Connexion of Item Response Theory to Decision Making in Chess

Presented by

Tamal Biswas

Research Advised by Dr. Kenneth Regan

Acknowledgement

- A few Slides have been taken from the following presentation and others with it:
- www.ice.ntnu.edu.tw/files/news/67_df937243.ppt

- Java applets and IRT basics are taken from
  - A visual guide to Item Response Theory
    - by I. Partchev

- Connexion: the British spelling connotes not just relation but a kind of union.

What is Measurement ?

- Abstractly, *measurement* is the assignment of numbers to objects or events.[1] The numbers intend to be unique “true values” of certain quantities, but how they carry that intent is the main issue.

- Context of Measurement:
  - Measurement in the Physical World
  - Measurements in the Psycho-social Science
  - Psychometrics

Formal Definition

In the Physical World

Assign numbers to objects or events

Ex: Time, Length, Weight, Temperature

In the Psycho-social Science

Measurement is the process of constructing lines and locating individuals on lines (Wright et. al, 1979)

Psychometrics

Assigning numbers to psychological characteristics
Examples – IQ, Opinion, Ranking

The Definition is Not Really Different

- *Measurement in psychology and physics are in no sense different. Physicists can measure when they can find the operations by which they may meet the necessary criteria; psychologists have but to do the same. They need not worry about the mysterious differences between the meaning of measurement in the two sciences. (Reese, 1943, p. 49)*

- *Still definition is important*
- http://rjlipton.wordpress.com/2013/06/06/psst-its-the-definition/

Levels of Measurement

- Nominal
- Ordinal
- Interval
- Ratio

http://en.wikipedia.org/wiki/Level_of_measurement

Purpose of Designing Tests

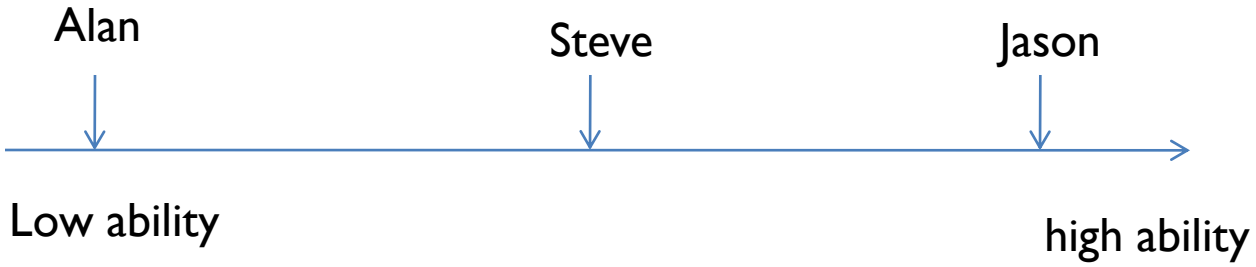
- To find out about ‘something’.
- What is it that we want to find out? About the items, or about the people?
  - In general, “measurement” is concerned with finding out about characteristics (latent traits) of people. The items are instrumental (or incidental) to achieve the measurement.
- We define measurement as measuring latent traits of people.

What do we need to achieve in the measurement?

- To tell us what people know about, or can do, certain things.
- So, we need to make sure that
  - Our measures are accurate (reliability)
  - Our measures are indeed tapping into the skills we set out to measure (validity)
  - Our measures are “invariant” even if different tests are used.

Ideal Measurement

- Scores we obtained are meaningful. What can each of these students do?

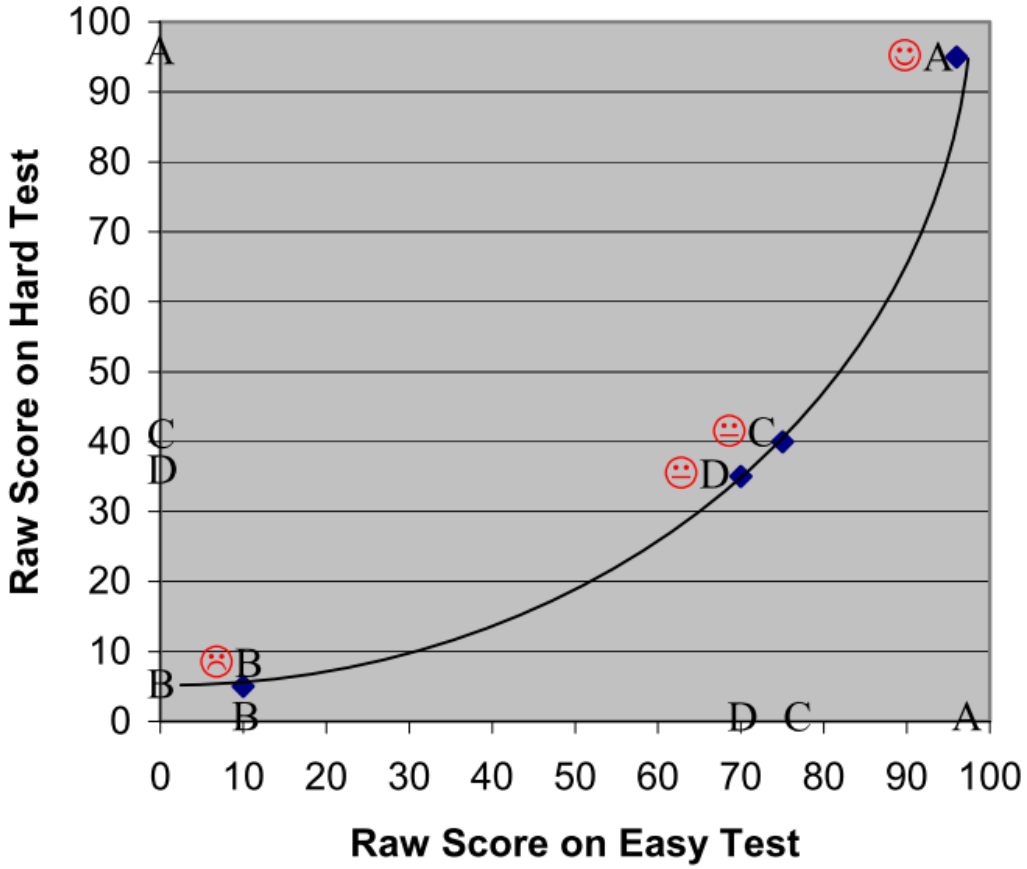


- Scores are independent of the sample of items used. If a different set of items are used, we will get the same results.

