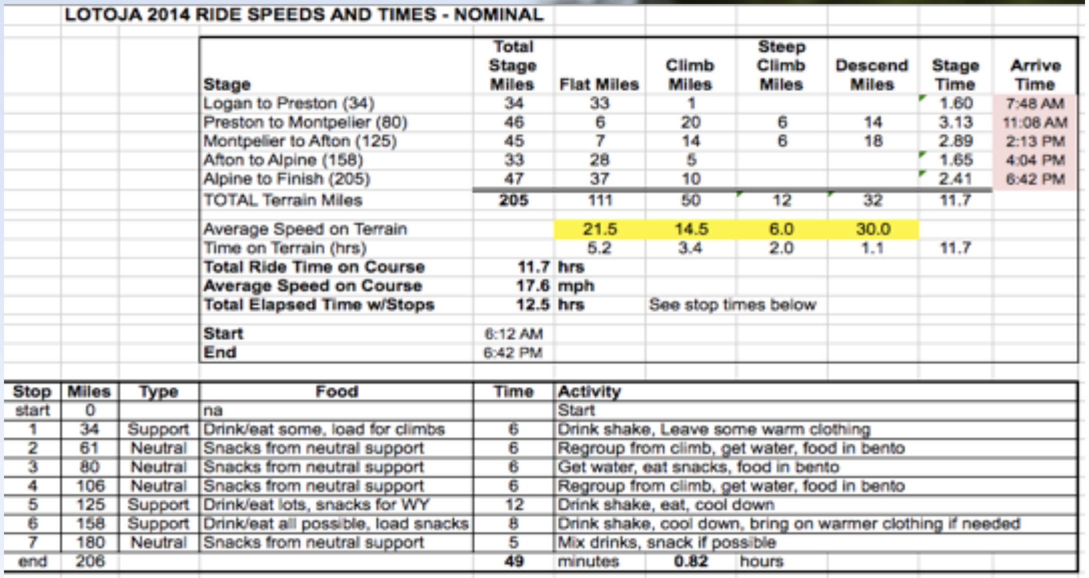
___ Men 15+
___ Men 25+
___ Men 35+
___ Men 45+
___ Men 55+
___ Men 60+
___ Women 15+
___ Women 35+
___ Women 45+
___ Mens/Womens Open
___ Tandem
___ Ride Relay - 2 person
___ Ride Relay - 3 to 5 person

norm-referenced

criterion referenced

If you’re competing against others, then it’s norm-referenced.

If you're competing against the clock, then it's criterion-referenced.



LOTOJA 2014 RIDE SPEEDS AND TIMES - NOMINAL

Stage	Total Stage Miles	Flat Miles	Climb Miles	Steep Climb Miles	Descend Miles	Stage Time	Arrive Time
Logan to Preston (34)	34	33	1			1.60	7:48 AM
Preston to Montpelier (80)	46	6	20	6	14	3.13	11:08 AM
Montpelier to Afton (125)	45	7	14	6	18	2.89	2:13 PM
Afton to Alpine (158)	33	28	5			1.65	4:04 PM
Alpine to Finish (205)	47	37	10			2.41	6:42 PM
TOTAL Terrain Miles	**205**	111	50	12	32	11.7	
Average Speed on Terrain		21.5	14.5	6.0	30.0		
Time on Terrain (hrs)		5.2	3.4	2.0	1.1	11.7	
Total Ride Time on Course	**11.7**	**hrs**					
Average Speed on Course	**17.6**	**mph**					
Total Elapsed Time w/Stops	**12.5**	**hrs**	See stop times below				
Start	6:12 AM						
End	6:42 PM						

Stop	Miles	Type	Food	Time	Activity
start	0		na		Start
1	34	Support	Drink/eat some, load for climbs	6	Drink shake, Leave some warm clothing
2	61	Neutral	Snacks from neutral support	6	Regroup from climb, get water, food in bento
3	80	Neutral	Snacks from neutral support	6	Get water, eat snacks, food in bento
4	106	Neutral	Snacks from neutral support	6	Regroup from climb, get water, food in bento
5	125	Support	Drink/eat lots, snacks for WY	12	Drink shake, eat, cool down
6	158	Support	Drink/eat all possible, load snacks	8	Drink shake, cool down, bring on warmer clothing if needed
7	180	Neutral	Snacks from neutral support	5	Mix drinks, snack if possible
end	206			**49**	minutes **0.82** hours

Do you know the secret to enjoying your job?

- Have a hobby that's even **worse**.

Computer Adaptive Tests…

- Can be used with norm-referenced tests and criterion-referenced tests

- With criterion-referenced tests, the items SHOULD BE DIRECTLY LINKED to the criteria or framework being tested.

Classical Test Theory

and why it shouldn't be used for CATs

vs.

Item Response Theory

and why it is suited for this purpose

Jump!

Challenge

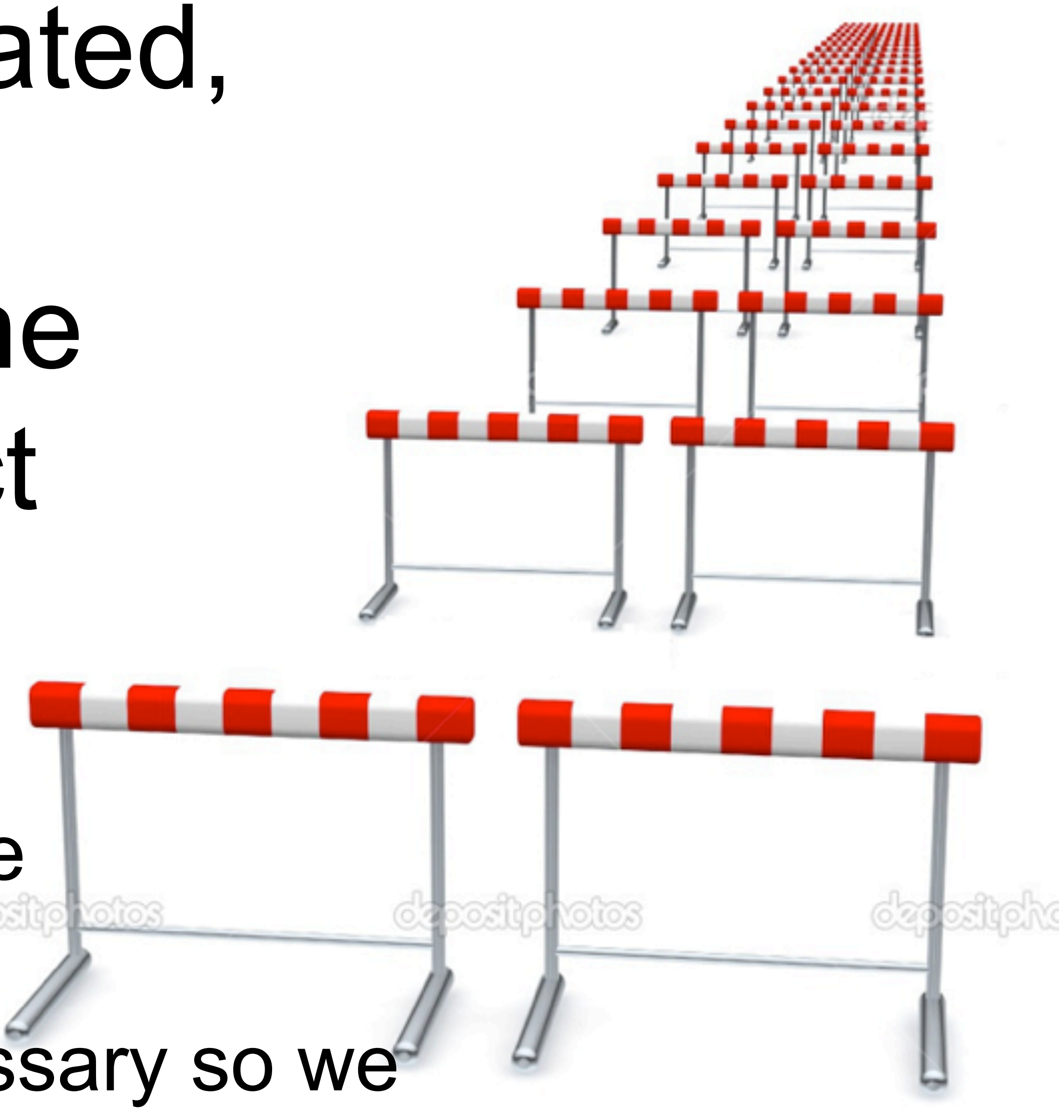
- Create an instrument to measure the “construct” of jumping ability.
- You have to be able to describe it to someone halfway across the world
- You can’t use standardized measures of length
  (No centimeters or inches allowed)

