

# **AN INTRODUCTION TO ITEM RESPONSE THEORY AND ITS APPLICATIONS TO HEALTH ASSESSMENTS**

Qiong (Joan) Wu

Harvard Center for Population and Development Studies

---

INDEPTH-SAGE WORKSHOP

April 20, 2010

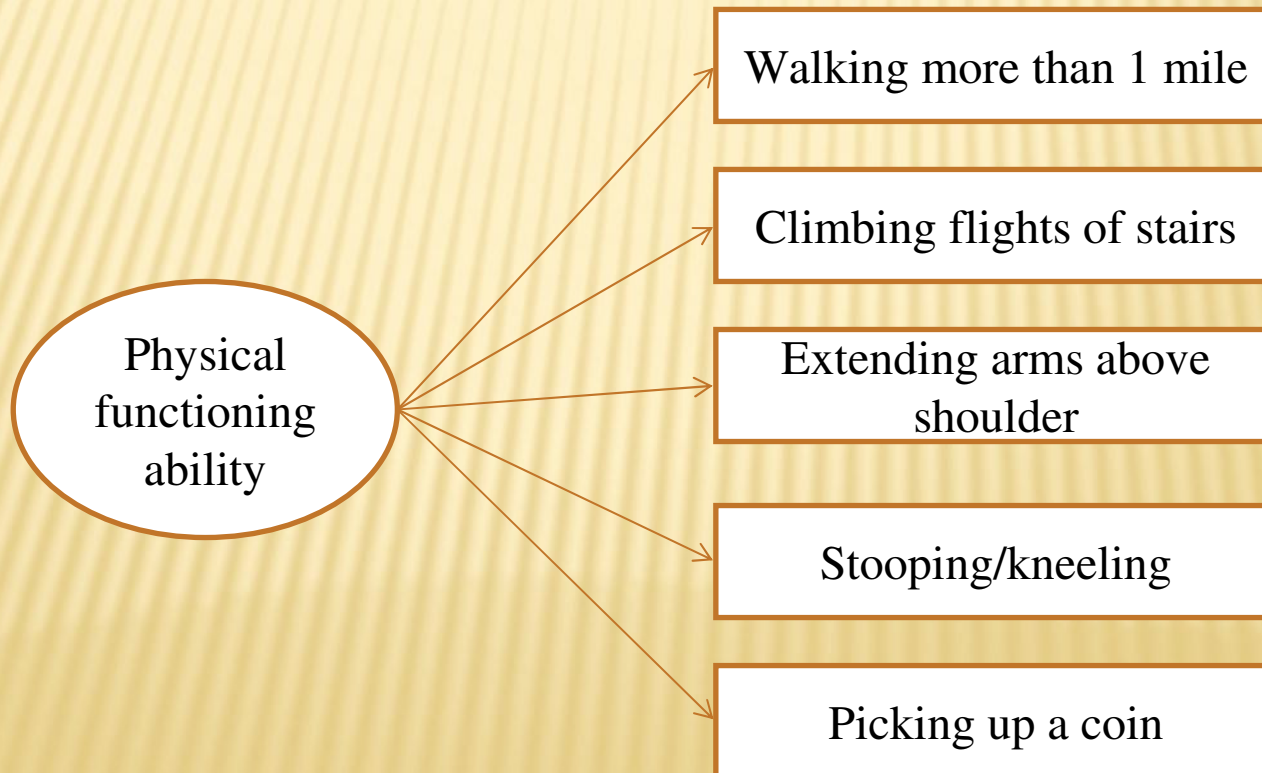
# ITEM RESPONSE THEORY

---

- ✘ IRT vs Classical test theory (CTT)
  - + CTT: focuses test scores  
observed score = true score + error ( $O=T+E$ )
  - + IRT: focuses on individual item characteristics
- ✘ IRT is a scaling method
  - + Assigns numerical scores based on a set of item responses
- ✘ IRT is an item analysis tool
  - + Evaluates quality of individual items based on estimated item parameters

# ITEM RESPONSE THEORY

- ✘ Also called latent trait models
  - + Many variables in health assessments cannot be measured directly, such as physical functioning ability, fatigue, depression etc.



# ALTERNATIVE WAYS OF DERIVING SUCH SCORES

- ✘ Raw total/Raw percentage
  - + Items weighted equally
  - + Item dependent
  
- ✘ Factor scores
  - + Usually assume continuous observed variables

# CHARACTERISTICS OF IRT SCALED SCORES

- ✘ Designed for dichotomous or polytomous items
- ✘ Pattern scoring instead of number scoring
- ✘ When the assumptions are met
  - + Scores are item invariant
- ✘ Considered equal-interval, so preferred scores for longitudinal analysis

# PARAMETERS IN DICHOTOMOUS IRT

- ✘ Person parameter (  $\theta$  )
  - + Latent scores, theta scores (mean=0, SD=1)
- ✘ Item parameters
  - + Item difficulty/location (b), usually  $-3 < b < 3$
  - + Item discrimination (a),  $\geq 0$
  - + Pseudo-guessing (c),  $0 \leq c \leq 1$
- ✘ Item difficulty/location and theta scores on the same scale