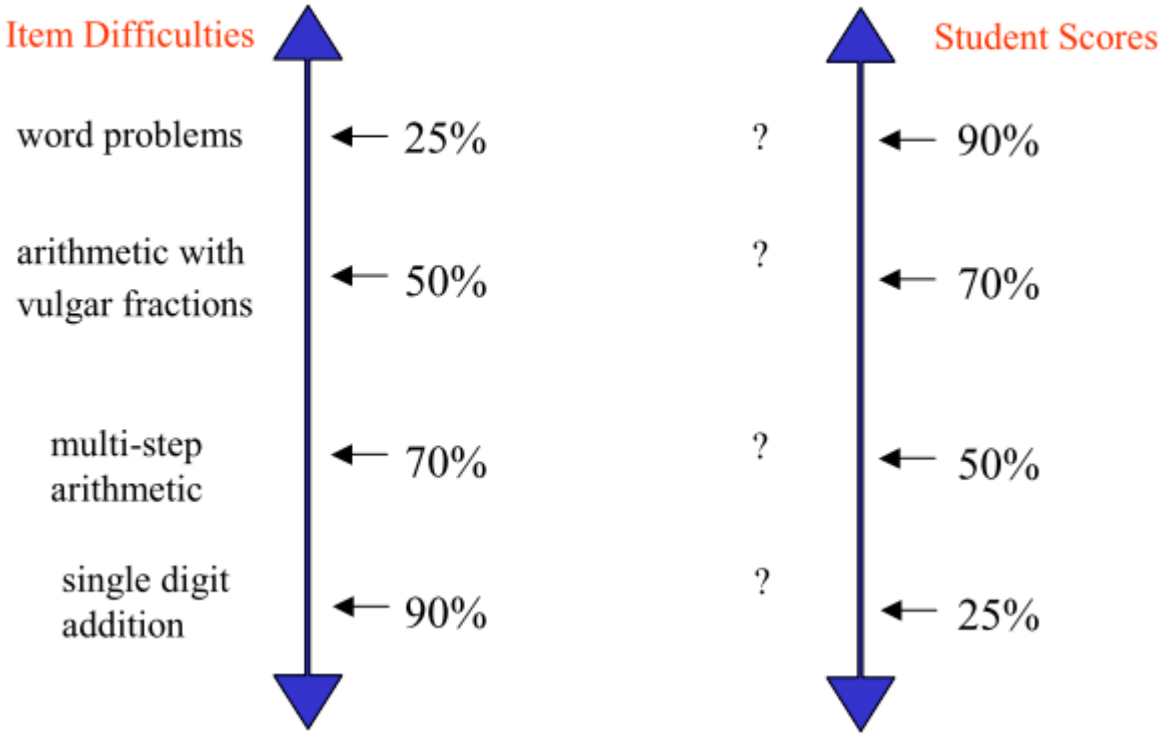
Problem of using Raw Scores

- Can raw scores provide the properties of an ideal measurement?
  - Distances between differences in scores are not easily interpretable.
  - Difficult to link item scores to person scores.

Raw Score Measurement



Link Raw Scores on Items and Persons



Item Response Theory

  - Item response theory helps us with achieving the goals of constructing the “best measurement”.
  - Meanings can be constructed to describe scores
  - Student scores should be independent of the particular set of items in the test.
- IRT provides tools to assess the extent to which good measurement properties are achieved.

Descriptions of IRT

- "IRT refers to a set of mathematical models that describe, in probabilistic terms, the relationship between a person's response to a survey question/test item and his or her level of the 'latent variable' being measured by the scale"
  - Fayers and Hays p55
  - Assessing Quality of Life in Clinical Trials. Oxford Univ Press:
  - Chapter on Applying IRT for evaluating questionnaire item and scale properties.

- This latent variable is usually a hypothetical construct [trait/domain or ability] which is postulated to exist but cannot be measured by a single observable variable/item.

- Instead it is indirectly measured by using multiple items or questions in a multi-item test/scale.

More about IRT

- IRT models give the *probability of success of a person on an item.*
- IRT models are not deterministic, but probabilistic.
- Given the item difficulty and person ability, one can compute the probability of success for each person on each item.

What IRT does

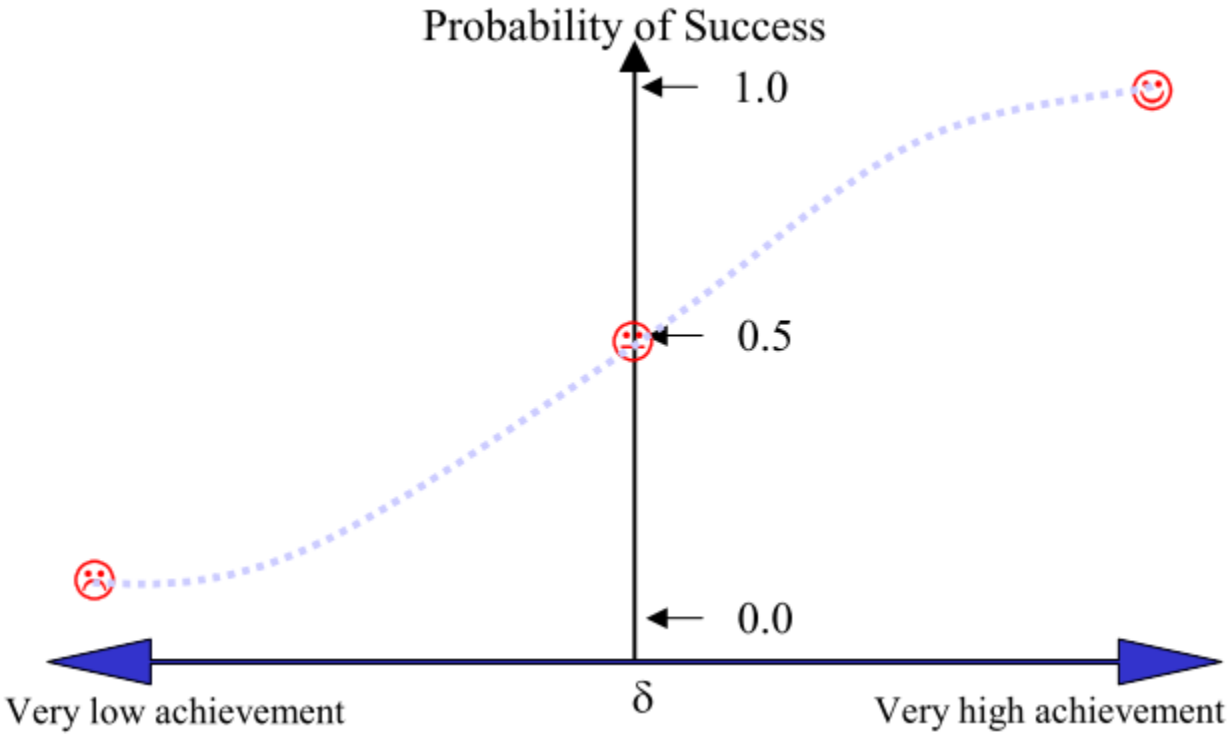
IRT models provide a clear statement [picture!] of the performance of each item in the scale/test and how the scale/test functions, overall, for measuring the construct of interest in the study population

The objective is to model each item by estimating the properties describing item performance characteristics hence Item Characteristic Curve or Symptom Response Function.