Need to determine the purpose

- Do I want to know which olympic medal podium they could jump to? (Criterion-referenced)

- Do I want to know their relative standing against each other? (e.g. Norm-referenced)

Series of repeated, independent measures of the same construct

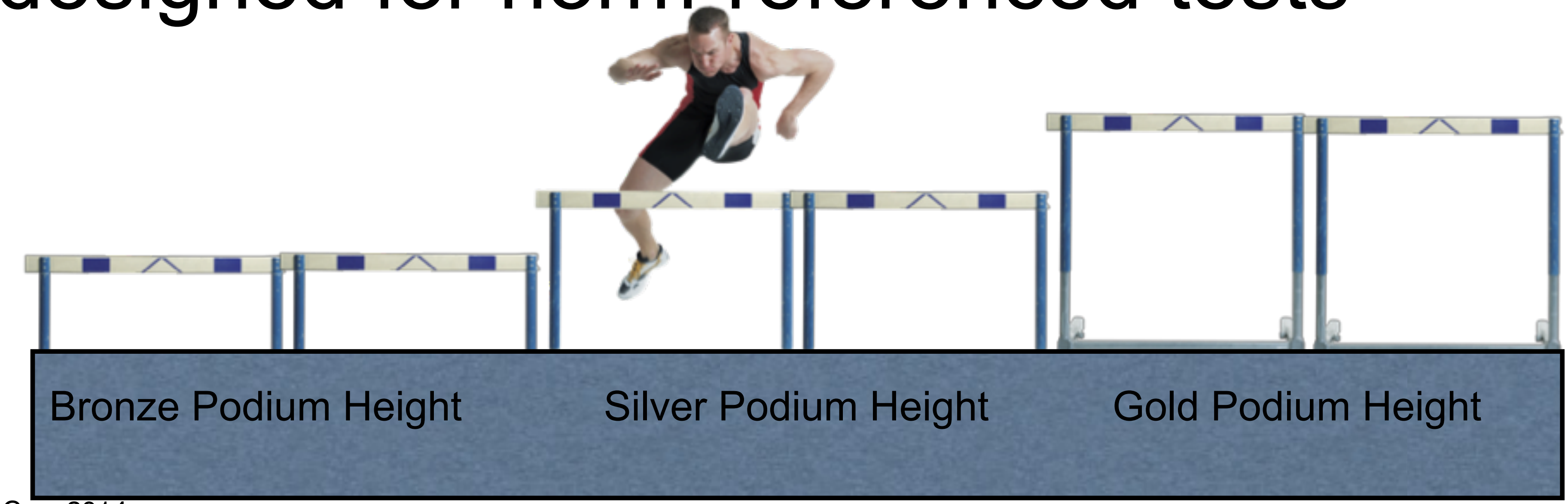


- Repeated performance increases confidence in reliability
- Independence necessary so we **add** the results into a single score

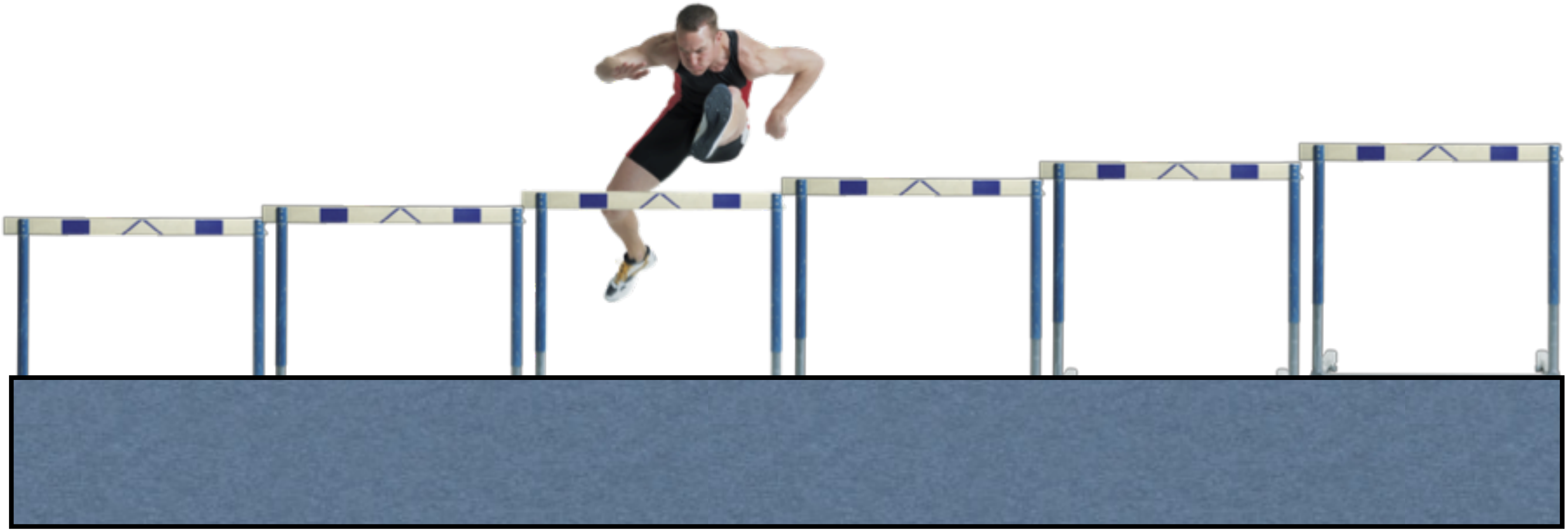
If it’s criterion-referenced, make the obstacles align with the criteria

BUT

Classical Test Theory was really designed for norm-referenced tests



So, for norm-referenced tests, have a range of obstacles that can differentiate the jumpers.



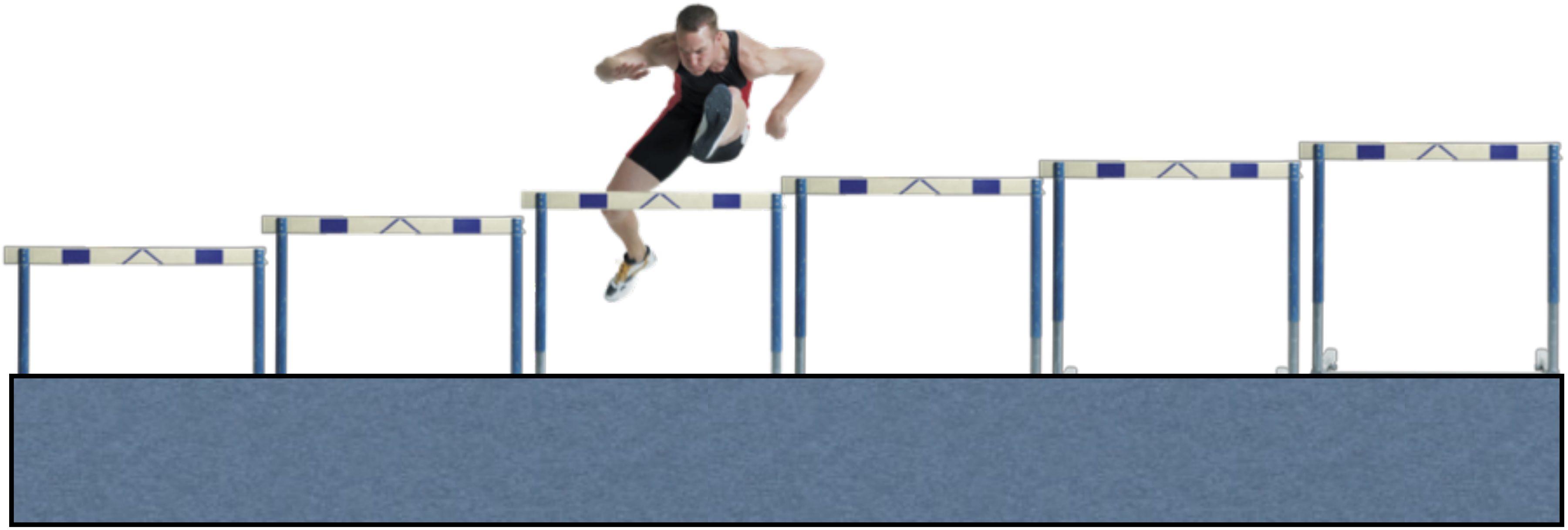
Classical Test Theory

Mathematical Model

True Score = Observed Ability (on entire test)

