International Symposium on the Applications of Item Response Theory in Health Measurement

IRT Concepts

Maria Orlando Edelen

RAND Corporation

September 17, 2013, Amsterdam, the Netherlands.

Overview

- Model assumptions
- Model fit
- Illustrative example
  - Evaluation of model assumptions and fit
- IRT scoring approaches
- DIF detection and evaluation
- Questions

Model Assumptions - Unidimensionality

- **Assumes covariance among the items can be explained by a single underlying dimension (the trait that is being measured by the item set)**
- **Can evaluate through examination of FA output**
  - **Magnitude of eigenvalues (# >1, ratio of first to second, scree plots)**
  - **Magnitude of item loadings on first factor**
  - **Fit indices, modification indices (from CFA)**
- **FA models should account for categorical response data**

Model Assumptions – Local Independence

- **Assumes no additional systematic covariance among items after accounting for their relationship with the underlying trait(s)**
- **Local Dependence (LD) can arise among items with similar stems, similar content, or that are placed sequentially**
- **Sometimes can identify potential LD by examining modification indices in CFA**
- **Problematic LD often results in inflated IRT slope parameter estimates for the violating item pair**
- **The Chen-Thissen χ^2 index indicates extent of LD for item pairs based on IRT calibration**

Model Fit

- **An implicit assumption of IRT applications is that the model fits the data**
- **Choose the right IRT model**
  - **Number and type of response options, need for varying slopes**
- **Evaluate model fit**
  - **Overall, LD**
  - **Item-level fit, person-level fit**

Evaluating Overall Model Fit

- **Overall model fit is difficult to evaluate directly**
- **Can evaluate fit of nested models (e.g., to see if varied slope values are needed)**
- **Can examine solution – make sure item parameter estimates and standard errors look reasonable**
- **Can examine matrix of LD indices**

Evaluating item- and person-level fit

- **Item-level fit**
  - **Can examine S-X^2 values (in IRTPRO)**
  - **Can generate graphical representations of item fit to identify obvious violations**
- **Person-level fit**
  - **Person fit indices are meant to evaluate the consistency of response patterns**
  - **Often used in educational testing to identify cheating**
  - **Have also been used in with personality items to identify socially desirable responding, malingering, random responding etc.**

Illustrative Example

- **Data come from 442 adolescents in treatment for substance abuse**
- **Items are intended to measure treatment process**
- **For purposes of this example, 8 items are analyzed**
  - **5-point Likert-type response scale indicating extent of agreement (*0=not at all – 4=completely*)**
  - **All items positively worded so higher score means more agreement and more 'progress' re treatment process**