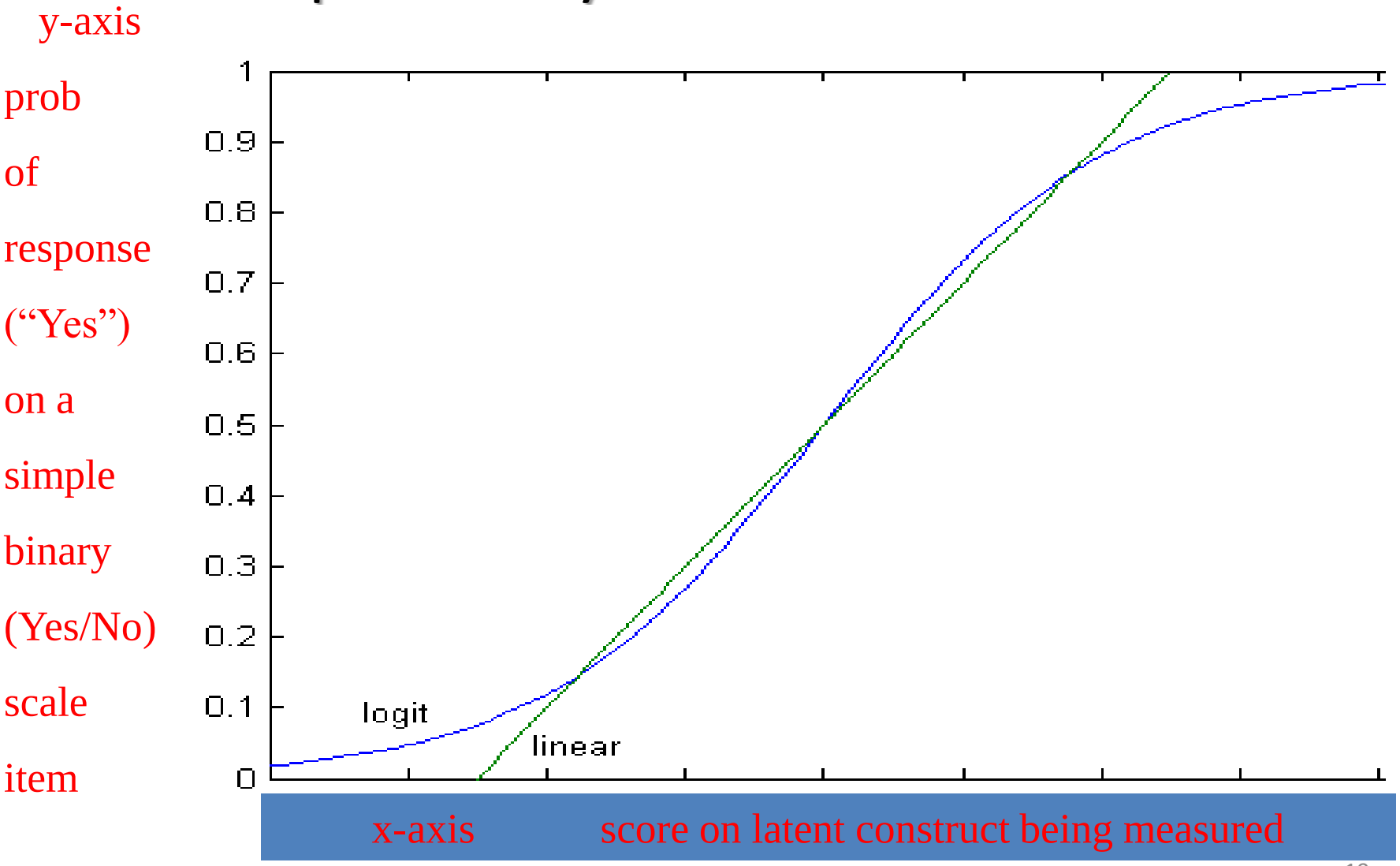
Can we use IRT in evaluating chess players?

- Yes.
- Each Position faced is nothing but a question asked.
- In chess, you cannot let a move ‘go’ (Pass)
- Evaluate ‘Trait’ of the move from computer analysis.

