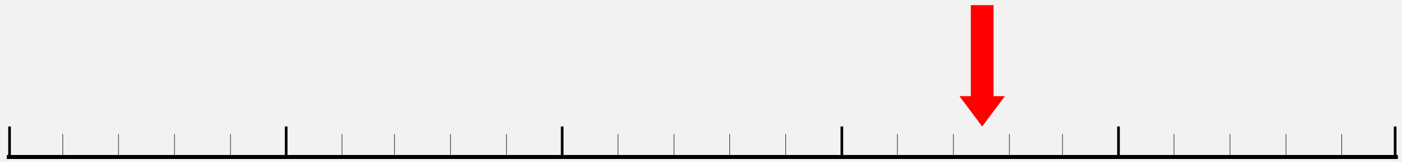
The dark arts – how to measure things we cannot see

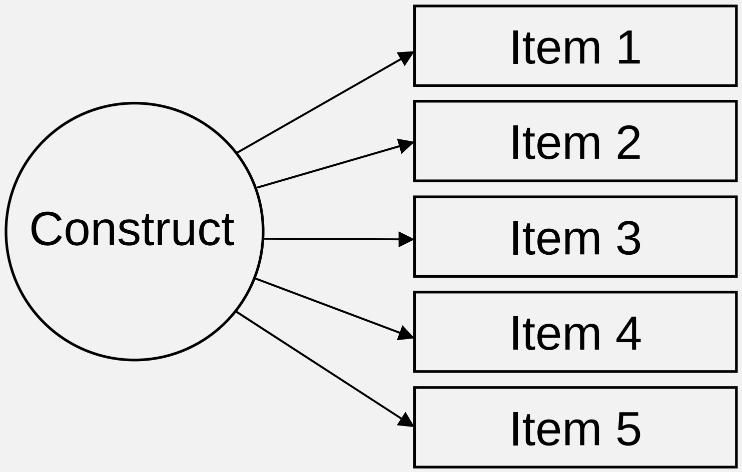
Part 2 – Current approaches to psychometric measurement

Andrew Mackinnon
Centre for Youth Mental Health
University of Melbourne

Fundamental concept



To locate individuals on a continuum/line representing a construct of interest



Responses

Correct/Incorrect

Endorsed/Not endorsed

‘Classical’ test theory

- **The classical test theory model:**

  Observed Score (O) = True *Score* (T) + *e*

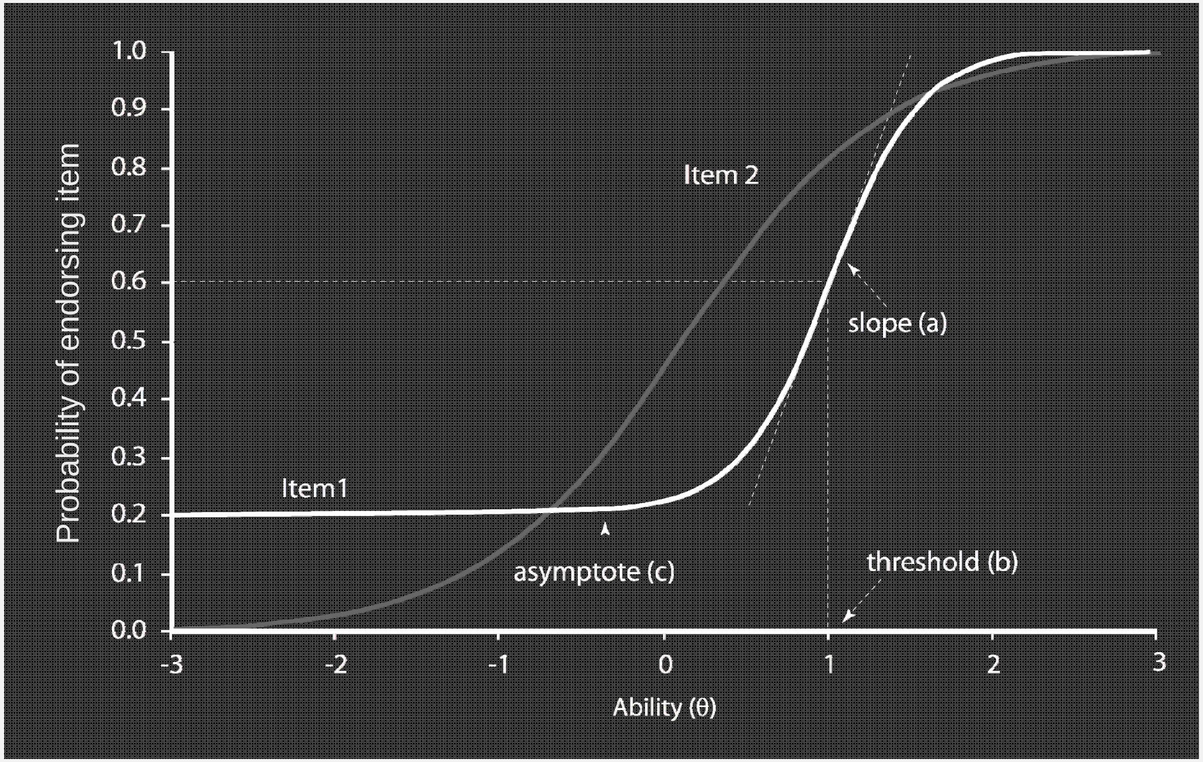
  Reliability $\rho_{OT}^{2} = \frac{\sigma_{T}^{2}}{\sigma_{O}^{2}}$