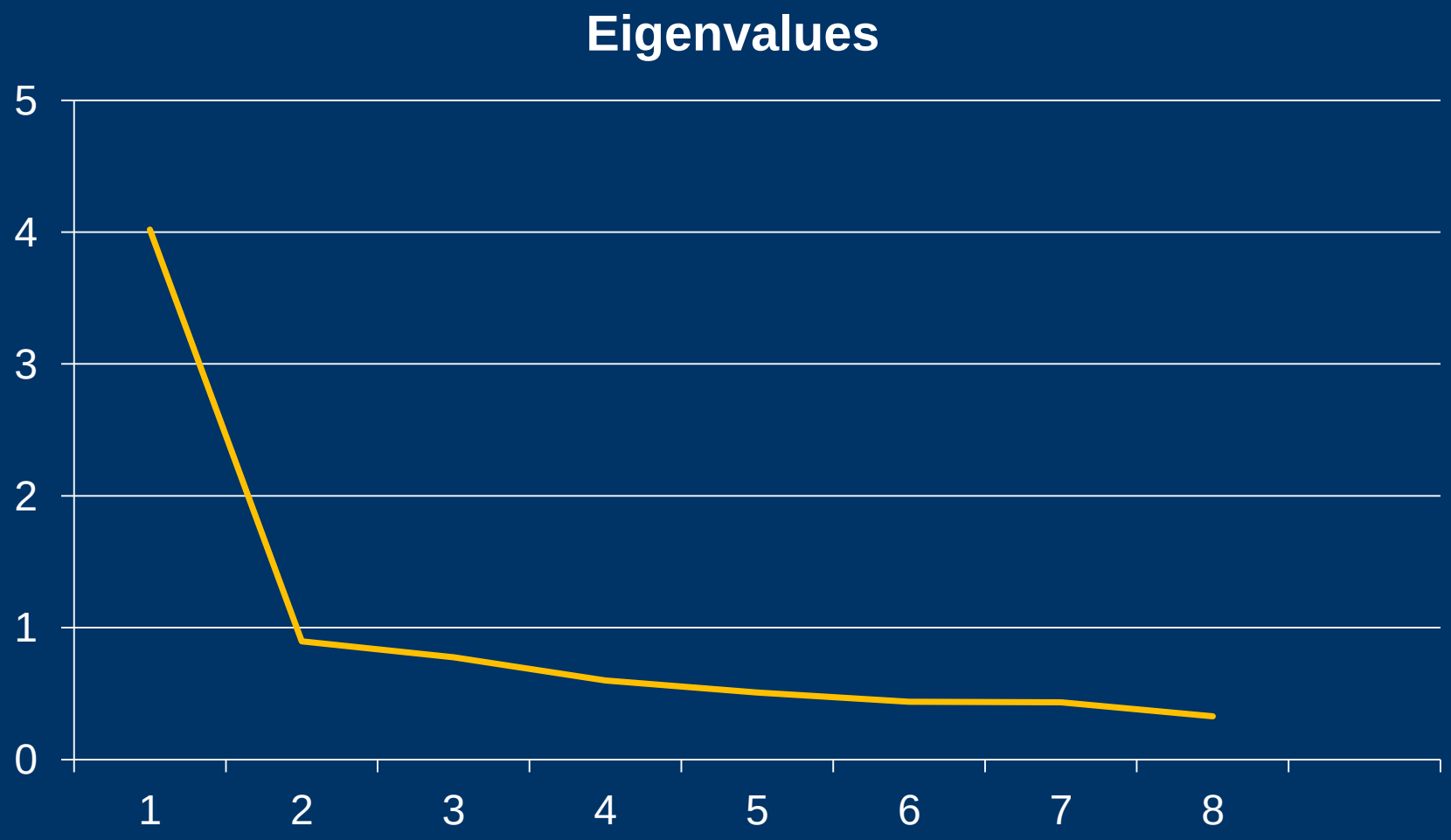
Items

1. *The people I care about expect me to make positive changes in my life*
2. *I can have fun without using drugs or alcohol*
3. *I notice that people here can support each other despite conflict*
4. *It is important for me to see my part in the problems I have*
5. *I am actively involved in my treatment plan*
6. *I am developing new interests*
7. *I don't need to use drugs or alcohol to relax and hang out with my friends*
8. *I try to help my peers in this program make better choices*

Approach

- EFA to check unidimensionality assumption
- IRT calibration to examine
  - Model solution
  - Item fit
  - LD
- Removal of items as needed based on above
- Repeat IRT calibration with reduced item set

EFA Results – Scree plot



EFA Results – Fit and Loadings

FIT
$\chi^2 = 63.048, p<.001$
RMSEA = 0.070 (0.051 – 0.090)
CFI = 0.979
TLI = 0.970

Item	Loading
1	0.50
2	0.63
3	0.56
4	0.72
5	0.81
6	0.74
7	0.54
8	0.74

IRT Calibration of 8 items - parameters

Item	*a (se)*	b_1 *(se)*	b_2 *(se)*	b_3 *(se)*	b_4 *(se)*
1	1.00 (0.15)	-5.15 (0.80)	-3.58 (0.48)	-2.82 (0.37)	-1.06 (0.16)
2	1.43 (0.15)	-2.57 (0.25)	-1.75 (0.17)	-0.75 (0.11)	0.17 (0.09)
3	1.19 (0.13)	-1.65 (0.18)	-0.52 (0.11)	0.78 (0.13)	1.98 (0.21)
4	1.91 (0.19)	-2.41 (0.20)	-1.57 (0.13)	-0.69 (0.08)	0.23 (0.08)
5	2.52 (0.26)	-1.95 (0.15)	-1.29 (0.10)	-0.46 (0.07)	0.35 (0.08)
6	2.07 (0.20)	-1.73 (0.14)	-1.13 (0.10)	-0.32 (0.07)	0.82 (0.10)
7	1.15 (0.13)	-1.64 (.19)	-1.04 (0.14)	-0.27 (0.10)	0.55 (0.12)
8	2.07 (0.20)	-1.89 (0.15)	-1.08 (0.10)	-0.25 (0.07)	0.7 (0.09)

Marginal Reliability: 0.85

IRT Calibration of 8 items – item fit

Item	X^2	*df*	*Probability*
1	50.13	43	0.2112
2	56.95	63	0.6911
3	73.35	67	0.2772
4	60.96	57	0.3347
5	53.29	50	0.348
6	52.28	58	0.6876
7	91.21	69	0.0379
8	49.16	55	0.697

IRT Calibration of 8 items – LD matrix

Item	*Marginal X^2*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	0.2							
2	0.4	0.8						
3	0.4	2.3	2.4					
4	0.5	1.4	3.1	2.9				
5	0.6	3.7	-0.5	5.1	1.2			
6	0.6	3.6	2.6	2.5	2.7	0		
7	0.4	3.6	12.4	0.8	1.8	5.8	5.3	
8	0.2	0.1	0.8	2.5	0.5	1.7	5.6	2.2

Items

1. *The people I care about expect me to make positive changes in my life*
2. *I can have fun without using drugs or alcohol*
3. *I notice that people here can support each other despite conflict*
4. *It is important for me to see my part in the problems I have*
5. *I am actively involved in my treatment plan*
6. *I am developing new interests*
7. *I don't need to use drugs or alcohol to relax and hang out with my friends*
8. *I try to help my peers in this program make better choices*