+

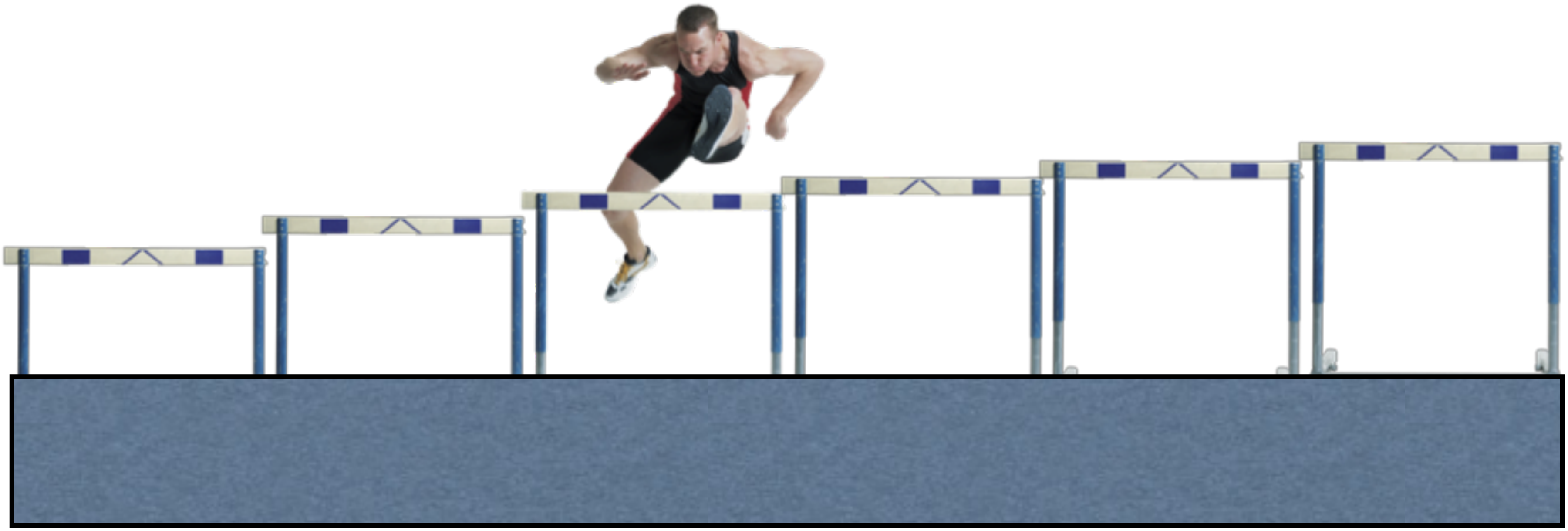
Error (single value for test and test-takers)



Classical Test Theory Limitations

Item Dependent (Person Score is additive result of performance on all items which contribute equally to the score)

Group Dependent (Item Difficulty is proportional result of population of test-takers)

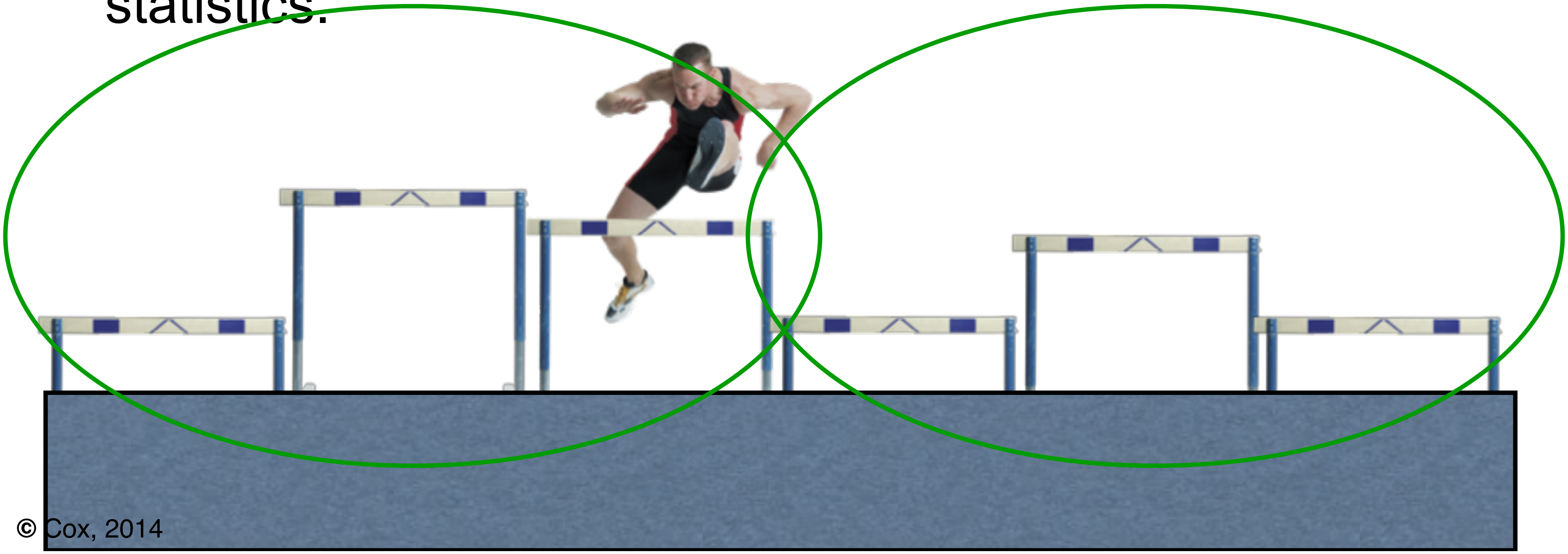


Classical Test Theory Limitations

Only applies to the test being administered.

EACH item is counted as a unit of measure (or interval) on the scale

For test forms to be equated, there need to be shared items. Test forms cannot be equated with just item statistics.

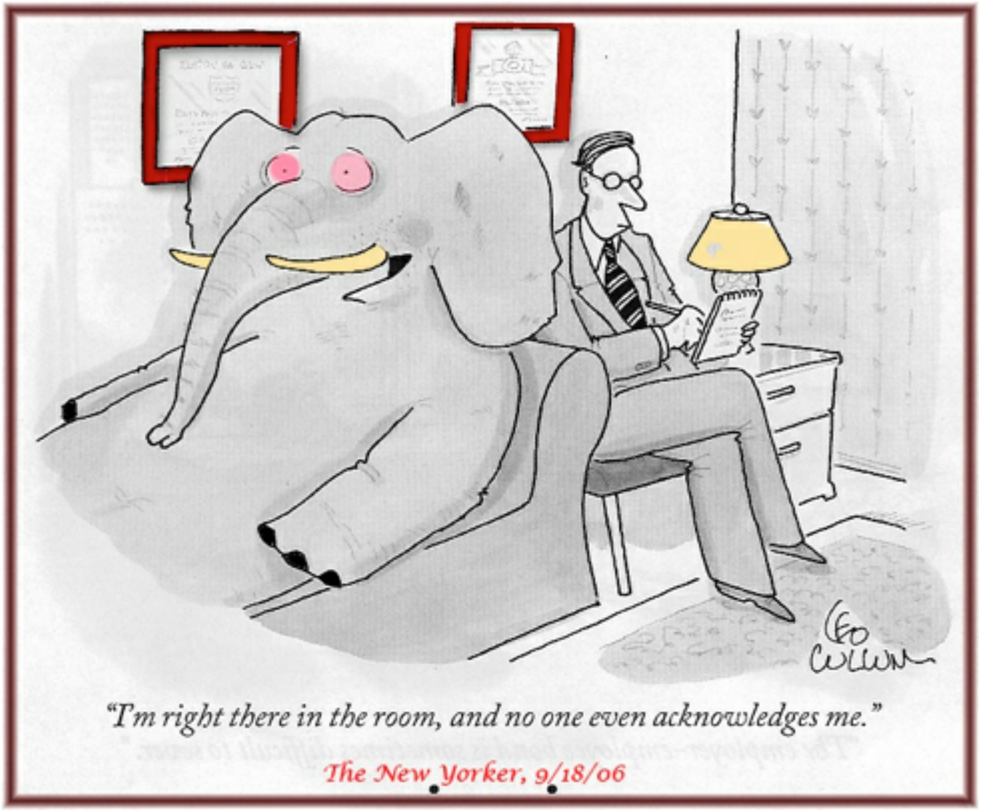


© Cox, 2014

Most Educational Tests

Are NOT interval (though everyone pretends they are)