# MATHEMATICAL MODELS FOR BINARY DATA

## ✦ One-parameter IRT (Rasch model)

$$P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}}$$



$$\log \frac{P_i(\theta)}{1 - P_i(\theta)} = \theta - b_i$$

ability score

item difficulty

## ✦ Two-parameter IRT

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$



$$\log \frac{P_i(\theta)}{1 - P_i(\theta)} = a_i(\theta - b_i)$$

item discrimination

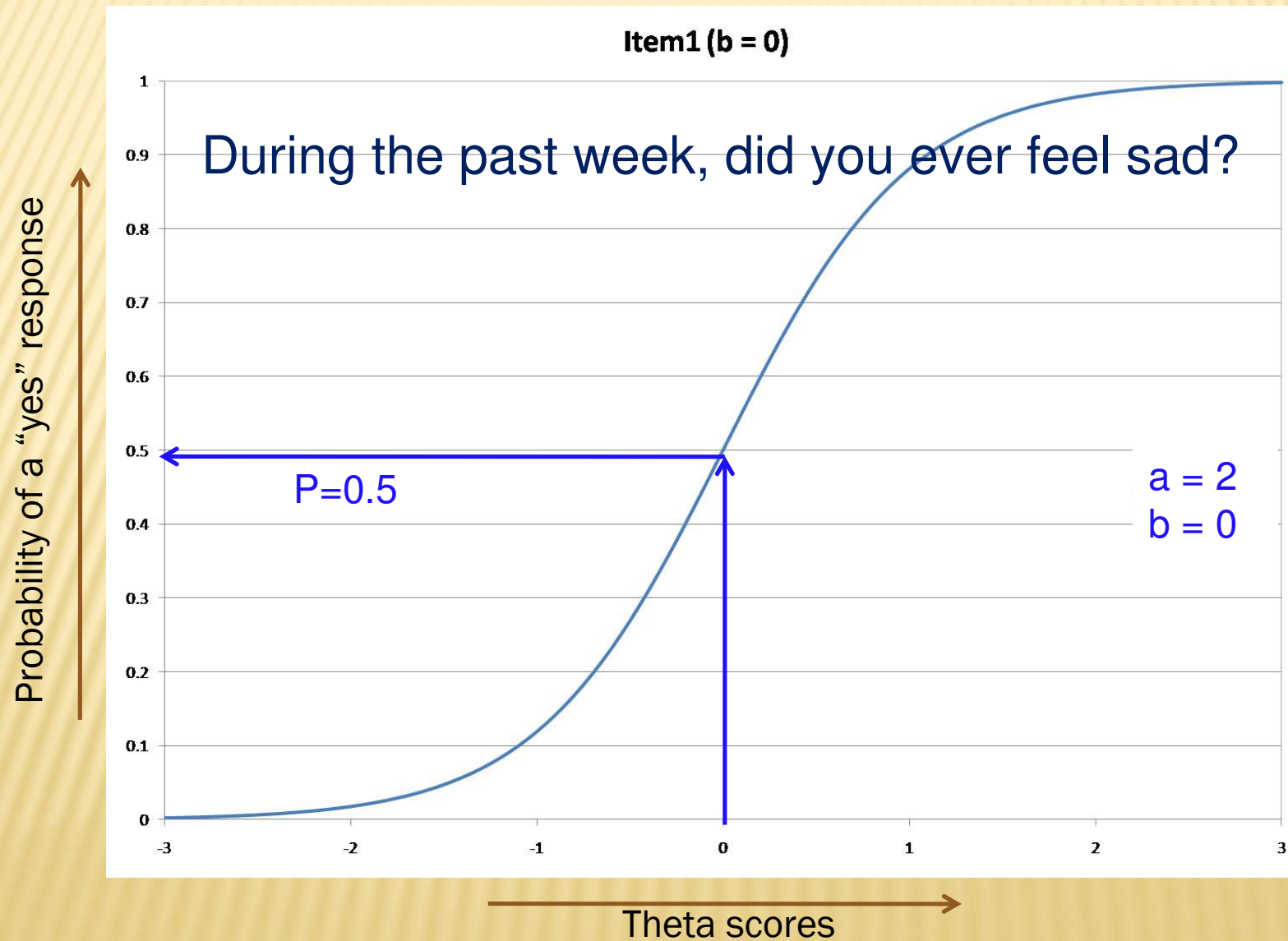