Items Reduced

1. ~~*The people I care about expect me to make positive changes in my life*~~
2. *I can have fun without using drugs or alcohol*
3. *I notice that people here can support each other despite conflict*
4. *It is important for me to see my part in the problems I have*
5. *I am actively involved in my treatment plan*
6. *I am developing new interests*
7. ~~*I don't need to use drugs or alcohol to relax and hang out with my friends*~~
8. *I try to help my peers in this program make better choices*

IRT Calibration of 6 items - parameters

Item	*a (se)*	b_1 *(se)*	b_2 *(se)*	b_3 *(se)*	b_4 *(se)*
2	1.34 (0.15)	-2.67 (0.27)	-1.82 (0.18)	-0.78 (0.11)	0.18 (0.10)
3	1.20 (0.13)	-1.64 (0.18)	-0.52 (0.11)	0.78 (0.12)	1.97 (0.21)
4	1.91 (0.19)	-2.41 (0.20)	-1.57 (0.13)	-0.68 (0.09)	0.24 (0.08)
5	2.64 (0.28)	-1.93 (0.14)	-1.27 (0.10)	-0.45 (0.07)	0.35 (0.08)
6	2.06 (0.20)	-1.73 (0.14)	-1.13 (0.10)	-0.32 (0.08)	0.82 (0.10)
8	2.03 (0.20)	-1.91 (0.15)	-1.09 (0.10)	-0.24 (0.07)	0.71 (0.10)

Marginal Reliability: 0.84

IRT Calibration of 6 items – item fit

Item	X^2	*df*	*Probability*
2	54.59	55	0.4911
3	55.39	54	0.4232
4	36.41	46	0.8438
5	52.16	43	0.1593
6	42.09	49	0.7475
8	64.88	48	0.0525

IRT Calibration of 6 items – LD matrix

Item	*Marginal X^2*	*2*	*3*	*4*	*5*	*6*
2	0.4					
3	0.4	2.2				
4	0.5	3.1	2.9			
5	0.5	-0.6	5.2	1.1		
6	0.5	2.6	2.4	2.7	0	
8	0.2	0.7	2.5	0.5	1.8	5.5

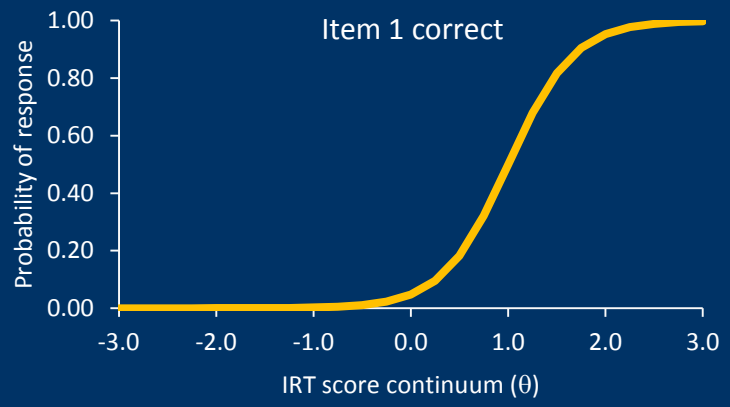
IRT Model-Based Score Estimation

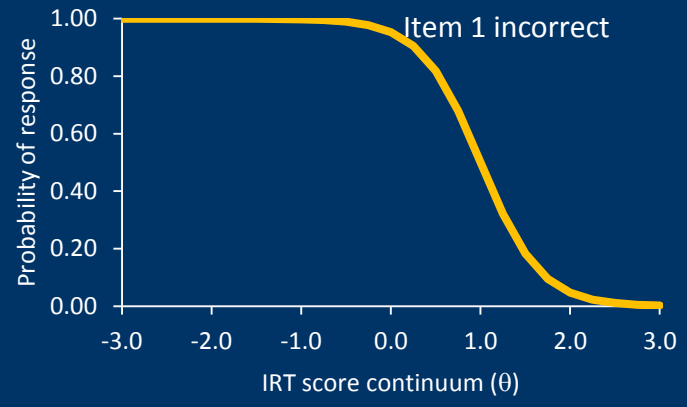
- **Essentially aggregate all of the item responses to form a posterior distribution and calculate score from that distribution**
- **Maximum likelihood (ML)**
  - **The mode of the likelihood function for the item responses**
- **Maximum *a posteriori* (MAP)**
  - **Incorporates a prior (typically normal) distribution and finds mode of posterior**
- **Expected *a posteriori* (EAP)**
  - **Incorporates (normal) prior but estimates the mean of the posterior**