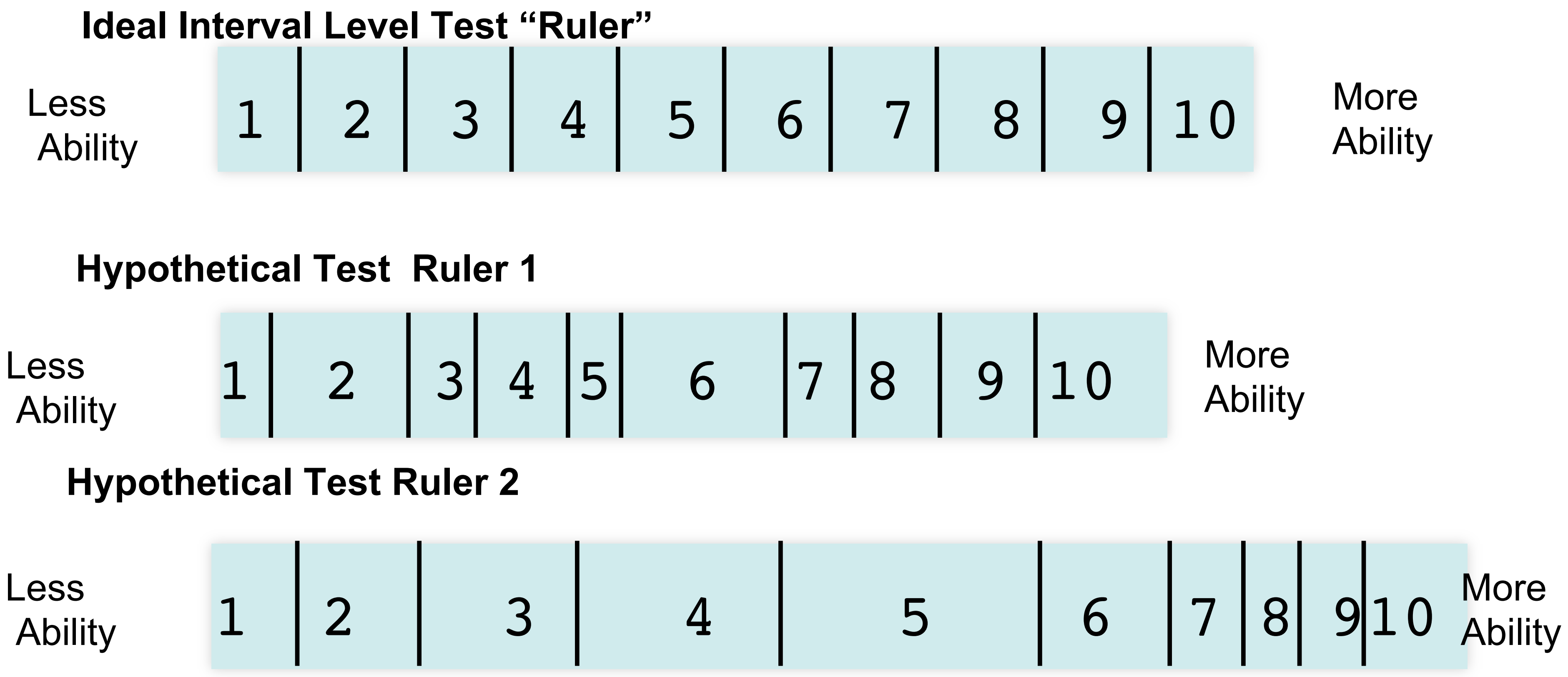
Are probably more ordinal than anything else



"I'm right there in the room, and no one even acknowledges me."

The New Yorker, 9/18/06

PROBLEM: There is NO external norm to validate the measurement instrument

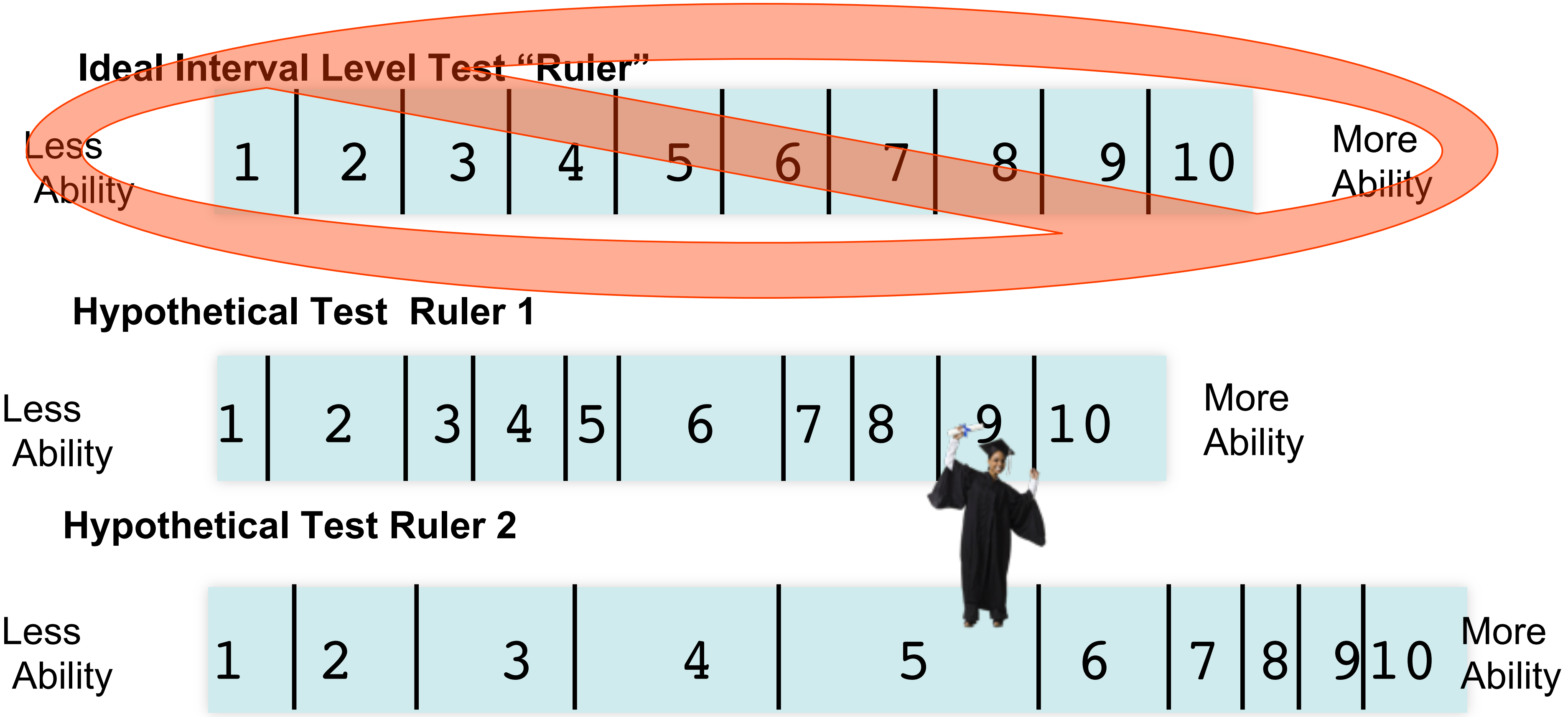


Ideal Interval Level Test "Ruler"

Hypothetical Test Ruler 1

Hypothetical Test Ruler 2

PROBLEM: There is NO external norm to validate the measurement instrument



Ideal Interval Level Test "Ruler"

Less Ability | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | More Ability

Hypothetical Test Ruler 1

Less Ability | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | More Ability