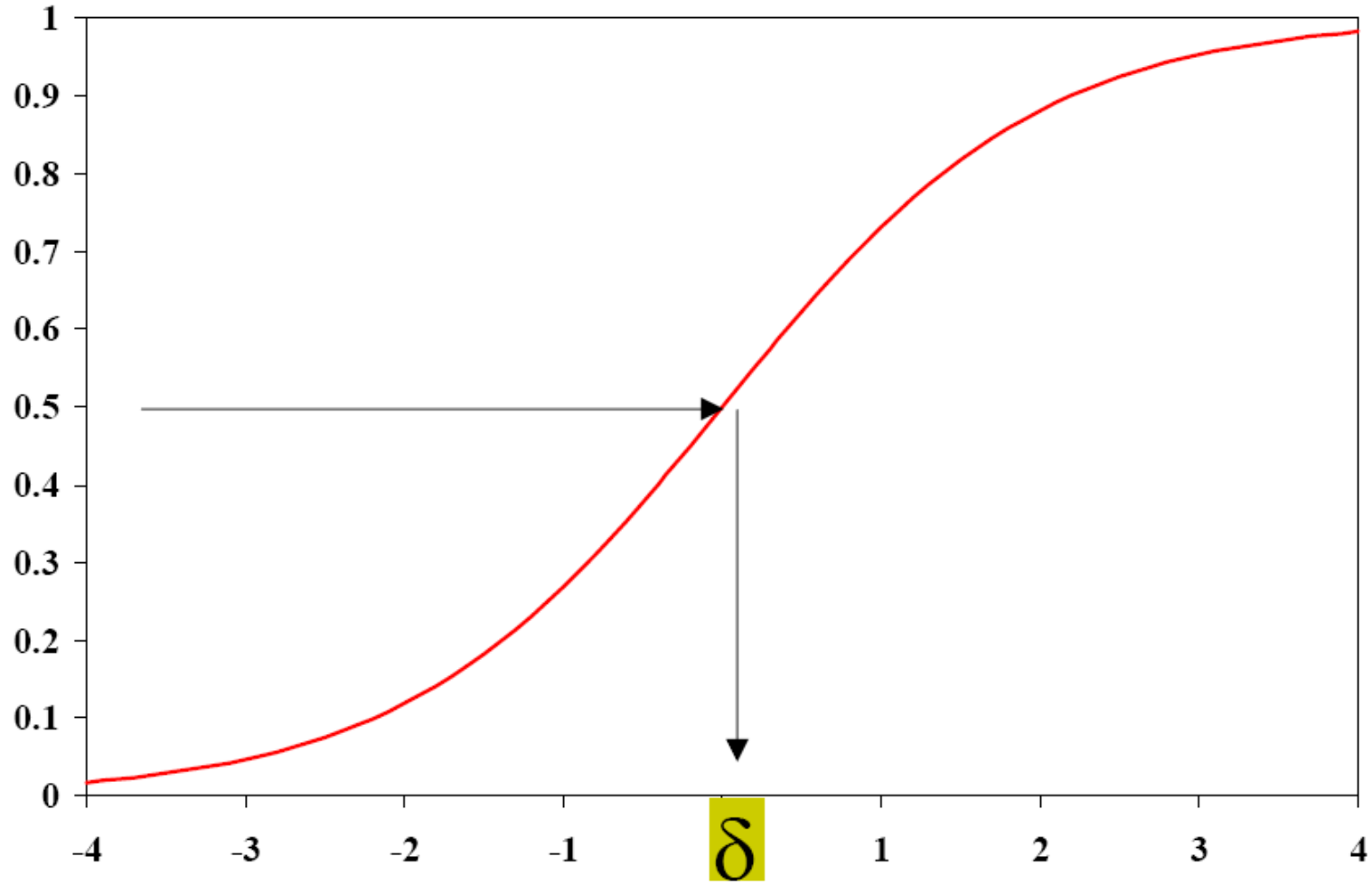
Item Characteristic Curve



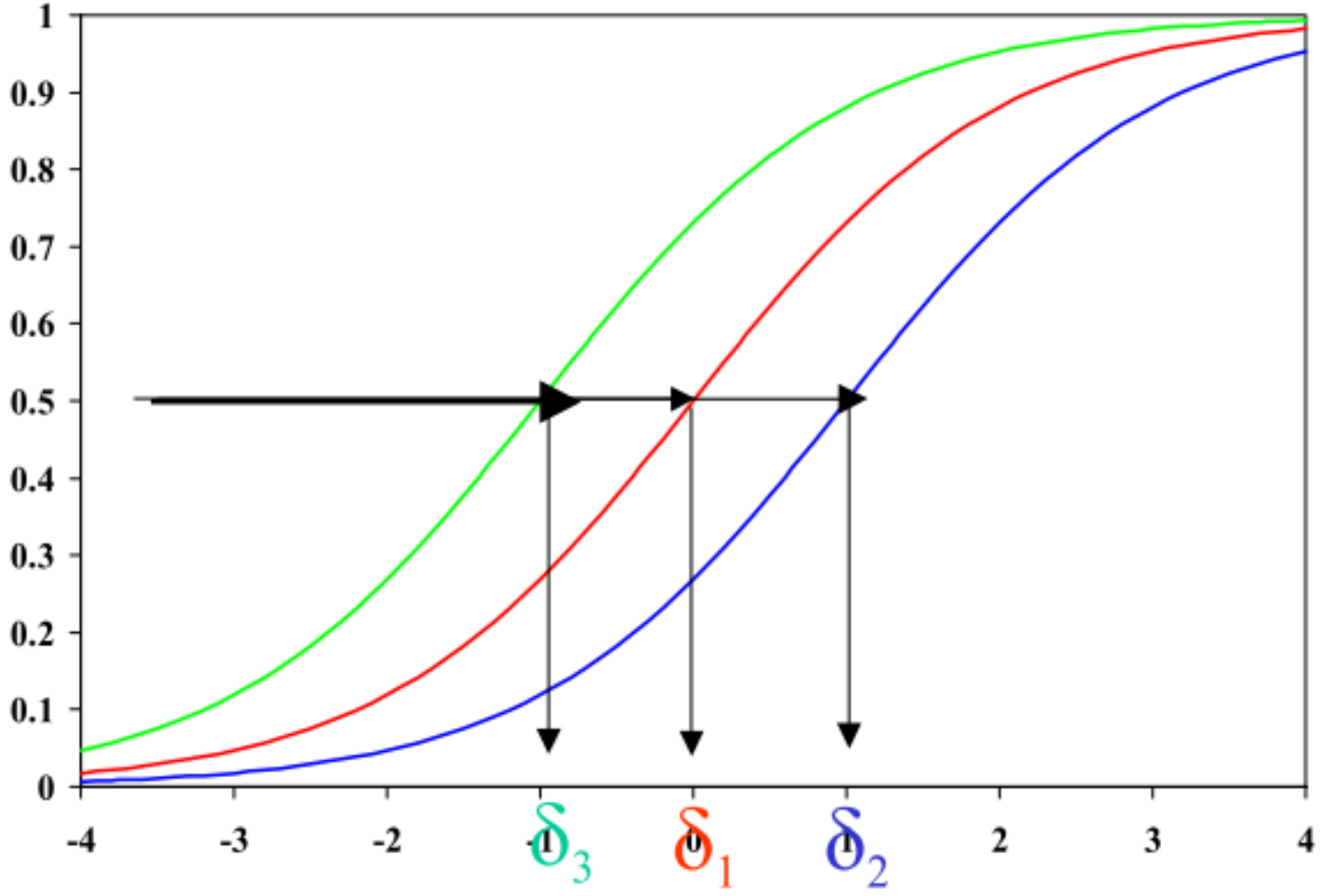
Linear vs non-linear regression of response probability on latent variable



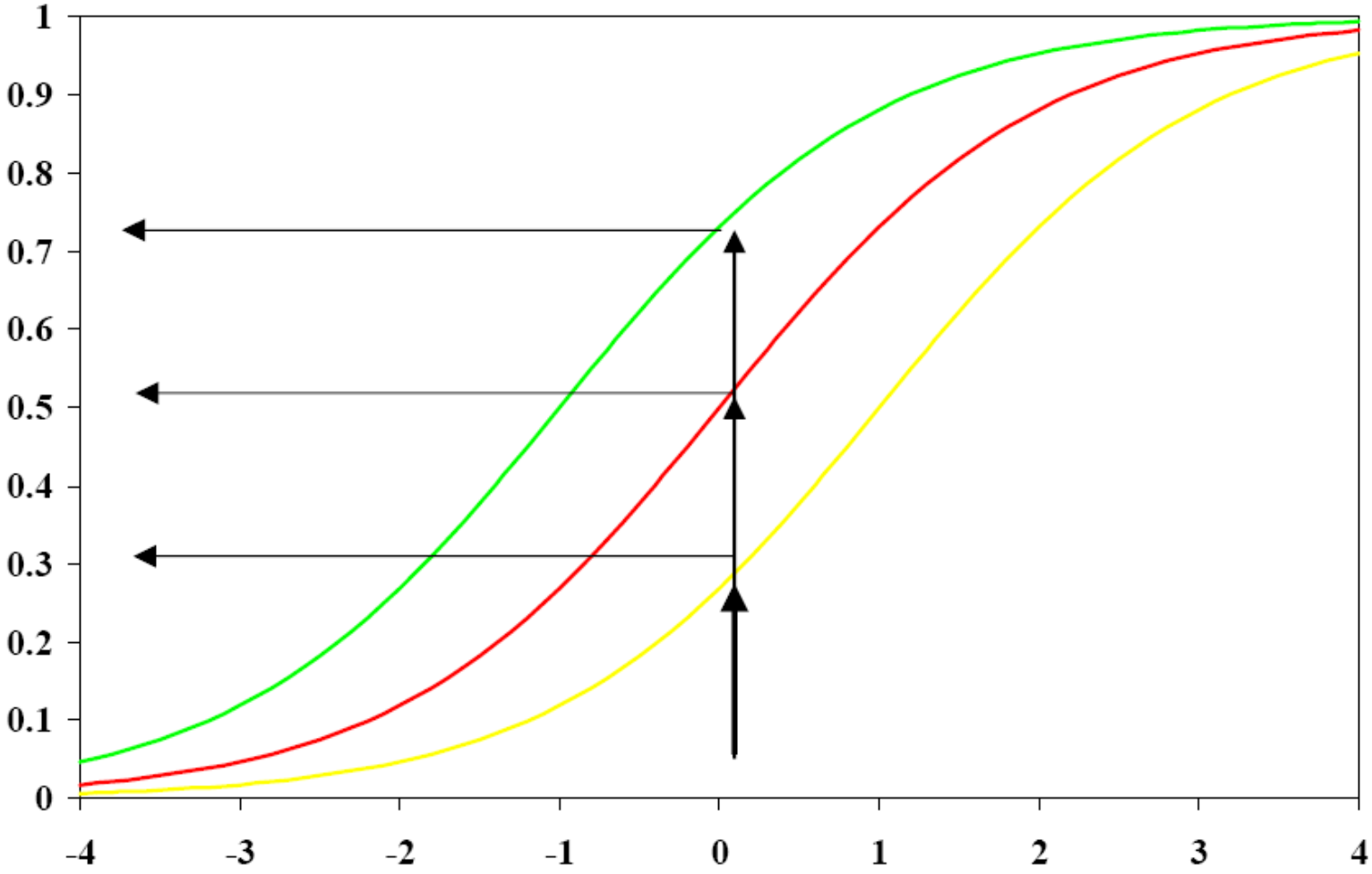
ICC example



ICC example 2



Probabilities of success for a person



Applet

- http://www.metheval.uni-jena.de/irt/ptb.html

Common IRT Models

Item Responses	Number of parameters	Model Developer	Software
Dichotomous	One PL model (Rasch)	Rasch(1960)	WINSTEPS (Rasch), BILOG, BILOG-MG PARSCALE MULTILOG
	Two PL model	Lord(1952)	
	Three PL model	Birnbaum(1968)	
Polytomous	nominal response model	Bock(1972)	MULTILOG, ConQuest PARSCALE WINSTEPS
	grade response model	Samejima(1969)	
	partial credit model	Wright & Masters(1982)	
	rating scale model	Andrich (1978)	

Assumptions for IRT (1)

- Dimensionality
  - Only one “dominant” ability is measured by a set of items in a test (unidimensional IRT model).
  - Factors that affects unidimensionality of a test
    - test anxiety
    - motivation
    - nuisance cognitive skills
    - etc.
  - However, IRT model can be extended with more than one ability (multidimensional IRT model).