Pattern Scoring and Summed Scoring

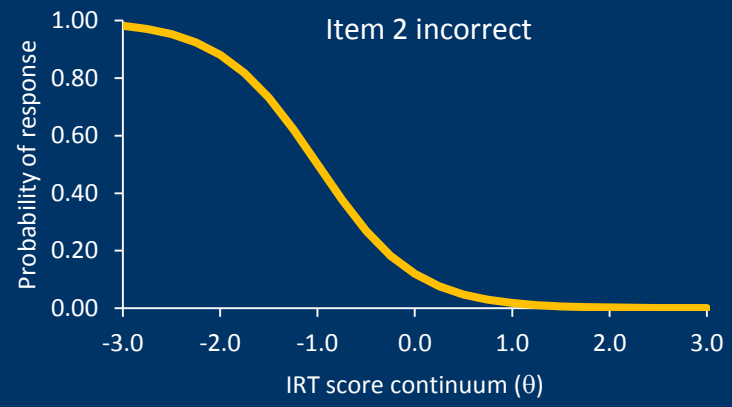
- **Each response pattern has a distinct score even for summed scores that are identical**
- **Scoring based on the full IRT model uses the pattern of responses to obtain the most precise trait score estimate**
- **For practical purposes it is often preferable to provide a single trait score estimate for each summed score rather than for each pattern score**
  - **The principles and estimation are the same, but the collection of item responses used to form the posterior distribution changes**