## ✦ Three-parameter IRT

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

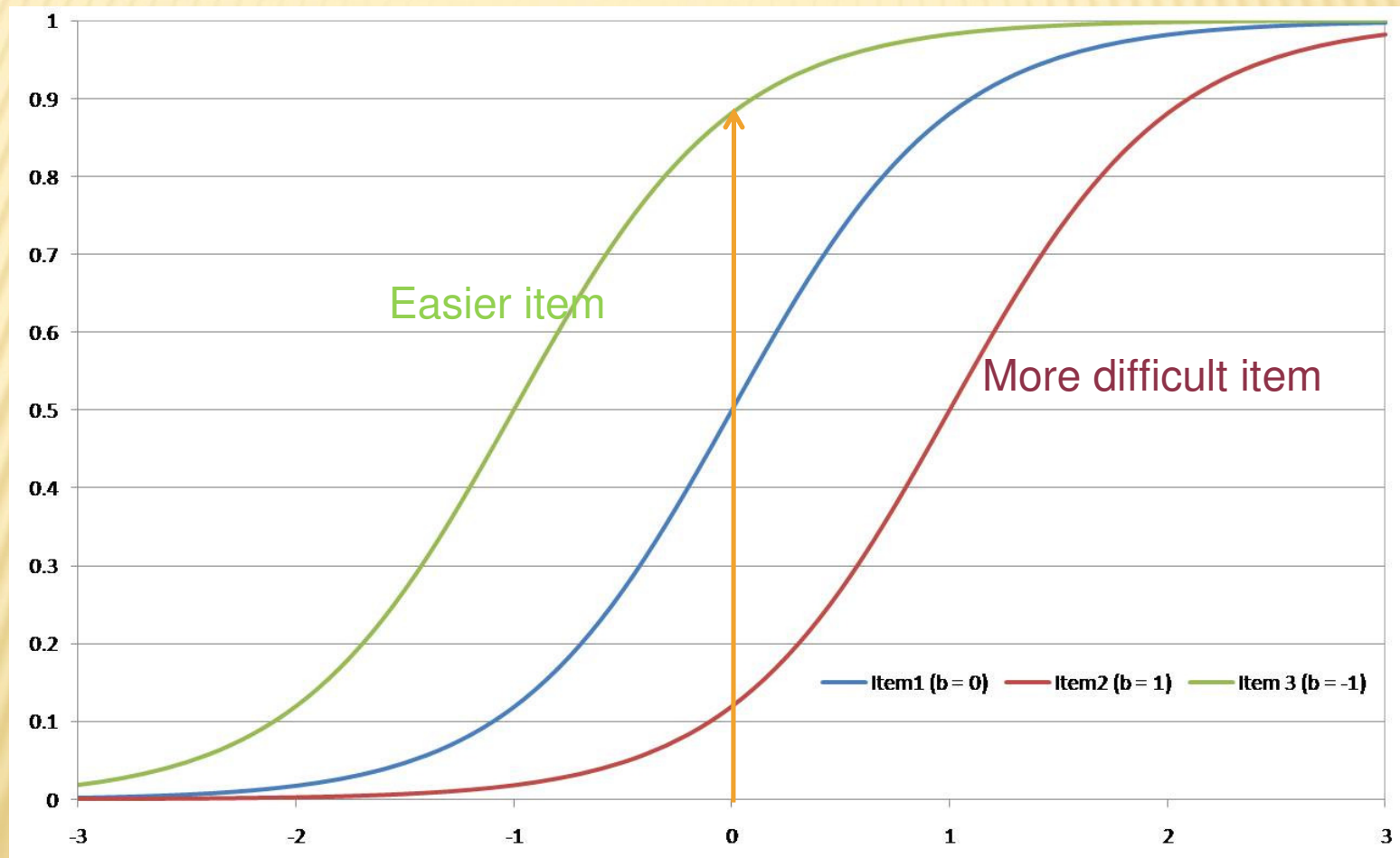


$$\log \frac{P_i(\theta) - c_i}{1 - P_i(\theta)} = a_i(\theta - b_i)$$

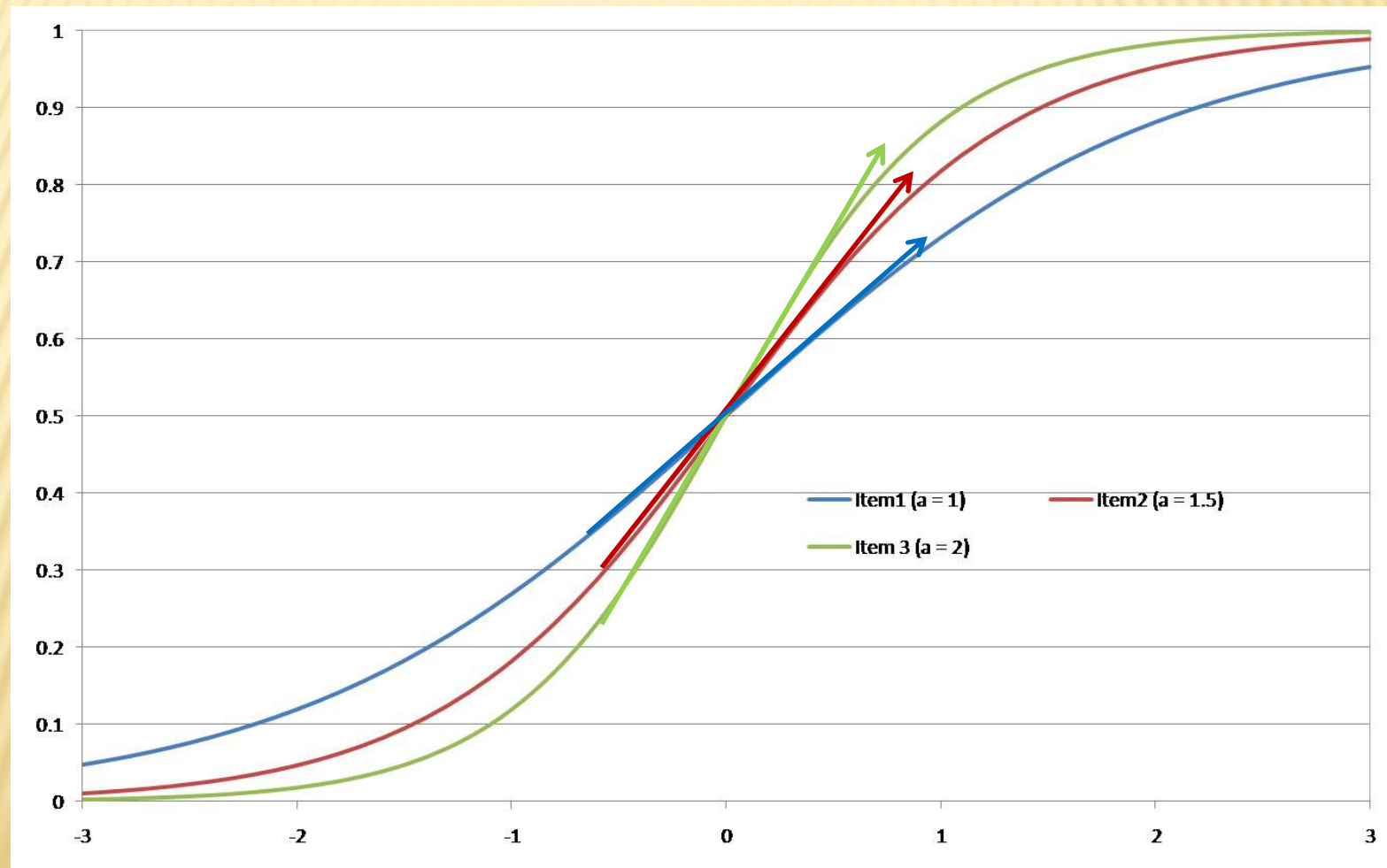
# ITEM CHARACTERISTIC CURVES (ICC)



# ITEMS WITH DIFFERENT DIFFICULTY/LOCATION



# ITEMS WITH DIFFERENT DISCRIMINATION



# POLYTOMOUS IRT

	In the last 30 days, how much difficulty did you have ...	None	Mild	Moderate	Severe	Extreme/ cannot do	N/A