Assumptions for IRT (2)

- Local Independence
  - When the abilities remain constant, there is no relationship between examinees' responses to different items.
  - When unidimensionality is satisfied, local independence assumption is automatically satisfied.

Assumptions of IRT

- Unidimensionality – only one ability is measured by a set of items on a test
- Local independence – examinee's responses to any two items are statistically independent
- 1-parameter model – no guessing, item discrimination is the same for all items
- 2-parameter model – no guessing

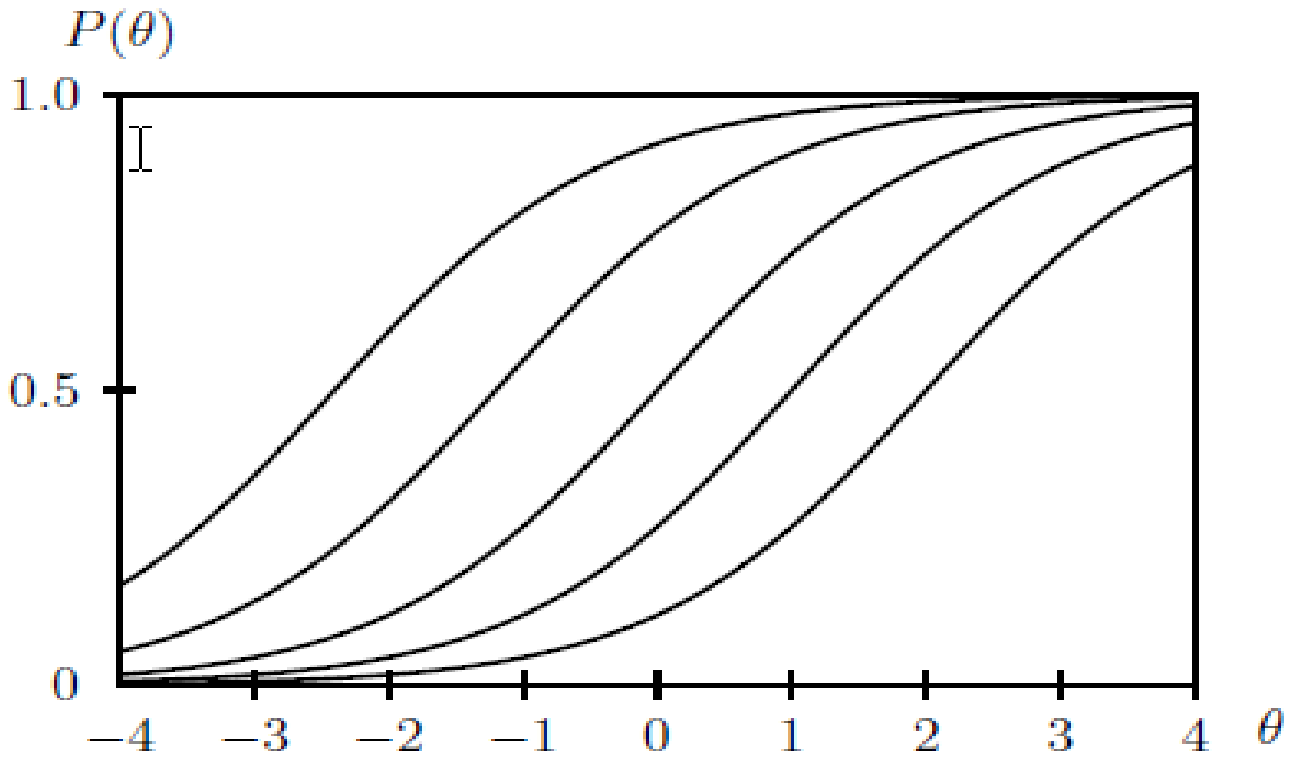
1-Parameter Logistic Model (The Rasch Model)

- Only item difficulty varies across items.
- Assume item discriminations are equal across items (a = 1)

$$P_j(\theta) = \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)} = \frac{1}{1 + \exp\left[-(\theta - b_j)\right]}$$

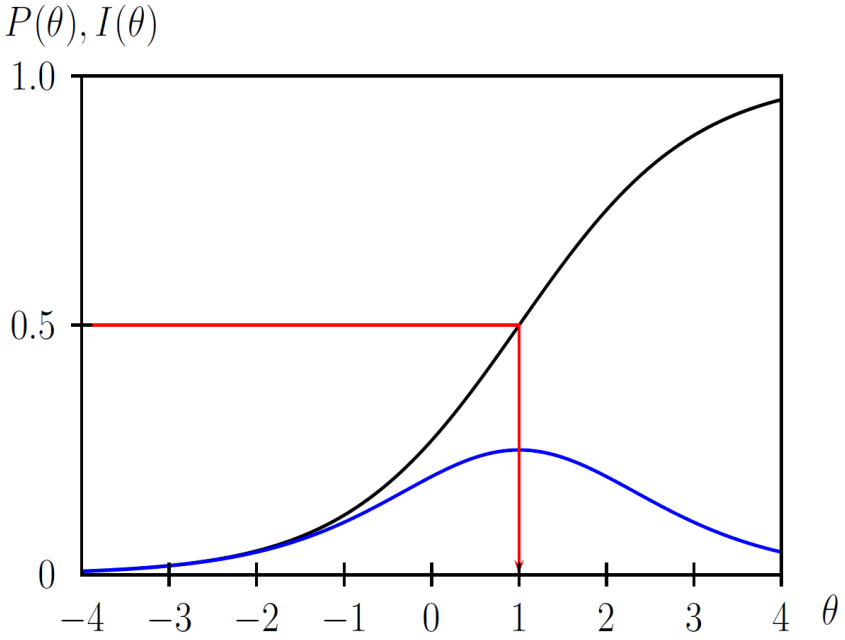
- In many cases, (q – b) is multiplied by a normalizing constant D = 1.7

Item characteristics curve for Rasch model



The Item Information Function of the 1PL model

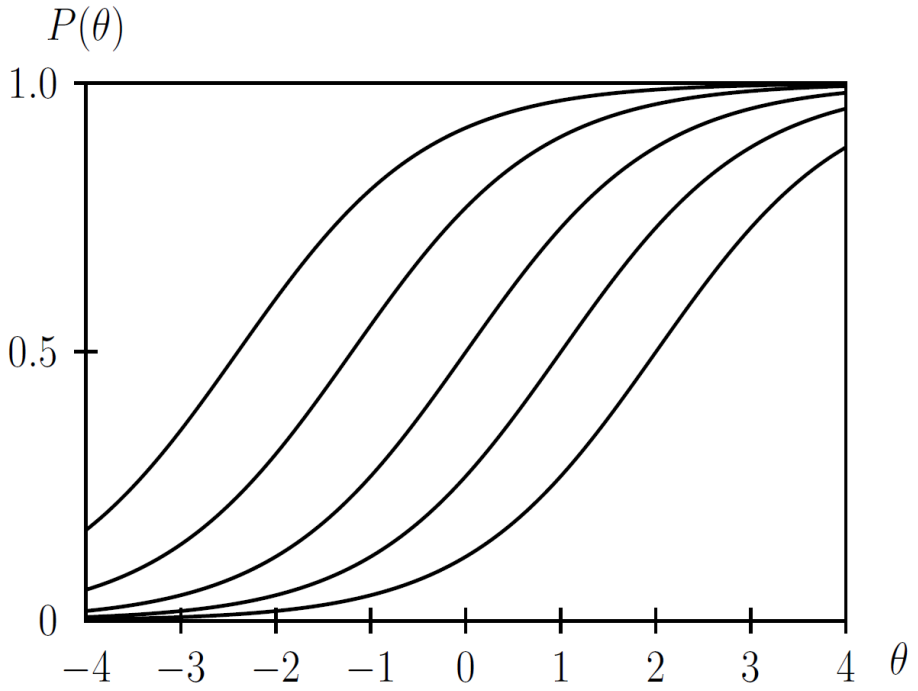
- $Ii(\Theta; bi) = Pi(\Theta; bi)Qi(\Theta; bi)$
- It is easy to see that the maximum value of the item information function is 0.25.