  A model, but not much of one!

- **‘Modern’ psychometric methods propose models seeking to predict the responses to items based on**
  - **characteristics of the item**
  - **the location of the respondent on the dimension**

(Parametric) Item Response Theory

Binary response model



$$P(x_i = 1 | \theta_i) = c_j + (1 - c_j)\frac{1}{(1 + e^{-Da_j(\theta_i - b_j)})}$$

IRT parameters

Parameter	Role
x_i	Observed response of person *j* (0=not endorsed/incorrect; 1=endorsed/correct)
θ_i	Location of person *j* on dimension (ability/severity) Scale of θ is arbitrary, distribution not assumed normal
a_j	Slope (discrimination) ($r_{item\text{-}dimension} = a/\text{sqrt}(1+a$^2) a=1 → r ~ .7)
b_j	Threshold (location)
c_j	Asymptote (pseudo-guessing parameter)
D	Scaling parameter. Equates logistic function to Normal ogive (D=1.7)

Family of IRT models

Parameter	Model
3 PL	$P(x_i = 1 \mid \theta_i) = c_j + (1 - c_j)\dfrac{1}{(1 + e^{-a_j(\theta_i - b_j)})}$ Location, discrimination and asymptote parameters for each item
2 PL	$P(x_i = 1 \mid \theta_i) = \dfrac{1}{(1 + e^{-a_j(\theta_i - b_j)})}$ Location and discrimination parameters for each item
1 PL	$P(x_i = 1 \mid \theta_i) = \dfrac{1}{(1 + e^{-a(\theta_i - b_j)})}$ Location parameter only (common discrimination)

For each logistic model, there is a corresponding Normal ogive model

All models assume unidimensionality – this must be established independently

Other IRT models

Model	**Details**
Guttman Scale	Item characteristic curve is a step function (slope is infinite)
Non-parametric IRT (e.g., Mokken scaling)	Form of item characteristic curve is not assumed logistic/normal but is inferred from data
Rasch Model	More later …

IRT example

Spot-the-Word Test

- **60 word - pseudo word pairs**
- **measure of IQ/ability**
- **resistant to subsequent change**