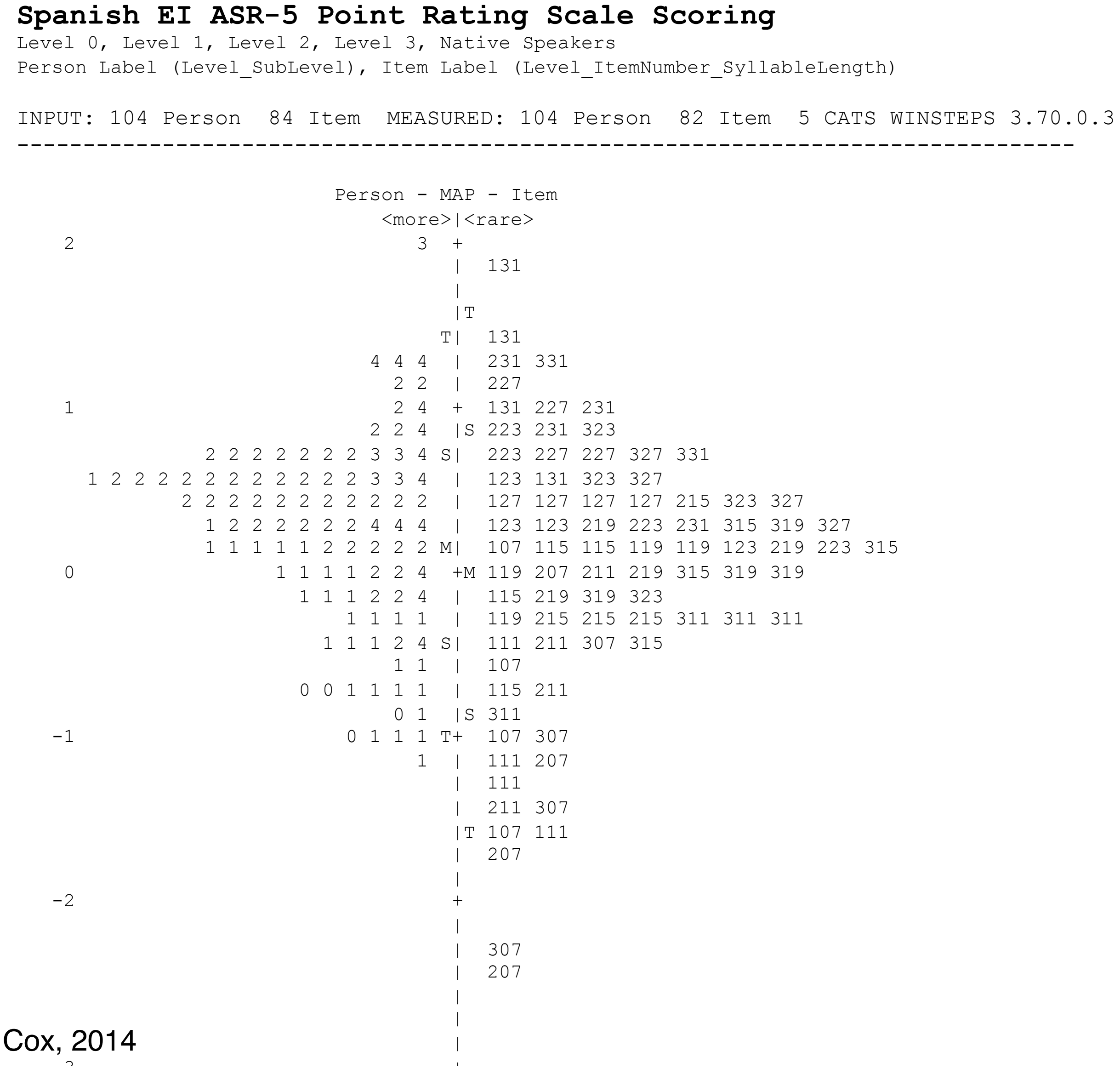
Hypothetical Test Ruler 2

Less Ability | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | More Ability

Classical Test Theory should not be used for CATs because it...

- is group dependent.
- is item dependent.
- does not produce interval data.
- is based on the assumption that there is a true score for an entire test that can predict future performance of individuals in the target language.
- assumes true test score vs. latent person ability

Item Response Theory



```
Spanish EI ASR-5 Point Rating Scale Scoring
Level 0, Level 1, Level 2, Level 3, Native Speakers
Person Label (Level_SubLevel), Item Label (Level_ItemNumber_SyllableLength)

INPUT: 104 Person  84 Item  MEASURED: 104 Person  82 Item  5 CATS WINSTEPS 3.70.0.3
--------------------------------------------------------------------------------------------

                           Person - MAP - Item
                               <more>|<rare>
    2                             3  +
                                     |  131
                                     |
                                     |T
                                    T|  131
                              4 4 4  |  231 331
                                2 2  |  227
    1                           2 4  +  131 227 231
                              2 2 4  |S 223 231 323
                16 2 2 2 2 2 2 3 3 4 S|  223 227 227 327 331
      1 2 2 2 2 2 2 2 2 2 2 2 3 3 4  |  123 131 323 327
              2 2 2 2 2 2 2 2 2 2 2  |  127 127 127 127 215 323 327
                1 2 2 2 2 2 2 4 4 4  |  123 123 219 223 231 315 319 327
                1 1 1 1 1 2 2 2 2 2 M|  107 115 115 119 119 123 219 223 315
    0                 1 1 1 1 2 2 4  +M 119 207 211 219 315 319 319
                        1 1 1 2 2 4  |  115 219 319 323
                            1 1 1 1  |  119 215 215 215 311 311 311
                          1 1 1 2 4 S|  111 211 307 315
                                1 1  |  107
                        0 0 1 1 1 1  |  115 211
                                0 1  |S 311
   -1                       0 1 1 1 T+  107 307
                                  1  |  111 207
                                     |  111
                                     |  211 307
                                     |T 107 111
                                     |  207
                                     |
   -2                                +
                                     |
                                     |  307
                                     |  207
                                     |
                                     |
                                     |
```

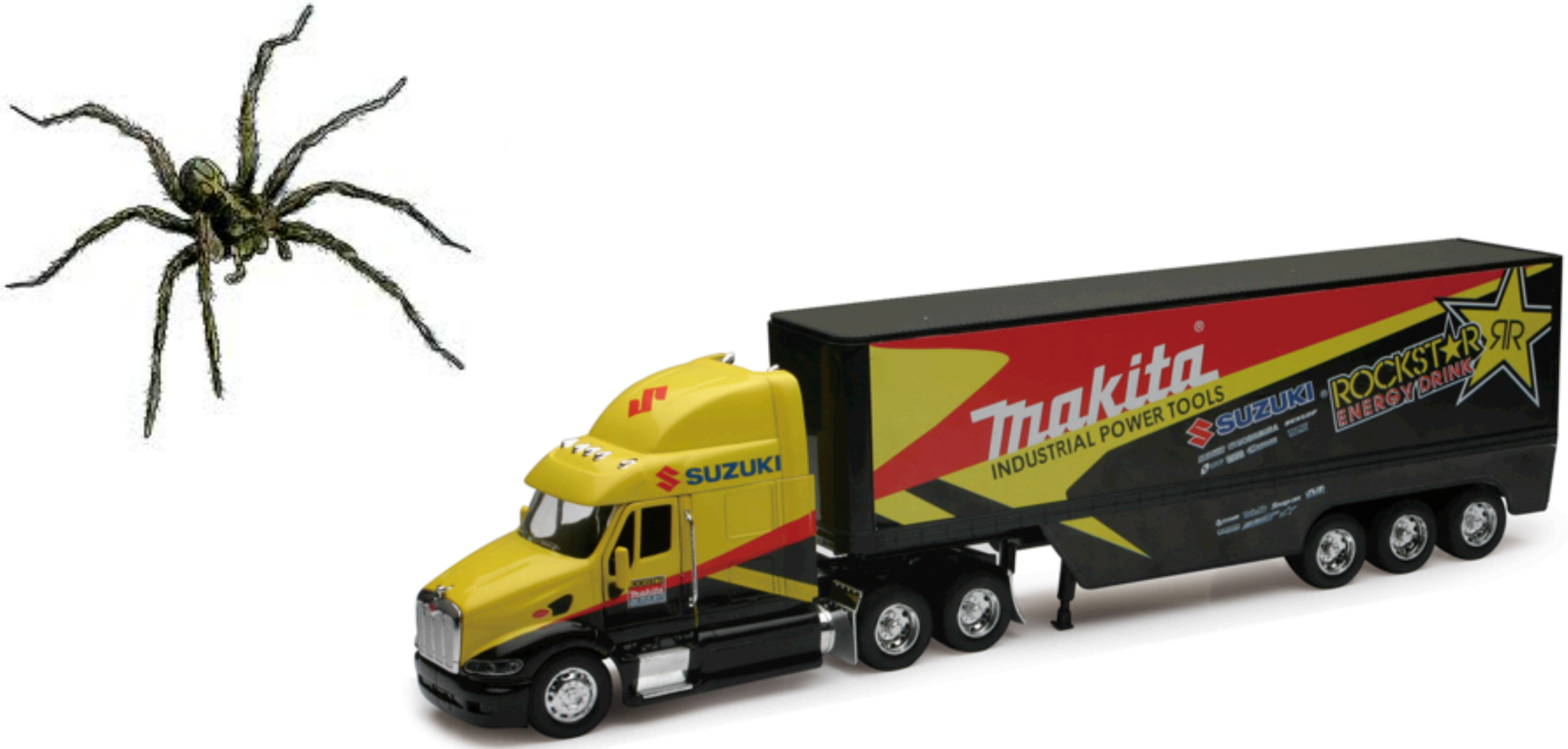
"Nothing is more practical than a good theory." *Kurt Lewin*