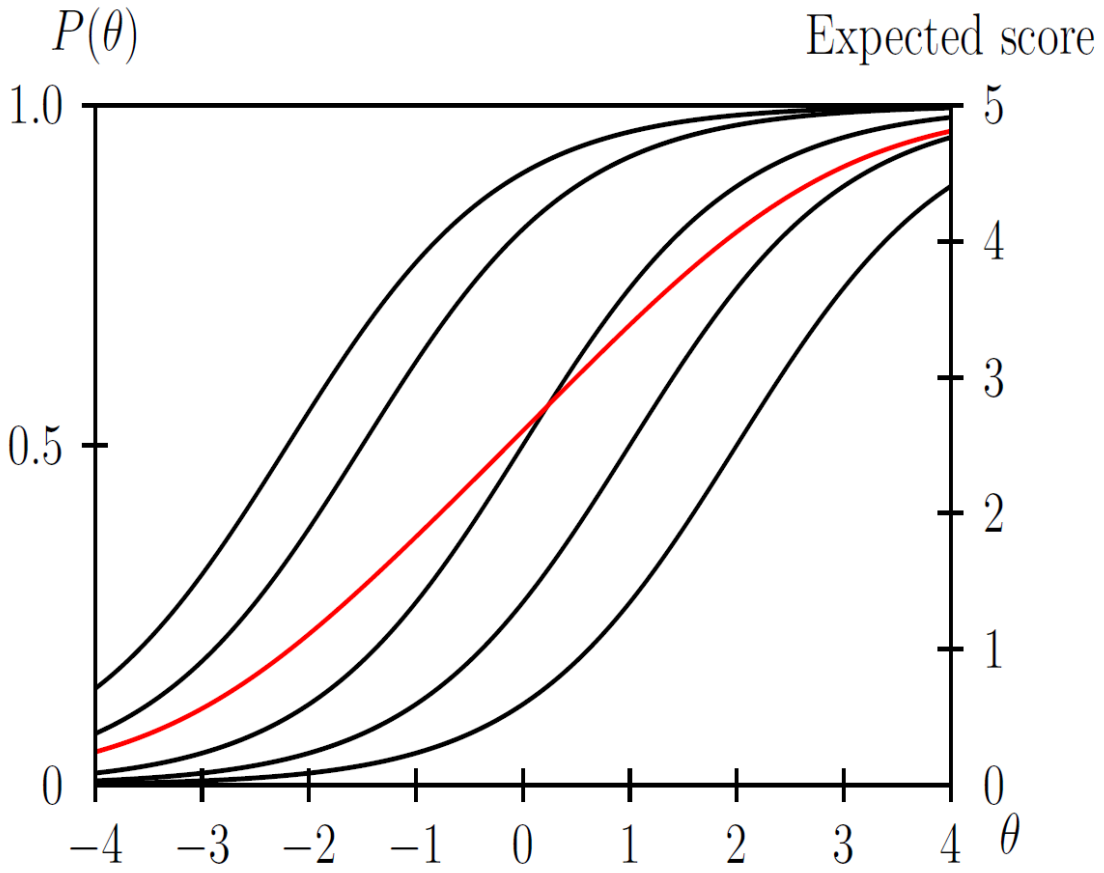
Applet

- http://www.metheval.uni-jena.de/irt/ii1pl.html

The test response function of the 1PL model



test response function (cont.)



Applet

- http://www.metheval.uni-jena.de/irt/trf1pl.html

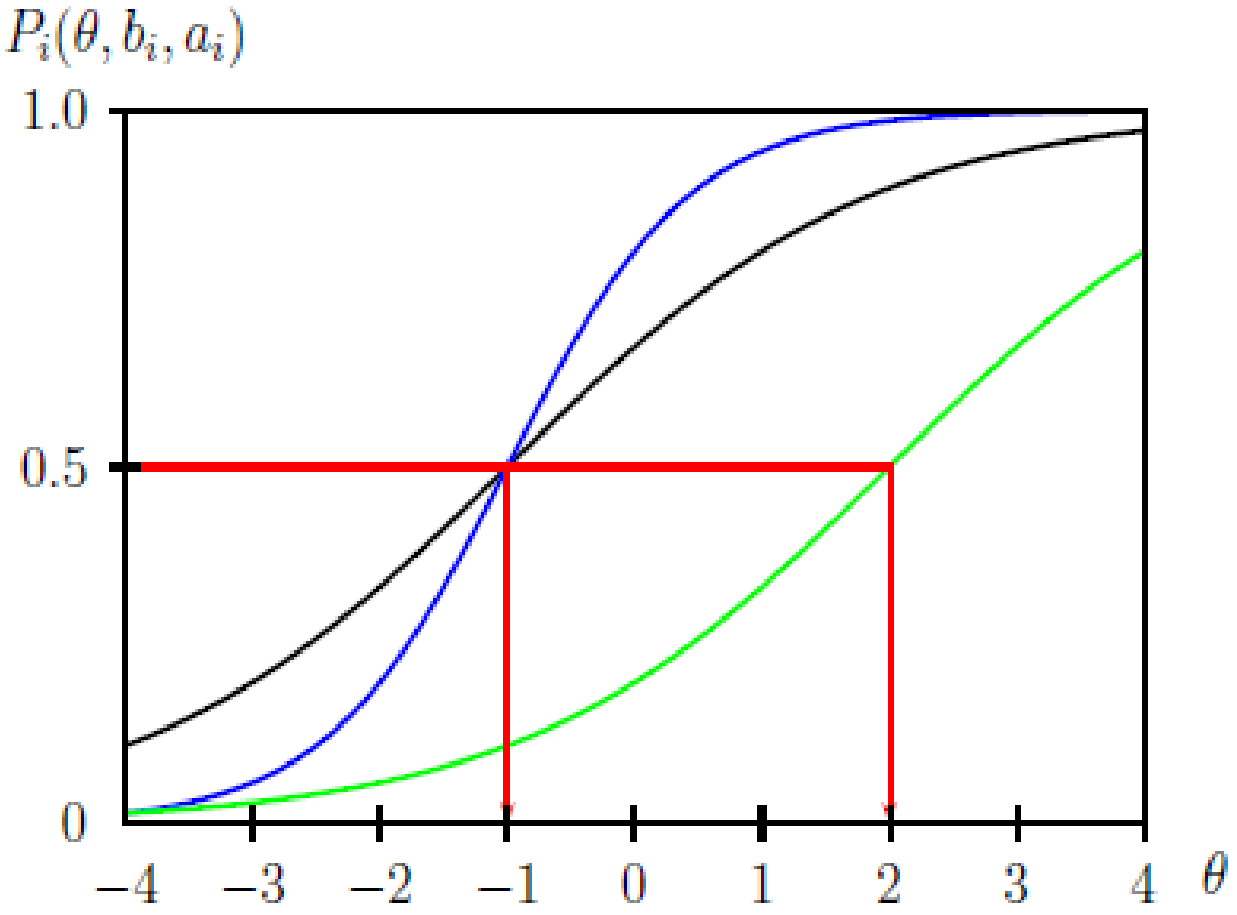
2-parameter logistic model (2-PL)

$$P_j(\theta) = \frac{\exp\left[a_j(\theta - b_j)\right]}{1 + \exp\left[a_j(\theta - b_j)\right]} = \frac{1}{1 + \exp\left\{-\left[a_j(\theta - b_j)\right]\right\}}$$

$$\ln\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right) = a_j(\theta - b_j)$$

- In many cases, $a(\theta - b)$ is multiplied by a normalizing constant $D = 1.7$

Item Characteristic Curves for Three 2PL Items



Applet

- http://www.metheval.uni-jena.de/irt/bbit.html

3-Parameter Logistic Model (3-PL)

- 3-PL assumes that the lower asymptote is more than zero.
- The non-zero lower asymptote represents the fact that anyone has more than zero probability of answering the item correct.
- The non-zero lower asymptote is represented by an additional parameter, c.
- Many people believe 3-PL is the best model for multiple choice test items.