plargen – savage

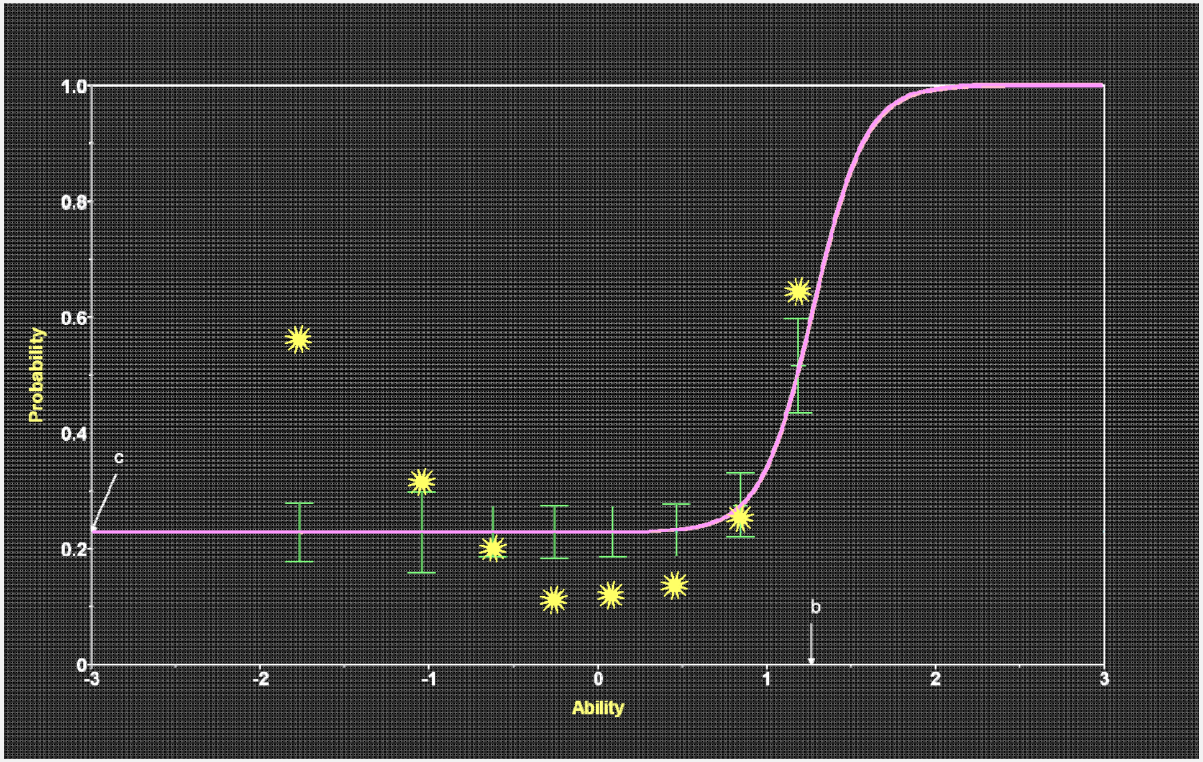
loxeme – legerdemain

threnody – epigrot

IRT parameters and fit statistics

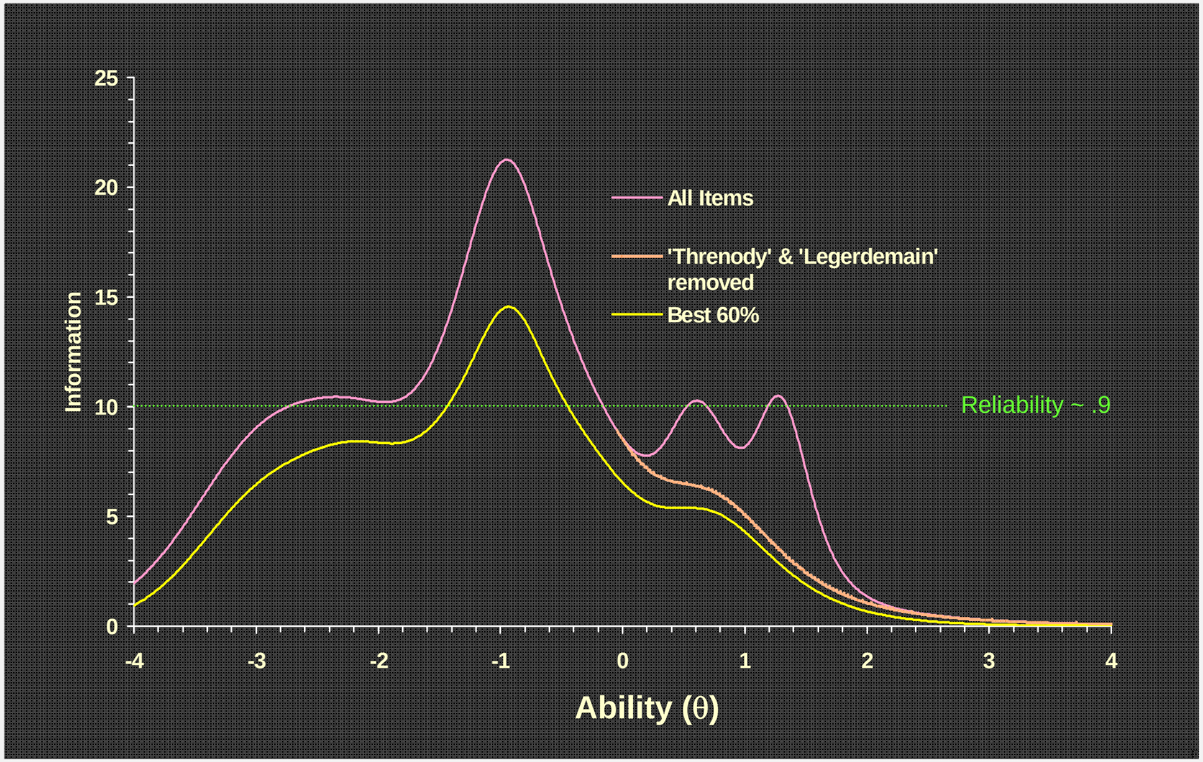
Word Pair	IRT parameters			Item fit
	Slope	Threshold	Asymptote	χ^2 / d*f* / P
plargen – savage	0.82	-3.50	0.32	3.4 / 5 / 0.64
trelding – rafters	1.04	-3.21	0.36	4.4 / 3 / 0.23
hilfren – domain	1.23	-3.05	0.32	5.6 / 2 / 0.06
broxic – oasis	1.20	-2.86	0.28	4.3 / 4 / 0.36
gibbon – wonnage	0.94	-2.42	0.38	2.8 / 5 / 0.73
pimple – brizzler	1.57	-2.34	0.44	2.7 / 3 / 0.45
livid – trasket	1.59	-1.96	0.41	1.7 / 4 / 0.79
venady – monad	0.46	-1.79	0.30	7.4 / 9 / 0.59
necromancy – ghoumic	0.96	-1.47	0.50	45.3 / 6 / 0.00
hipple – osprey	1.63	-1.16	0.35	6.6 / 6 / 0.36
brastome – banshee	2.35	-1.11	0.29	1.4 / 5 / 0.92
archipelago – zampium	2.67	-1.06	0.50	14.5 / 4 / 0.01
clavanome – bestiary	0.59	0.06	0.50	19.1 / 9 / 0.02
canticle – grammule	1.24	0.31	0.50	10.2 / 9 / 0.33
viridian – psynoptic	0.87	0.32	0.29	10.4 / 9 / 0.32
loxeme – legerdemain	3.32	0.55	0.37	52.6 / 7 / 0.00