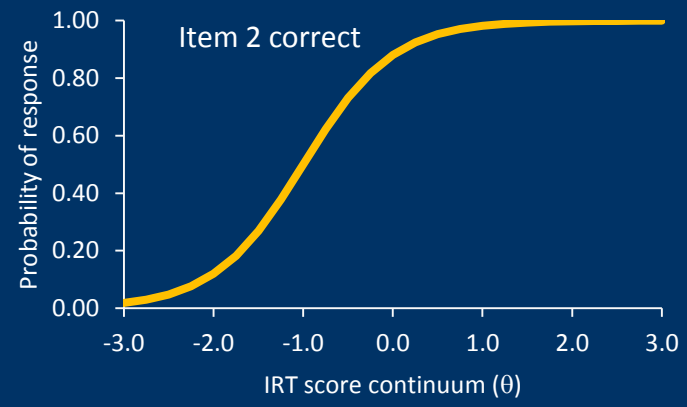
a = 3
b = 1

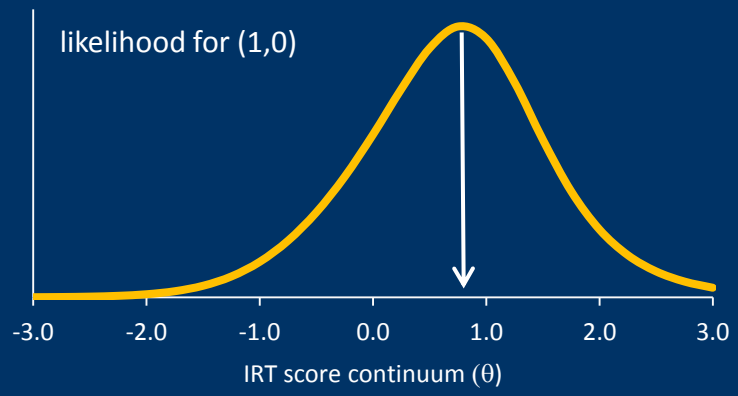


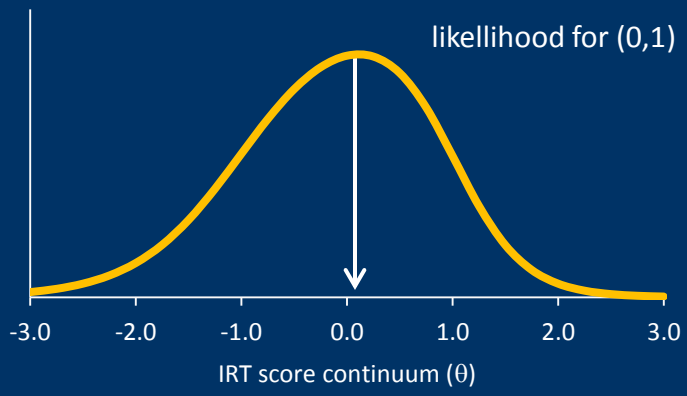


a = 2
b = -1



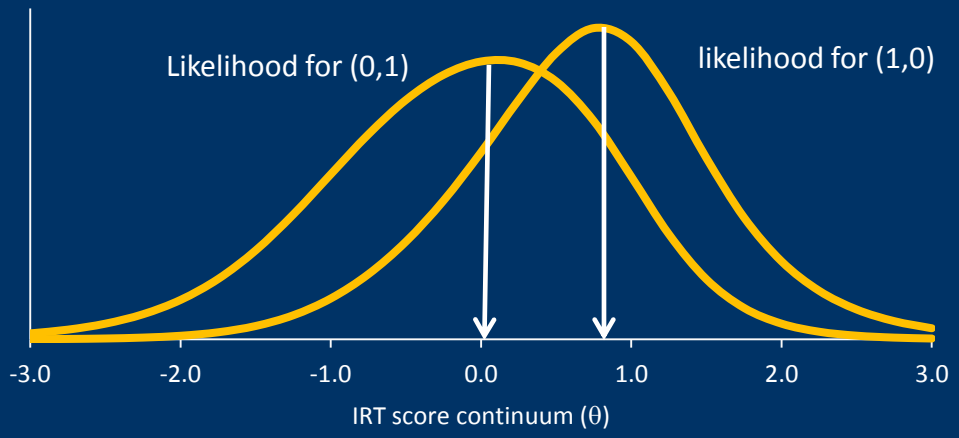


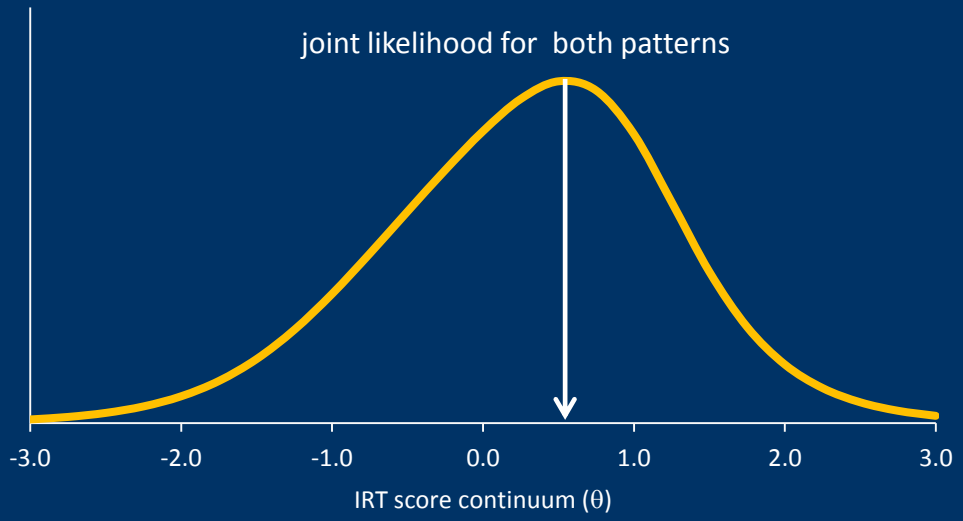




RAND

Pattern vs. Summed Scoring





Differential Item Functioning (DIF)

- **DIF indicates a difference in the measurement properties of an item that is associated with membership in a particular group**
  - **The item response (and expected score) function is different for the two groups**
  - **The model-based item score is different for members of the two groups even if they have the same level of the underlying trait**
- **For example, depression items assessing frequency of crying typically exhibit gender DIF**
  - **Men endorse this symptom less frequently than women – this difference is not related to depression but to the group membership**

Differential Item Functioning (DIF)

- **Strong DIF can impact inferences about group mean differences.**
- **Scale and item bank development efforts typically include checks for DIF according to basic demographic groups (e.g., gender, age, ethnicity, education)**
- **DIF approach can also be used to evaluate**
  - **quality and comparability of scale translations**
  - **Effect of administration mode on item properties**

DIF Detection

- **DIF is present if the item parameters for the two groups are significantly different from one another given the overall group mean difference**
- **Typically evaluated in IRT context via likelihood ratio or WALD χ^2 tests**
- **It is challenging to get an unbiased estimate of the group mean difference in the presence of DIF items**
  - **All – other likelihood ratio χ^2 approach**
  - **Two-step WALD approach**
  - **Purified anchor approach as second step to either of these initial 'sweeps'**

DIF Evaluation

- **IRT DIF tests tend to be powerful**
- **χ^2 significance is influenced by sample size**
- **Helpful to adopt some p-value correction method to adjust for multiple tests (Bonferroni, Benjamini-Hochberg)**
- **Examination of item characteristic or expected score curves for the two groups is also helpful**
  - **a visual representation of the identified DIF**
  - **Can sometimes help discern whether significant DIF is 'ignorable'**