Formulas for 3PL model

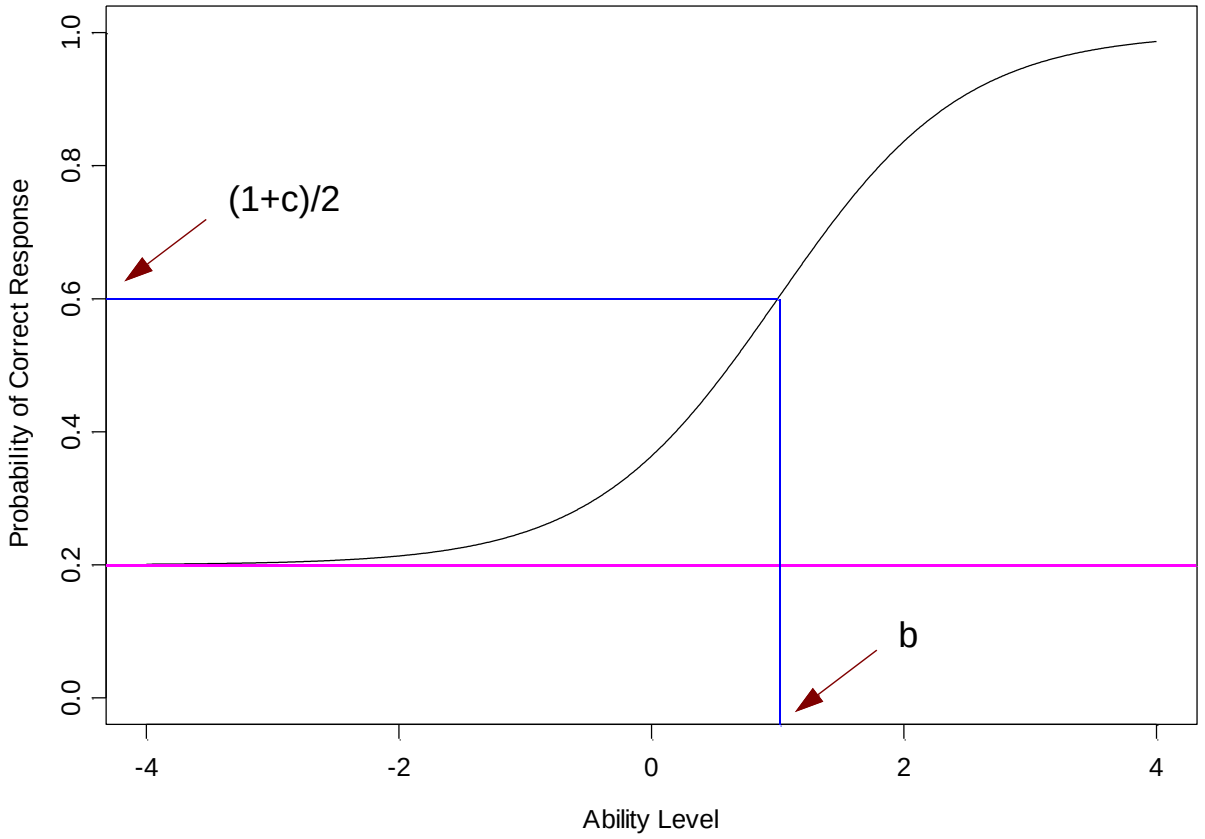
$$P_j(\theta) = c_j + (1 - c_j)\frac{\exp\left[a_j(\theta - b_j)\right]}{1 + \exp\left[a_j(\theta - b_j)\right]}$$

$$= c_j + \frac{1 - c_j}{1 + \exp\left\{-\left[a_j(\theta - b_j)\right]\right\}}$$

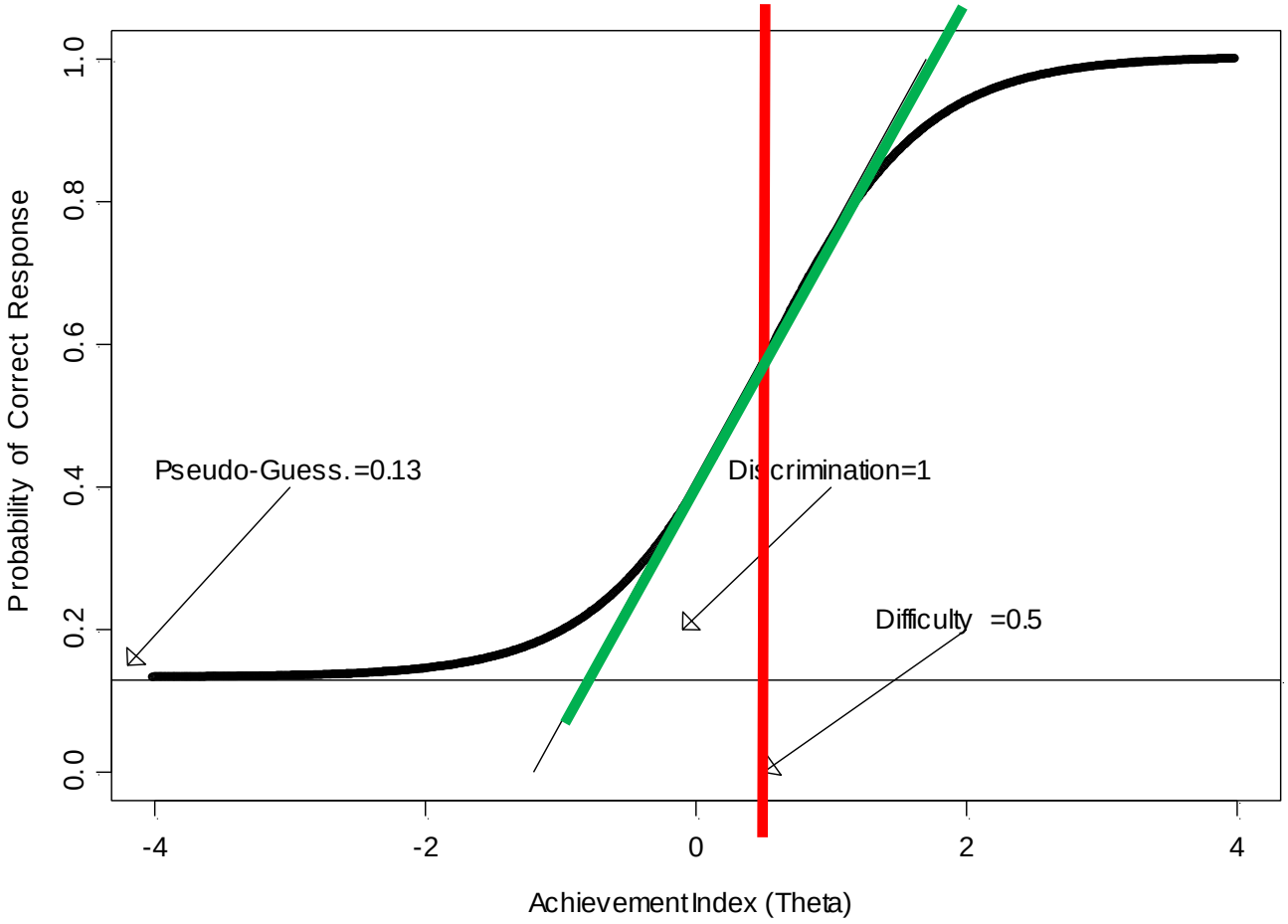
- *c*-parameter is called “pseudo-guessing” parameter.
- The reason it is “pseudo-guessing”, instead of “guessing” is that it could involves more than guessing.
- Again, $a(\theta - b)$ can be multiplied by the normalizing constant $D = 1.7$.