narwhal – epilair	1.69	0.66	0.17	38.0 / 8 / 0.00
hoyden – clinotide	1.93	0.67	0.33	33.1 / 8 / 0.00
pinnace – strummage	0.45	0.69	0.50	59.6 / 9 / 0.00
shako – strubbage	1.30	0.88	0.34	7.7 / 9 / 0.56
threnody – epigrot	3.82	1.27	0.23	242.2 / 8 / 0.00
bellissary – cyan	0.69	1.46	0.36	6.1 / 9 / 0.73

Item fit – threnody–epigrot



a=3.81, b=1.26, c=0.23; *but* χ^2=242.2,df=8, p=0.00

Precision of measurement



$$se(\theta) = 1/\sqrt{I(\theta)}$$

$$I_{Test}(\theta) = \sum I_i(\theta)$$

Item Response Theory

≡

Factor Analysis

- **IRT models can be shown to be equivalent to (nonlinear/generalized) single factor models**
  (**Does not apply easily to models with ‘guessing’)**
- **FA parameters may be converted IRT parameters and *vice versa***
  - loadings → slopes/discrimination parameters
  - item thresholds → dimension thresholds
- **CFA in SEM packages is often easier and more flexible than IRT-specific routines**

A rant about Rasch

3 par IRT model —
$$P(x_i = 1 | \theta_i) = c_j + (1 - c_j)\frac{1}{(1 + e^{-Da_j(\theta_i - b_j)})}$$