How do we determine *increasing* and *decreasing* “X”?

Is this animal large or small?

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

And this one, large or small?



DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Best to measure it, but rulers don’t exist in the social sciences.

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Many Sorts of Trucks

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

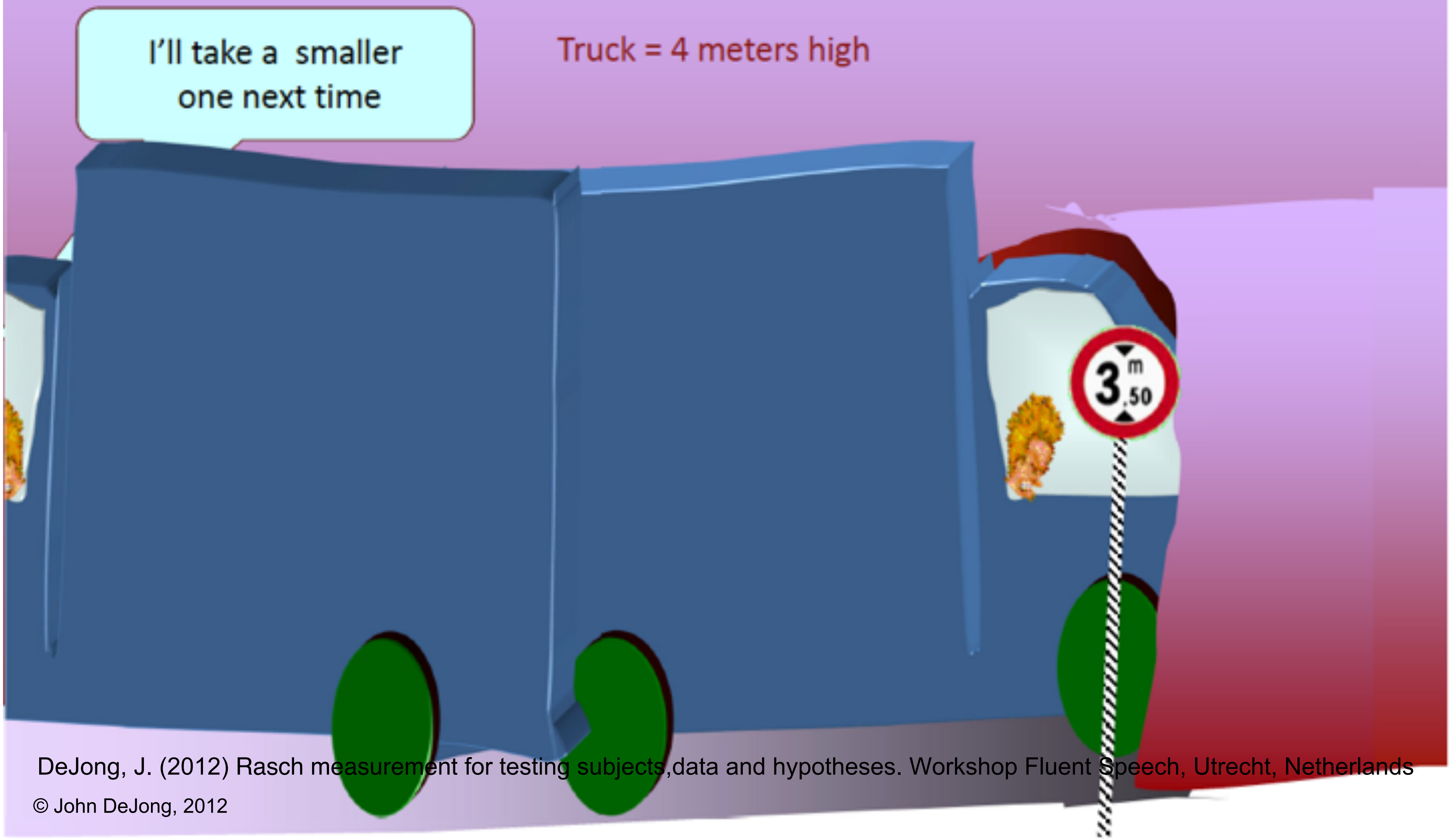
Many Sorts of Tunnels

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Truck = 3 meters high

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Trucks & Tunnels B



Truck = 4 meters high

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Trucks & Tunnels C

Truck = ± 3.5 meters high

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Trucks & Tunnels: Conclusion

- If height of truck < height of tunnel, then Pass=1
- If height of truck > height of tunnel, then Pass=0
- If height of truck = height of tunnel, then Pass= 50/50

- The most precise information about the height of truck and tunnel comes from the third equation.

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

To pass…

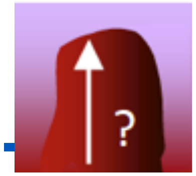
- The probability (Greek letter pi: Π) that we will observe a pass is a function of the difference in height between the truck and the tunnel

- $\Pi_{\{Pass=yes\}}$ = Function(

)

…or not to pass

- The probability that we will NOT observe a pass is also a function of the difference in height between the truck and the tunnel

- $\Pi_{\{Pass=no\}}$ = Function(- )

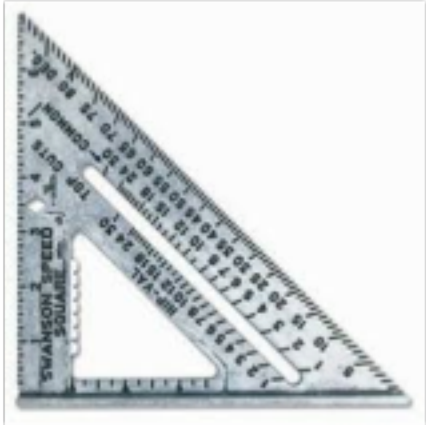
Truck and Tunnel Measurement

We can use these findings in two ways:

1. If we know the height of all the tunnels in Europe…
    1. We can measure the height of the trucks by sending them through Europe and seeing which tunnels they can pass through.
2. If we know the height of our trucks…
    2. We can measure the height of the tunnels in Europe by sending the trucks through Europe and see which tunnels they can pass.

Note: Either tunnels give us information about trucks, or trucks give us information about tunnels.

With physical objects, we use standardized measurements.



With things you can’t see,

- you need to make hypotheses and observations.

Latent Trait Theory

We cannot see the constructs we are measuring.

Since we cannot see them, they are latent.

We can talk about constructs, and form an opinion, but to measure the construct we need a theory to explain our observations.

STANAG provides an operational theory of real world language use.

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Persons and Items

- Persons and Items are like Trucks and Tunnels
- We have seen how we can get information on the height of a truck if we send it through a tunnel with known height, by observing whether the truck can pass through the tunnel.
- Likewise we can get information on the traits of people if we observe the result of confronting them with an item of known difficulty.

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Item Response Theory

- The observable result of a ‘person-by-item’ confrontation is the response given by the person.