Q1025	... in standing for long periods (such as 30 minutes)?	1	2	3	4	5	98

- ✘ Deals with items with more than 2 response options
- ✘ Model probability of each response option conditional on latent trait

# POLYTOMOUS IRT (GRADED RESPONSE MODEL)

## ✘ Parameters

- + Latent trait: theta
- + Discrimination parameters
- + Threshold parameters

## • Dichotomous IRT

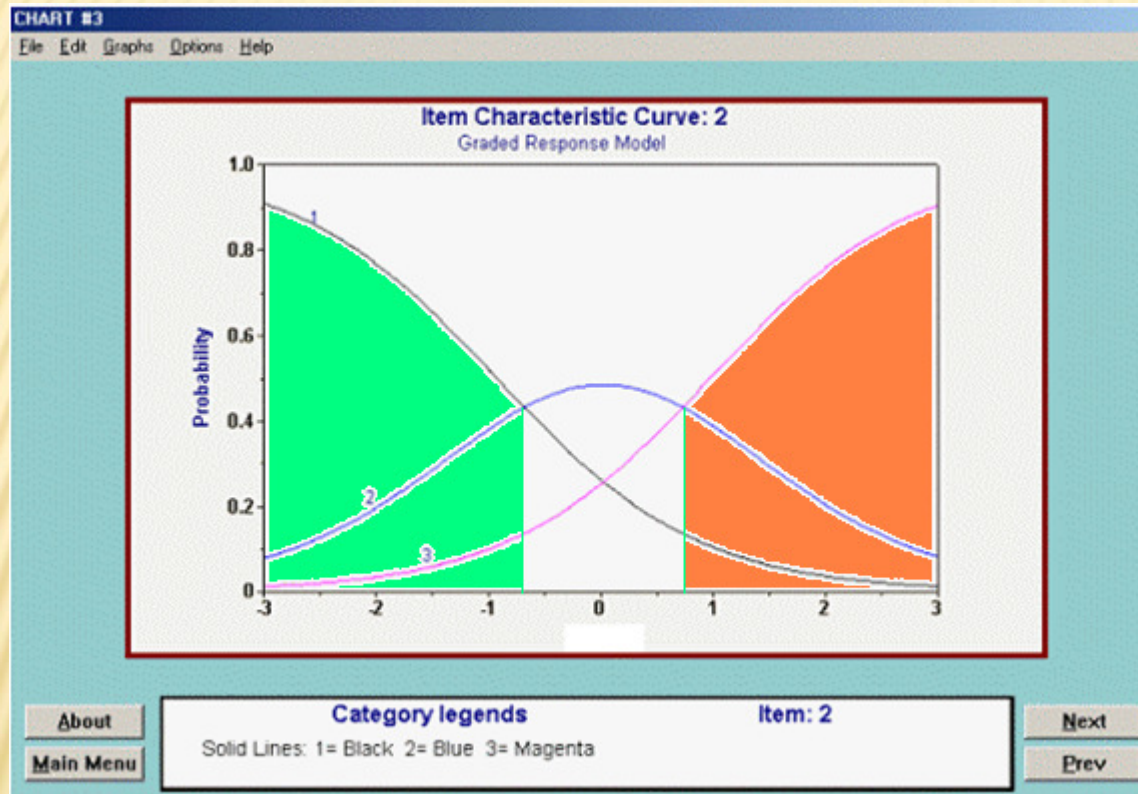
- Theta
- Discrimination par
- Difficulty/location par