Rasch model —
$$P(x_i = 1 | \theta_i) = \frac{1}{(1 + e^{-(\theta_i - b_j)})}$$

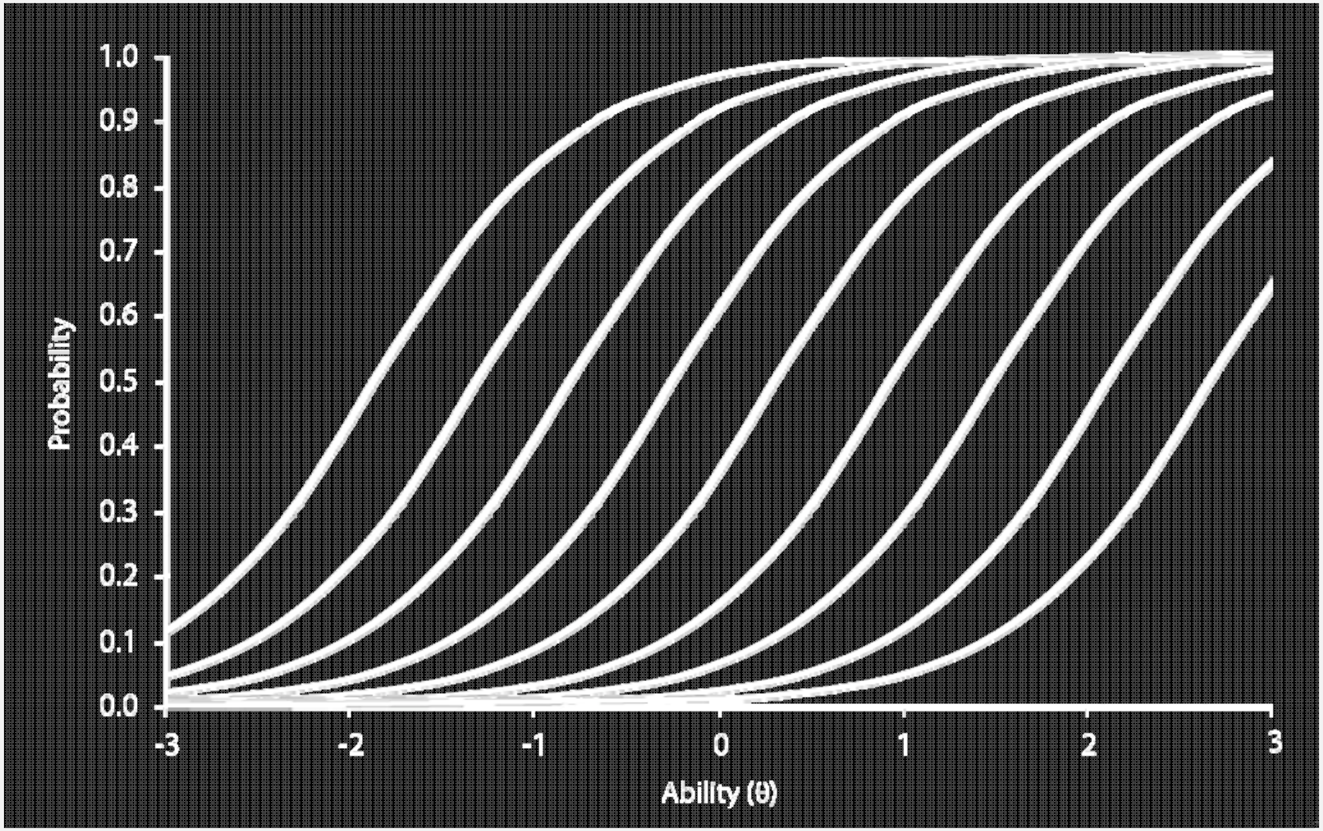
Pragmatically, Rasch models are one parameter IRT models (with the common discrimination parameter factored out)

or

single-factor factor analyses with all loadings constrained to be equal, and the value of the loadings ignored

Rasch measurement has a philosophical basis which aims to develop instruments that ensure ‘objective measurement’ of psychological constructs