- Item Response Theory (IRT) was originally called “Latent Trait Theory”

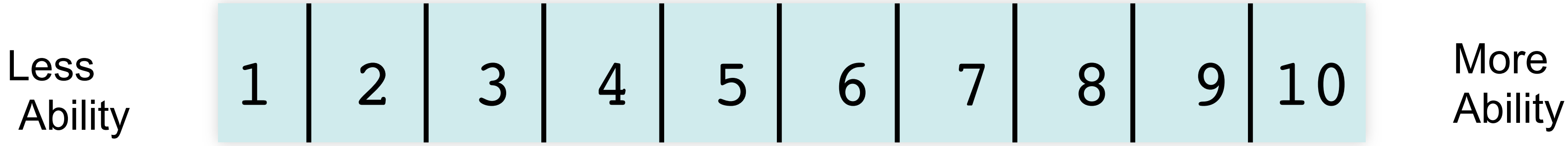
DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands

Assumptions for IRT

- Unidimensionality
  - Remember the shirts
- Local independence
  - Remember the hurdles
- Sufficient statistics
- Similar to CTT but more stringent

What about these assumptions?

- They are assumptions—not facts; we use the theory to check whether we can maintain the assumption.
- If the test meets the assumptions, than we know the test can be a measurement instrument.

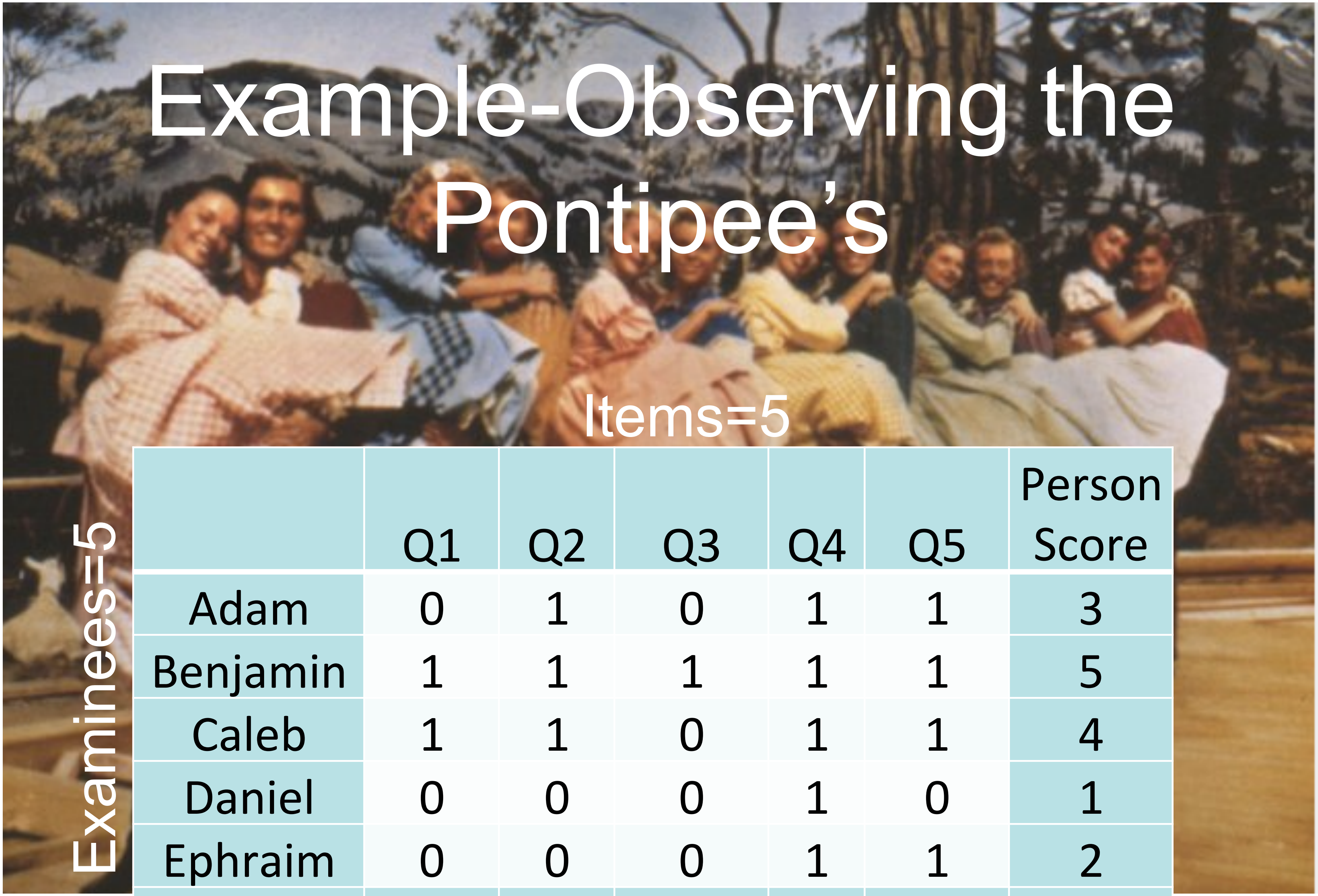


Less Ability

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

More Ability

DeJong, J. (2012) Rasch measurement for testing subjects,data and hypotheses. Workshop Fluent Speech, Utrecht, Netherlands



Example-Observing the Pontipee's

Items=5

Examinees=5

	Q1	Q2	Q3	Q4	Q5	Person Score
Adam	0	1	0	1	1	3
Benjamin	1	1	1	1	1	5
Caleb	1	1	0	1	1	4
Daniel	0	0	0	1	0	1
Ephraim	0	0	0	1	1	2
Item Score	2	3	1	5	4	15

Marginal totals

	Q1	Q2	Q3	Q4	Q5	Person Score
Adam	0	1	0	1	1	3
Benjamin	1	1	1	1	1	5
Caleb	1	1	0	1	1	4
Daniel	0	0	0	1	0	1
Ephraim	0	0	0	1	1	2
Item Score	2	3	1	5	4	15

These are the marginal totals. They contain all information about items and persons.

Interpreting Marginal Totals

	Q1	Q2	Q3	Q4	Q5	Person Score
Adam	0	1	0	1	1	3
Benjamin	1	1	1	1	1	5
Caleb	1	1	0	1	1	4
Daniel	0	0	0	1	0	1
Ephraim	0	0	0	1	1	2
Item Score	2	3	1	5	4	15

We see that Benjamin answered all the items correct.
We see that Q4 was answered correctly by all persons.

Sort by person score marginal totals

	Q1	Q2	Q3	Q4	Q5	Person Score
Benjamin	1	1	1	1	1	5
Caleb	1	1	0	1	1	4
Adam	0	1	0	1	1	3
Ephraim	0	0	0	1	1	2
Daniel	0	0	0	1	0	1
Item Score	2	3	1	5	4	15

Sort by item score marginal totals

	Q4	Q5	Q2	Q1	Q3	Person Score
Benjamin	1	1	1	1	1	5
Caleb	1	1	1	1	0	4
Adam	1	1	1	0	0	3
Ephraim	1	1	0	0	0	2
Daniel	1	0	0	0	0	1
Item Score	5	4	3	2	1	15

Predicting from marginal totals

	Q4	Q5	Q2	Q1	Q3	Person Score
Benjamin	1	1	1	1	1	5
Caleb						4
Adam	1	1	1	0	0	3
Ephraim	1	1	0	0	0	2
Daniel	1	0	0	0	0	1
Item Score	5	4	3	2	1	

Caleb has a total score of 4. Which item did he most likely get wrong?

Predicting marginal totals by item response

	Q4	Q5	Q2	Q1	Q3	Person Score
Benjamin	1	1	1	1	1	5
Caleb	1	1	1	1	0	4
Adam	1	1	1	0	0	3
Ephraim	1	1	0	0	0	2
Daniel	1	0	0	0	0	1
Frank	0					?
Milly					1	?

What marginal totals would you predict Frank and Milly to have based on their item responses?

What happens when items or people don't cooperate with the model?

Remember the assumptions!

Items and Persons can be examined on their model fit

	Q4	Q5	Q2	Q1	Q3	Q6	Person Score
Milly	1	1	1	1	1	1	6
Benjamin	1	1	1	1	1	0	5
Caleb	1	1	1	1	0	0	4
Adam	1	1	1	0	0	0	3
Ephraim	1	1	0	0	0	1	3
Frank	0	0	0	1	1	1	3
Daniel	1	0	0	0	0	1	2
Item Score	6	5	4	4	3	3	

Frank doesn’t seem to fit the expected profile. Why? (Tall & skinny, short & fat, cheating)
Q6 doesn’t seem to fit. Why? (Dimensionality? Quality?)