$$P(X_i = k | \theta) = \frac{1}{1 + \exp(\alpha_i(\theta - \delta_{ix}))} - \frac{1}{1 + \exp(\alpha_i(\theta - \delta_{ix-1}))}$$

## ✘ Measurement questions

- + How many categories are needed?
- + Are response categories really different one another?

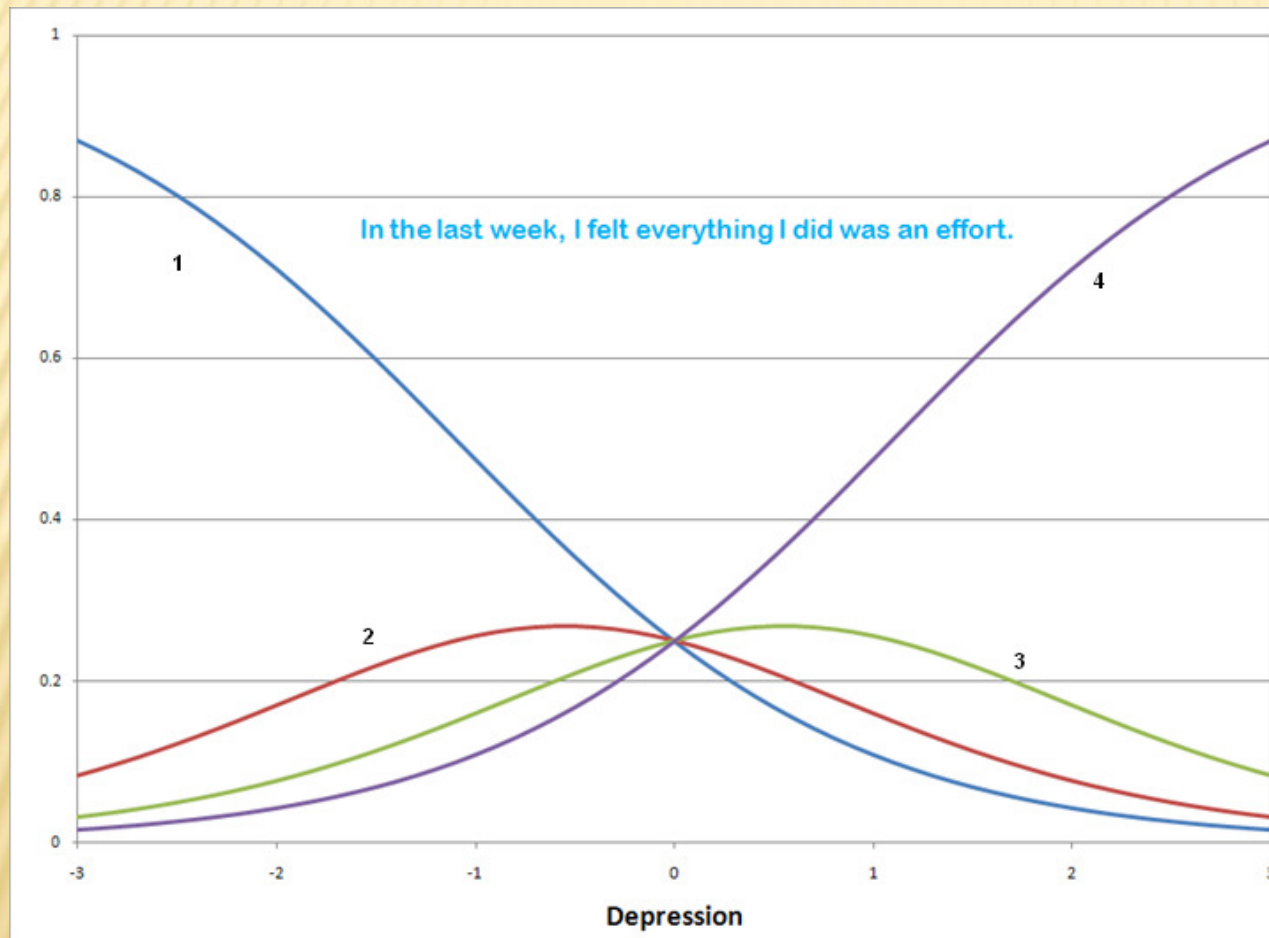
# AN IDEAL ICC FOR A 3-OPTION ITEM



1: rarely  
2: sometimes  
3: most or all of the time

How often do you feel sad?

# A REAL ITEM



**1: rarely or none  
of the time;**

**2: some or a little  
of the time**

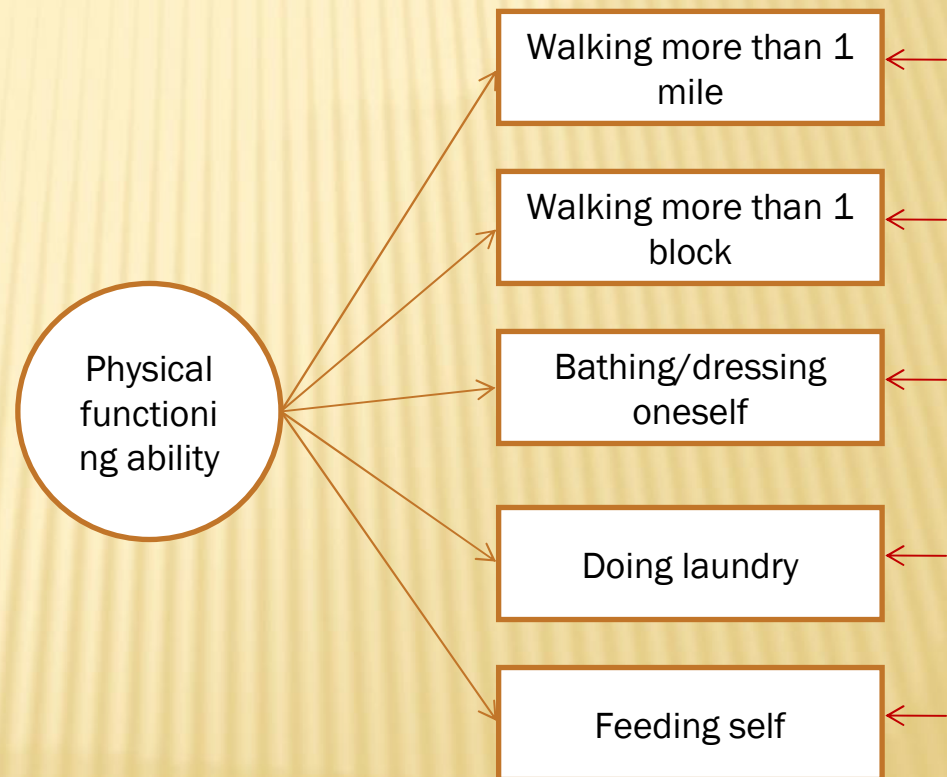
**3: occasionally or a  
moderate amount;**

**4: most or all of the  
time**

# ASSUMPTIONS OF IRT

- ✘ Unidimensionality
- ✘ Local independence  
(residual covariance=0)

Unidimensionality implies  
LI, not vice versa.



# APPLICATIONS OF IRT

---

- ✘ Test construction and item analysis
- ✘ Computerized adaptive testing (CAT) and item banking
- ✘ Test equating/linking
- ✘ Differential item functioning/item bias

# TEST CONSTRUCTION AND ITEM ANALYSIS

## ✘ Basic rule

- + Select items with appropriate  $b$  & high  $a$  for dichotomous items
- + Check ICCs for polytomous items

# CAT AND ITEM BANKING

---

- ✘ Basic idea
  - + Items tailored to individuals' trait levels
- ✘ Why do we need it?
  - + Reduce test length, minimize fatigue
  - + Minimize floor/ceiling effects
- ✘ Challenges
  - + Large item pool
  - + Large sample of subjects for initial item calibration