Example DIF Item from Assessment of Affective/Hedonic Benefits of Smoking

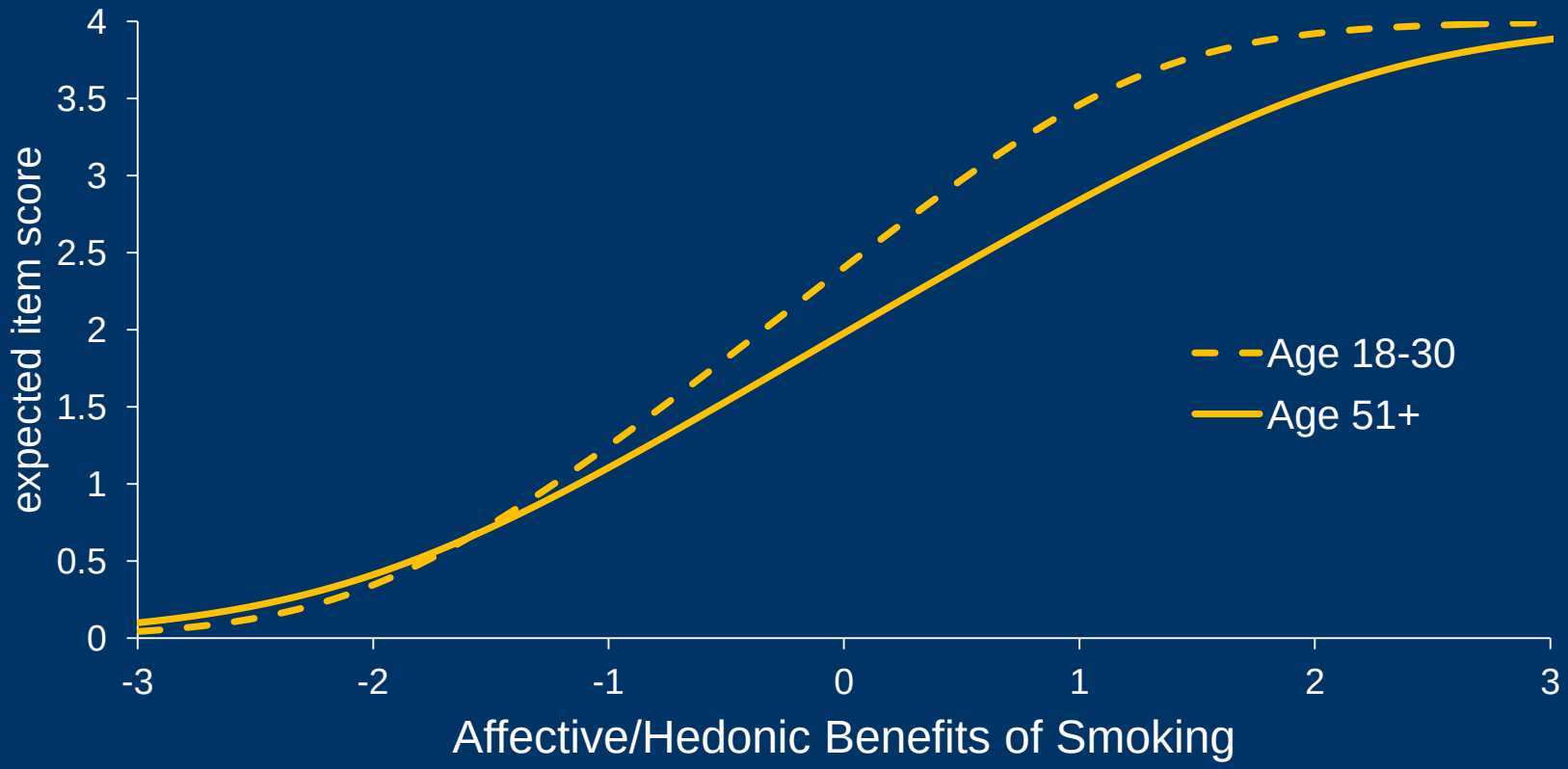
I enjoy the sensations of a long slow exhalation of smoke

	a	b_1	b_2	b_3	b_4
Age 18-30	2.36	-1.59	-.73	.08	.9
Age 51+	1.67	-1.52	-.54	.58	1.6