Strengths of IRT

- Because “Item” is part of the mathematical model, items can be looked at separately and scaled separately
- If items are written to specific criterion, they are INDEPENDENT of the test-takers

What is the mathematical model?

Probability of Success

=

Function (Person Ability-Item Difficulty)

“In IRT models, trait scores are estimated separately for each score or response pattern, controlling for the characteristics (e.g., difficulty) of the items that were administered. **Standard errors are smallest when the items are optimally appropriate for a particular trait score level...”**

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*

Some Symbols

Π (the uppercase Greek letter *pi*) indicates Probability

θ (the lowercase Greek letter *theta*) indicates the ability of the person

δ (the lowercase Greek letter *delta*) indicates the difficulty of the item

x indicates the score on an item

Formula

$$\Pi_{\{x=1\}} = (\theta-\delta)$$

English translation: The probability that the item score will be 1 is a function of the difference between the person ability and the item difficulty.

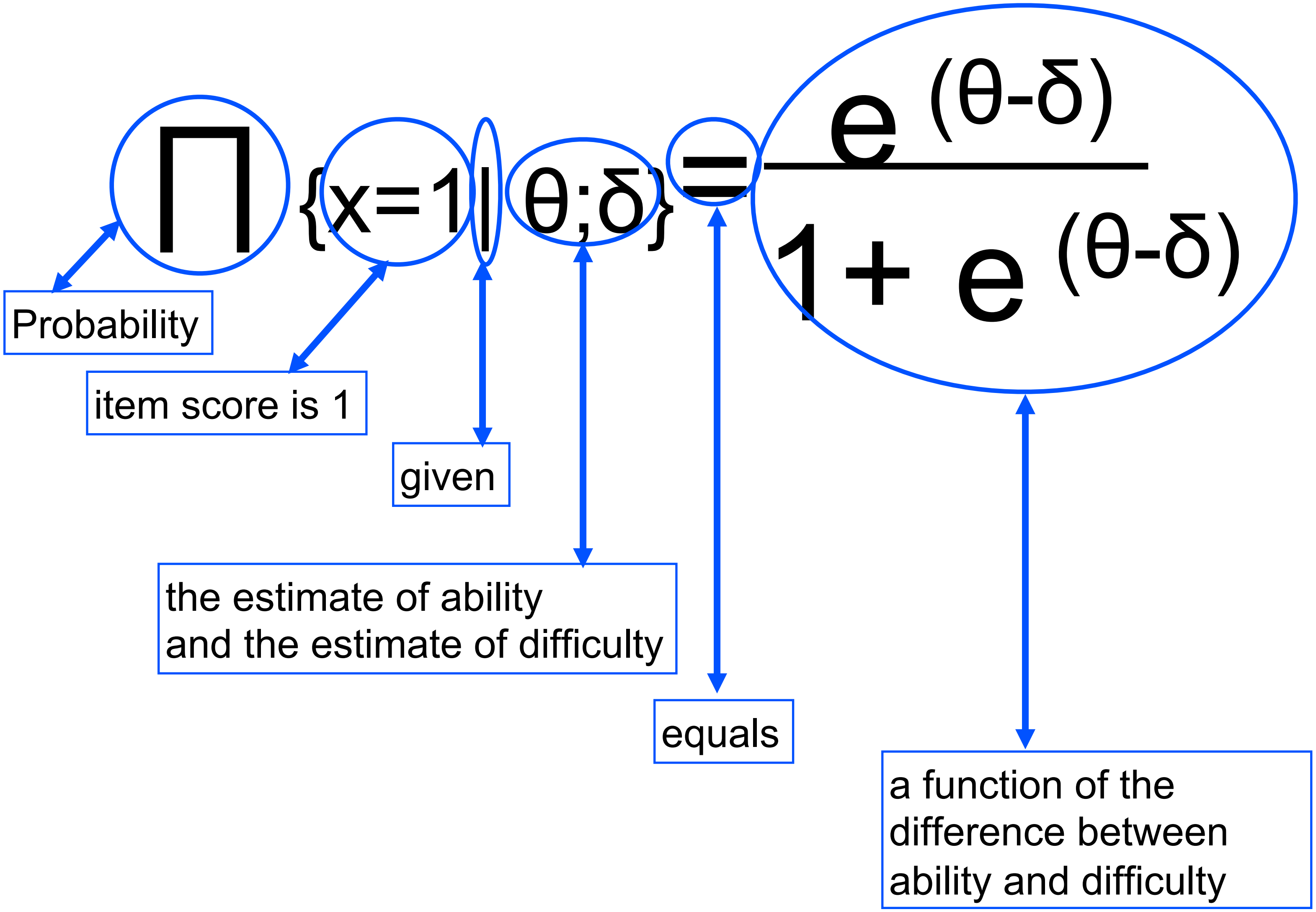
Additional Conclusions

If $\theta > \delta$, then $\Pi_{\{x=1\}} > 50/50$

If $\theta < \delta$, then $\Pi_{\{x=1\}} < 50/50$

If $\theta = \delta$, then $\Pi_{\{x=1\}} = 50/50$

Rasch Formula



$$\Pi\{x=1|\theta;\delta\}=\frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}}$$

Probability

item score is 1

given

the estimate of ability
and the estimate of difficulty

equals

a function of the
difference between
ability and difficulty

Person ability estimate independent of items

$$\prod \{x=1 \mid \theta_{Millie};\delta_i\} = \frac{e^{(\theta_{Millie}-\delta_i)}}{1+e^{(\theta_{Millie}-\delta_i)}} \qquad \prod \{x=1 \mid \theta_{Adam};\delta_i\} = \frac{e^{(\theta_{Adam}-\delta_i)}}{1+e^{(\theta_{Adam}-\delta_i)}}$$

$$\prod \{x=1 \mid \theta_{Millie};\delta_i\} > \prod \{x=1 \mid \theta_{Adam};\delta_i\}$$

$$\theta_{Millie} > \theta_{Adam}$$

Item difficulty estimate independent of person

$$\prod \{x=1 \mid \theta_{Adam};\delta_{Q1}\} = \frac{e^{(\theta_{Adam}-\delta_{Q1})}}{1+e^{(\theta_{Adam}-\delta_{Q1})}} \qquad \prod \{x=1 \mid \theta_{Adam};\ \delta_{Q2}\} = \frac{e^{(\theta_{Adam}-\delta_{Q2})}}{1+e^{(\theta_{Adam}-\delta_{Q2})}}$$

Q1 **Q2**

$$\prod \{x=1 \mid \theta_{Adam};\ \delta_{Q1}\} > \prod \{x=1 \mid \theta_{Adam};\ \delta_{Q2}\}$$

$$-\ \delta_{Q1} > -\delta_{Q2}$$

$$\delta_{Q1} > \delta_{Q2}$$

Item Response Theory…

Is person independent

Is item independent

Puts person and item on the same scale

Allows items to be targeted to person ability level

Is ideal for CATs

Differences Between CCT and IRT

Area	Classical Test Theory	Item Response Theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e., easy to meet with test data)	Strong (i.e., more difficult to meet with test data)
Item-ability relationship	Not specified	Item characteristic functions
Ability	Test scores (estimated true scores) are reported on a test-score scale	Ability scores are reported on the scale -∞ to + -∞
Invariance of item and person statistics	No—item and person parameters are sample dependent	Yes—item and person parameters are sample independent, if the model fit the data

Hambleton, R. K., & Jones, R. W. (2005). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. Retrieved from Google Scholar.

What do we learn from shirts, bikes, hurdles and tunnels?

- Shirts
  - Hardly anything is truly unidimensional.

    There needs to be clear communication between test-creators and test-users on what is being measured.
- Bikes
  - Is it criterion or norm-referenced?

    If criterion-referenced, how do items/rubric relate to the criteria?
- Hurdles
  - There needs to be independent, repeated measures.
- Tunnels and Trucks
  - Conjoint measurement and Rasch IRT (tunnels and trucks; persons and items).

Recipe: Computer Adaptive Test

- Computer
  - Programming
  - Equipment
- Adaptive
  - Algorithm
- Test
  - Psychometrics