# CAT AND ITEM BANKING (2)

- ✘ Different types of adaptive testing
  - + Two-stage testing

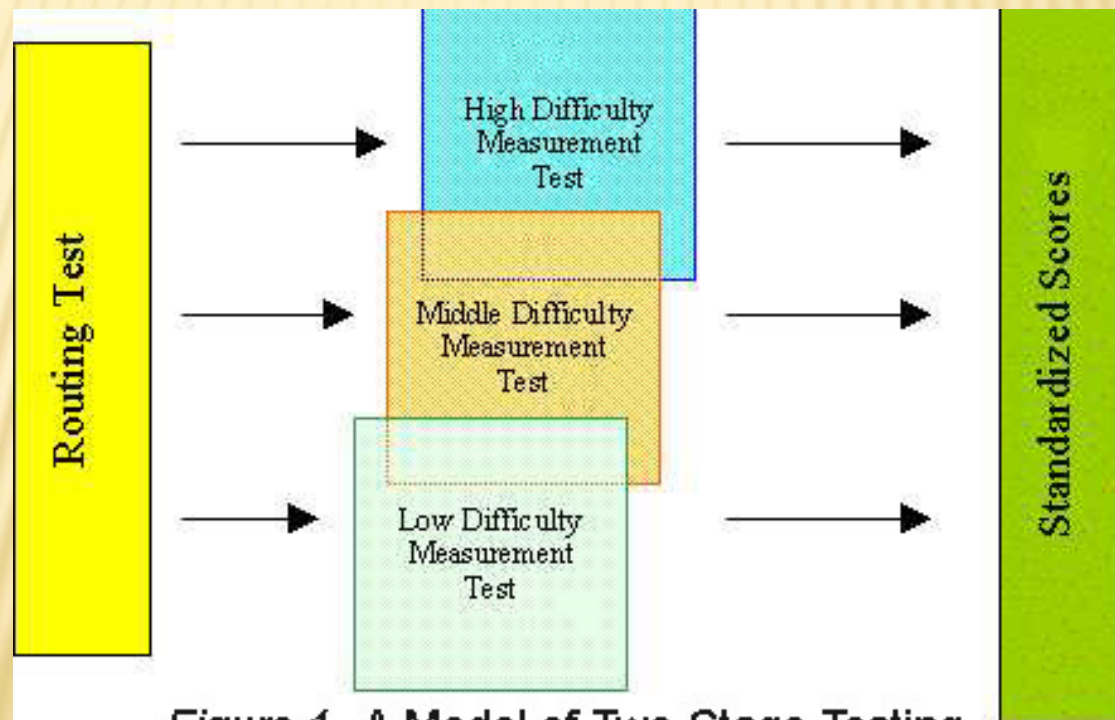
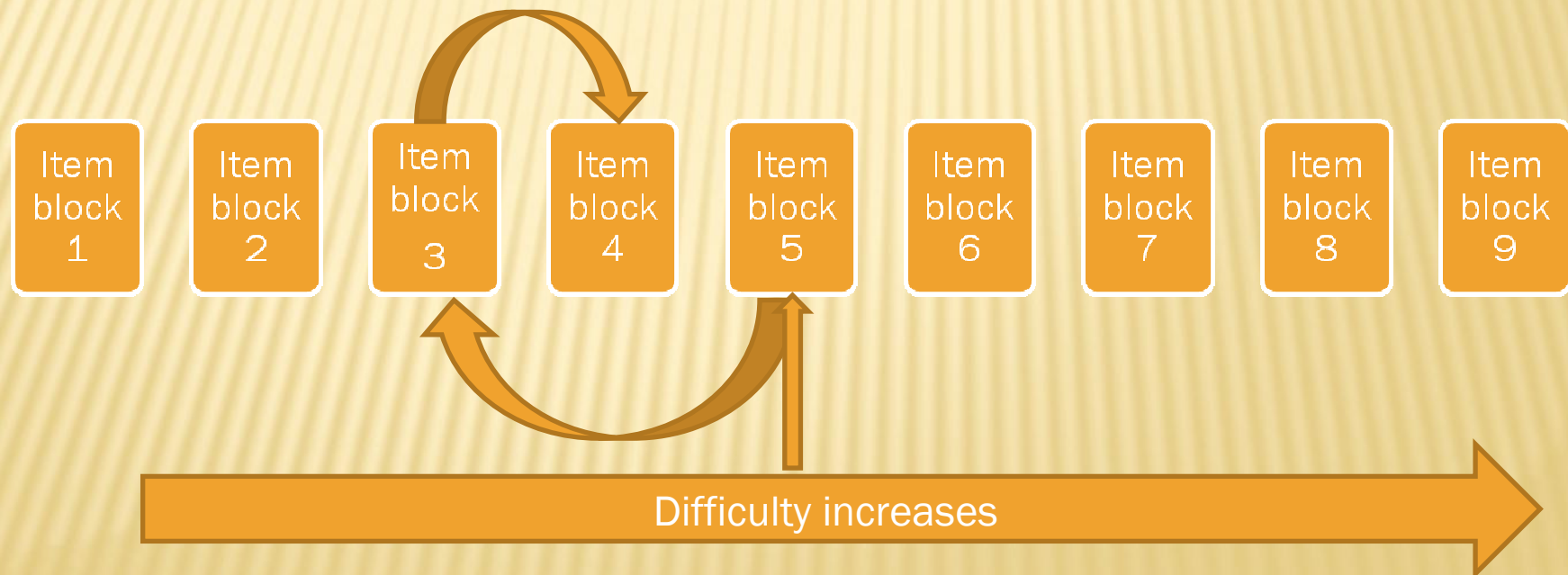


Figure 1. A Model of Two-Stage Testing

## CAT AND ITEM BANKING (2)

- ✦ Different types of adaptive testing
  - + Multi-stage testing



# TEST EQUATING/LINKING

---

- × Definition

- × Setting a common metric for scores from tests composed of different sets of items

- × When do we need it?