$\chi^2_{(5)} = 23.9 \qquad p < .0003$

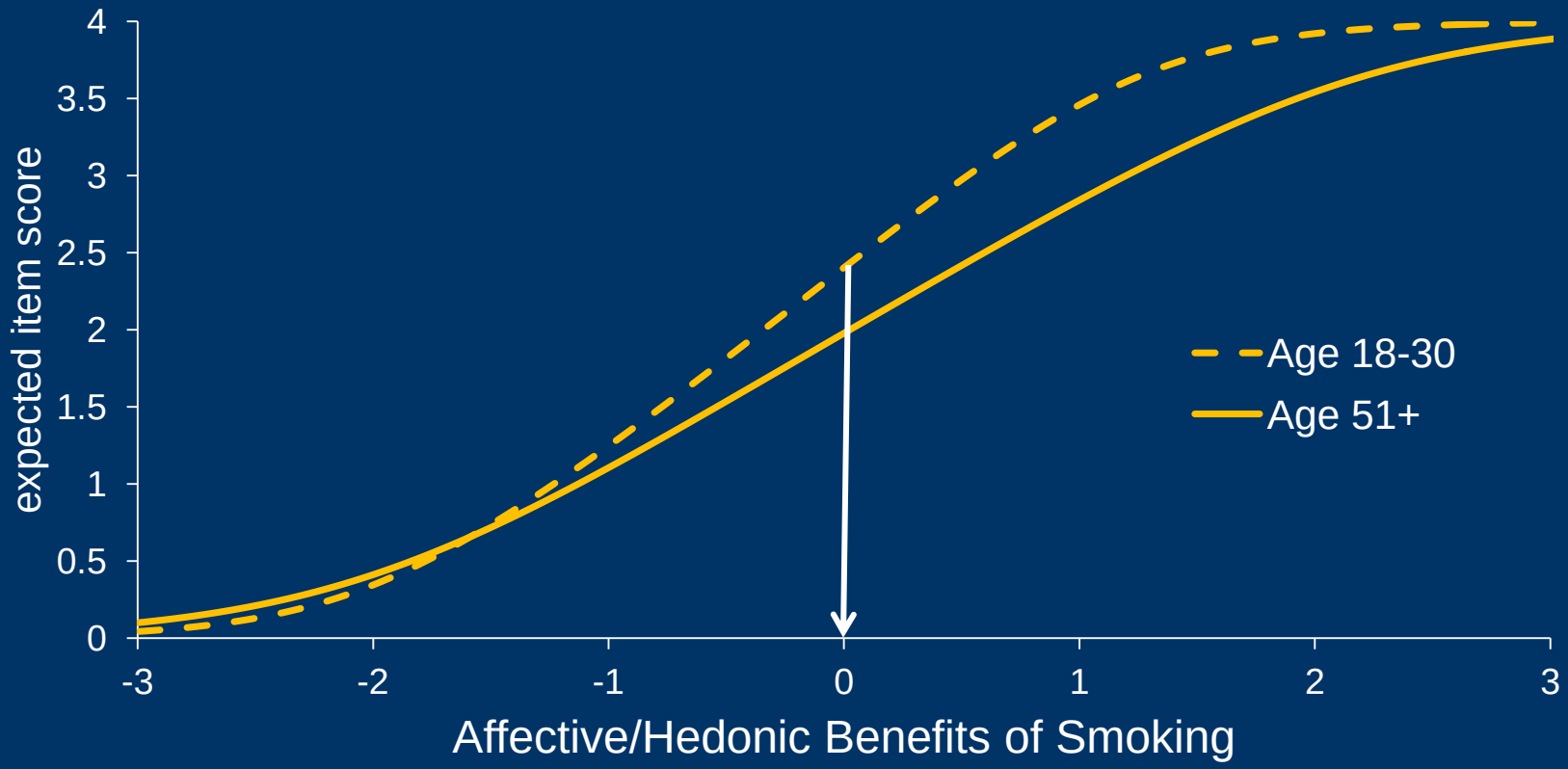
Example DIF Item from smoking inventory

I enjoy the sensations of a long slow exhalation of smoke



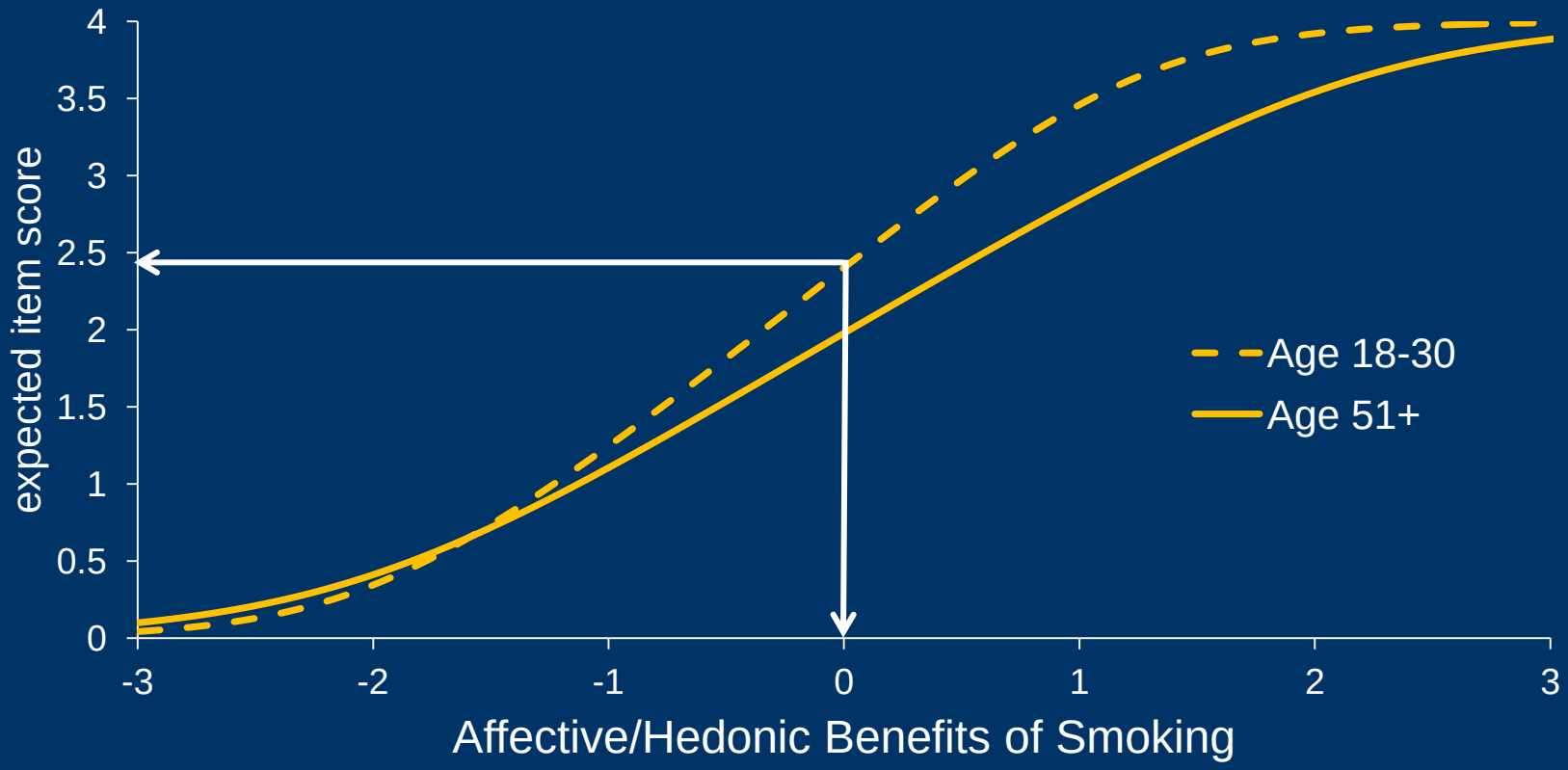
Example DIF Item from smoking inventory

I enjoy the sensations of a long slow exhalation of smoke