ICC for 3-PL

$a = 0.8, b = 1.0, c = 0.2$



ICC for 3PL


Probability of Correct Response
0.0
0.2
0.4
0.6
0.8
1.0
Pseudo-Guess. =0.13
Discrimination=1
Difficulty =0.5
-4
-2
0
2
4
Achievement Index (Theta)

Item Response Theory
Item Parameters

- As a set, an item's parameters define its relation to the latent trait (Θ).
- Discrimination parameter (a)
  - one per item
  - higher values mean better discrimination among individuals
- Difficulty parameters (b)
  - each item has one for each response category
  - indicates point on Θ where 50% choose that category

How Does It Work ?

- The IRT model has to be (more or less) true, and the item parameters known. Any attempt at testing is therefore preceded by a *calibration study:*
  - the items are given to a succinct number of test persons whose responses are used to estimate the item parameters.

But in Chess ?

- Chess games are fairly unique.
- Every single day, chess games are reaching positions with novel features unseen in the whole history of chess.
  - But

We have chess engines at our disposal to analyze them reliably enough.

Using Engine to evaluate Parameter(s)

- When Evaluating Chess Games / Position:
  - Engine thinks at different depth.
  - We can store that information, and use it to identify the item (here position) parameters.

Configure Chess Engine

- We can configure chess engine to run at multi PV and store the move evaluation at each depth.
- In the next slide we have presented the legal moves evaluation by StockFish for a random position:
- "6k1/5R2/3p1R1p/2pP2p1/1p3nP1/1P5P/P4K2/3r4 w - - 10 44“
- The position appeared in a match between Khuseinkhodzhaev and Gorovykh in 2008.01.09

All relevant information is taken to come from evaluation of each legal moves at each search depth

Depth of Engine Evaluation

Legal Moves

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Rd7	+231	+231	+231	+221	+293	+293	+283	+264	+259	+259	+259	+309	+343	+344	+338	+384	+372	+411	+428
Rf8	+246	+246	+246	+246	+246	+293	+283	+284	+302	+302	+302	+302	+343	+344	+338	+384	+372	+411	+428
h4	+151	+151	+151	+138	+132	+132	+256	+284	+302	+302	+302	+302	+302	+302	+302	+302	+302	+302	+386
Ra7	+236	+222	+222	+194	+182	+163	+235	+235	+235	+235	+266	+266	+266	+300	+300	+305	+337	+354	+365
Rc7	+206	+206	+206	+206	+202	+202	+223	+244	+244	+269	+266	+266	+266	+300	+300	+313	+313	+354	+358
Rb7	+234	+234	+220	+192	+180	+209	+209	+211	+211	+209	+266	+266	+266	+266	+296	+313	+349	+354	+346
Re7	+219	+196	+231	+208	+199	+180	+181	+221	+231	+223	+266	+266	+266	+278	+296	+313	+295	+295	+344
Rf5	+103	+097	+097	+097	+153	+128	+118	+118	+113	+113	+113	+113	+113	+126	+126	+131	+159	+159	+159
Kg3	+093	+093	+093	+093	+093	+093	+093	+091	+072	+072	+072	+085	+090	+096	+096	+096	+113	+113	+113
Kf3	+145	+145	+165	+145	+145	+145	+145	+108	+091	+091	+091	+091	+091	+091	+091	+096	+113	+113	+113
a4	+013	+033	+040	+040	+053	+053	+053	+053	000	+061	+070	+070	+048	+048	+048	+019	+035	+035	+035
a3	-018	+033	+002	+040	+053	+053	+053	+053	+011	000	+070	+070	+028	+048	+048	+019	+035	+035	+035
Rxf4	+057	+056	+068	+068	+059	+033	+011	+011	+001	+001	000	-017	000	+011	+011	000	+021	+021	+003
Ke3	-007	-036	-036	-036	-039	-039	-039	-075	-071	-104	-107	-115	-118	-148	-139	-136	-136	-136	-136
Rxd6	-340	-340	-369	-369	-359	-362	-341	-347	-347	-398	-412	-416	-417	-441	-440	-454	-454	-454	-454
Rxh6	-333	-333	-333	-333	-459	-367	-375	-411	-411	-444	-441	-441	-441	-454	-454	-454	-454	-454	-454
Re6	-380	-380	-414	-421	-460	-460	-475	-475	-458	-467	-494	-502	-537	-537	-537	-523	-548	-548	-572
Rg6	-419	-414	-421	-460	-460	-475	-426	-455	-458	-467	-494	-502	-537	-537	-523	-523	-548	-548	-572
Rh7	-390	-390	-390	-390	-423	-423	-445	-460	-509	-505	-482	-456	-487	-487	-522	-522	-522	-621	-621
Rg7	-420	-420	-420	-453	-453	-468	-468	-523	-523	-523	-523	-523	-529	-559	-552	-581	-585	-585	-631

How to calculate the item parameters

- For Chess,
- **b** represents difficulty of the chess position.
- **a** represents how good the move is for using it as a discriminator for identifying the strength of a player.
- **c** represents the pseudo-guessing parameter.

Estimating Position Difficulty or ‘b’

- From the Move Evaluation Matrix, For any particular depth, we can say the move is difficult if the scaled delta values for sub-optimal moves are close to the best move at that depth.
- Lower Variation indicates Higher difficulty, where higher Variation indicates a clear choice move and corresponds to lower difficulty.
- The difficulty of a move is a function of difficulty at each depth with non decreasing weight for each depth.

Estimating Position Discrimination or ‘a’

- A position can discriminate players better, if there is a *swing move*.
- A swing move is an apparently attractive move that proves to be worse at higher depth, or a good move that looks poor until high depth.
- This value can be calculated by checking from right to left in a row and summing up the absolute changes in value.

Positions to Drop or Prune

- Scenarios like having only one legal move, or having only one reasonable move (such as an “obvious” recapture).
- Positions all of whose reasonable moves have little value distinction. Can be in a dead-drawn game, or a game where one side is way ahead of the opponent.

When can be applied

- When the model is appropriate and the estimates of the item parameters are reasonably accurate, IRT promises that the testing will have certain attractive properties. Most important, we can ask different examinees (chess players or engines) about different items (positions), and get comparable estimates of ability.

The Goal “Connexion”

- Use well-researched properties of the world chess rating system and its population, together with features of the chess model, to carry over a *natural* grading and skill-rating system for multiple-choice examinations and similar human decision instances.

- [discuss possible concrete applications]

Extra: Detecting Chess Cheating

- Is it possible ?
- Yes, by this model
  - The accused will be asked to solve few positions to find the best move in isolated condition.
  - The position will be selected that best matches the performance of those game for which the player is accused for.
  - By test response function, we can detect the player's true ability.