- + When two groups of subjects take different test forms

- × How do we do it?

- + Needs a valid link, either through common-item design, or common subject sample design

# TEST EQUATING/LINKING

+ Linking test scores from two test forms (common item design)

✘ Group A



✘ Group B



✘ Fixing item parameters to be equal across two forms

# DIFFERENTIAL ITEM FUNCTIONING

---

## × Definition

- + different groups of subjects display different probabilities of endorsing a response option conditional on latent trait

## × Why do we care?

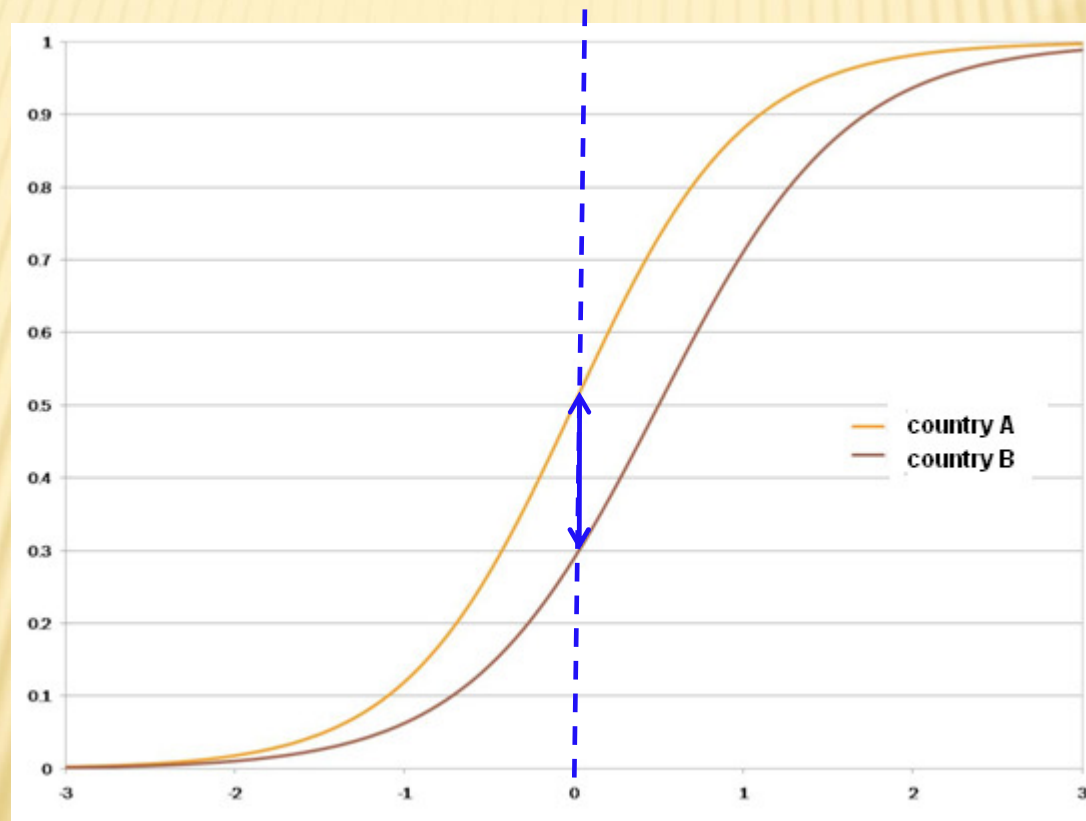
- + Test/survey bias when those items are included

## × Anchor items

- + Use anchor items to equate the groups on the latent trait.  
IRT often use rest items

# DIF ACROSS GENDER GROUPS

- ✘ “Do you have any difficulties with making phone calls?”



Functional ability →

# DEMYSTIFYING IRT

---

- ✘ High correlation between number right scores and IRT scaled scores
- ✘ IRT estimates can be sample dependent (DIF)
- ✘ Large-sample technique
  - + >200 for 1PL models; >400 for 2PL models; >600 for 3PL models

# WHEN IS IRT RECOMMENDED

---

- ✘ Computerized adaptive testing and test construction
  - + A large item pool AND
  - + A large number of subjects
- ✘ Test scoring when item discrimination vary a lot
- ✘ Cross-cultural comparison
- ✘ Longitudinal analysis

# OTHER ISSUES

---

- ✘ Multidimensional IRT
- ✘ Specialized software program
  - + BILOG, MULTILOG, Mplus
  - + R module