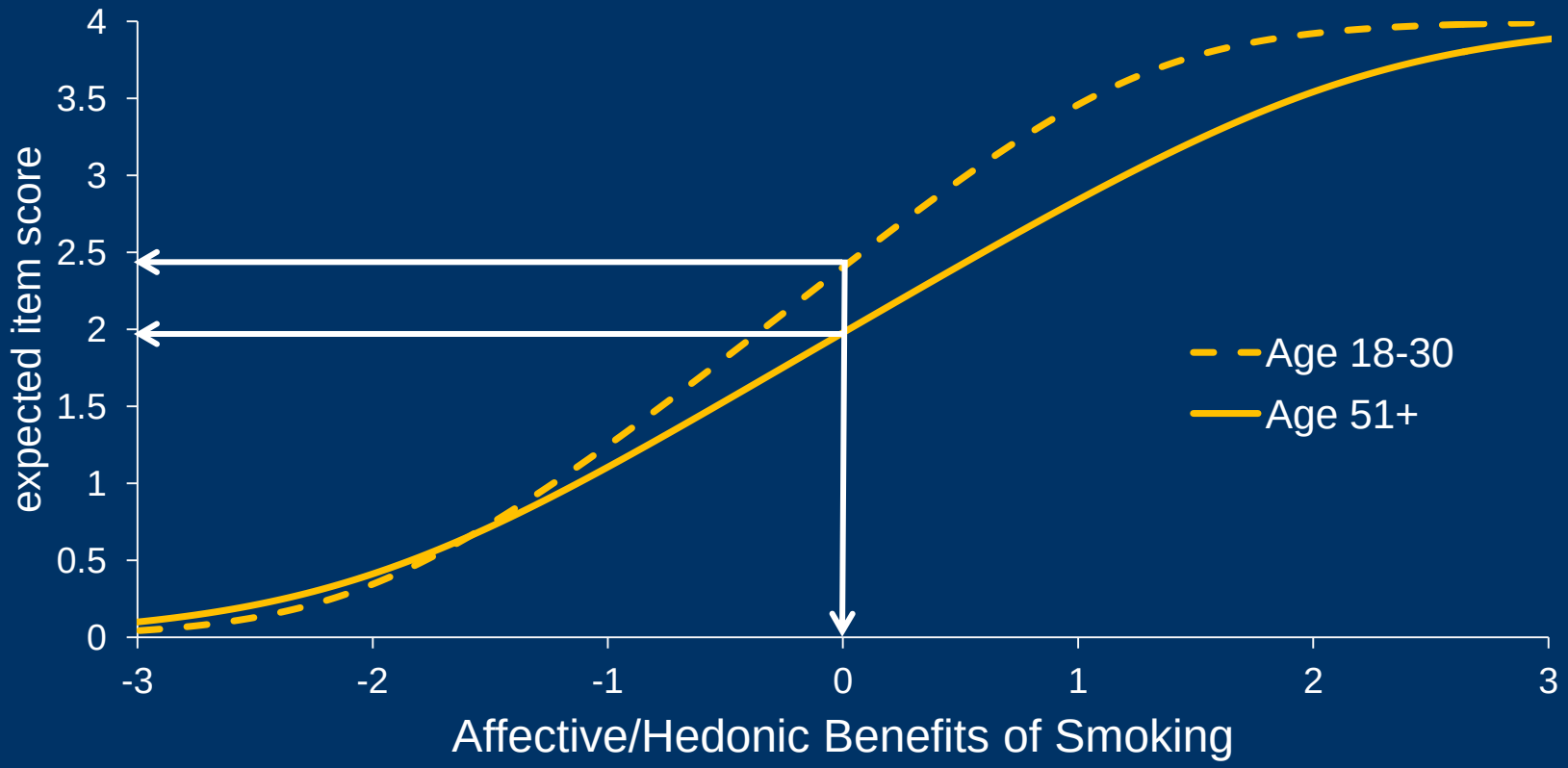
Example DIF Item from smoking inventory

I enjoy the sensations of a long slow exhalation of smoke



Example DIF Item from smoking inventory

I enjoy the sensations of a long slow exhalation of smoke



Thank you

QUESTIONS?