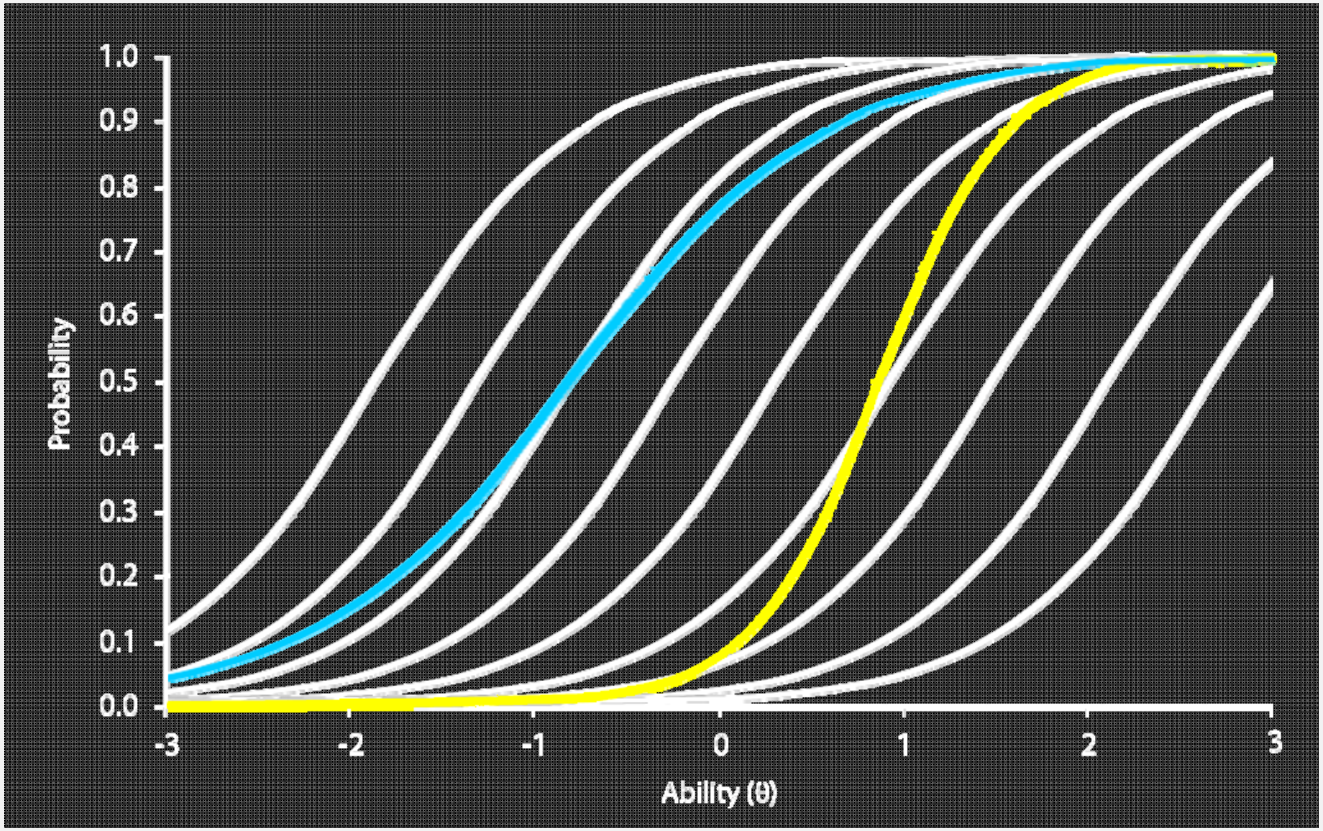
Rasch - invariance



Relative endorsement patterns of individuals is invariant regardless of which items are chosen.

Relative behaviour of items is invariant regardless of the respondents chosen.

Rasch - invariance



Relative endorsement patterns of individuals is invariant regardless of which items are chosen.

Relative behaviour of items is invariant regardless of the respondents chosen.

Rasch – the good

- **Rasch scales have lovely measurement properties**
- **Rasch models have lovely statistical properties**
- **Models are economical in terms of parameter estimation**
- **Can be used successfully with modest sized samples**
- **Basic models can be fitted with standard software**
- **In previously investigated tests with relatively homogeneous loadings/slopes, Rasch models can focus item selection on item location or severity**

Rasch – the less good

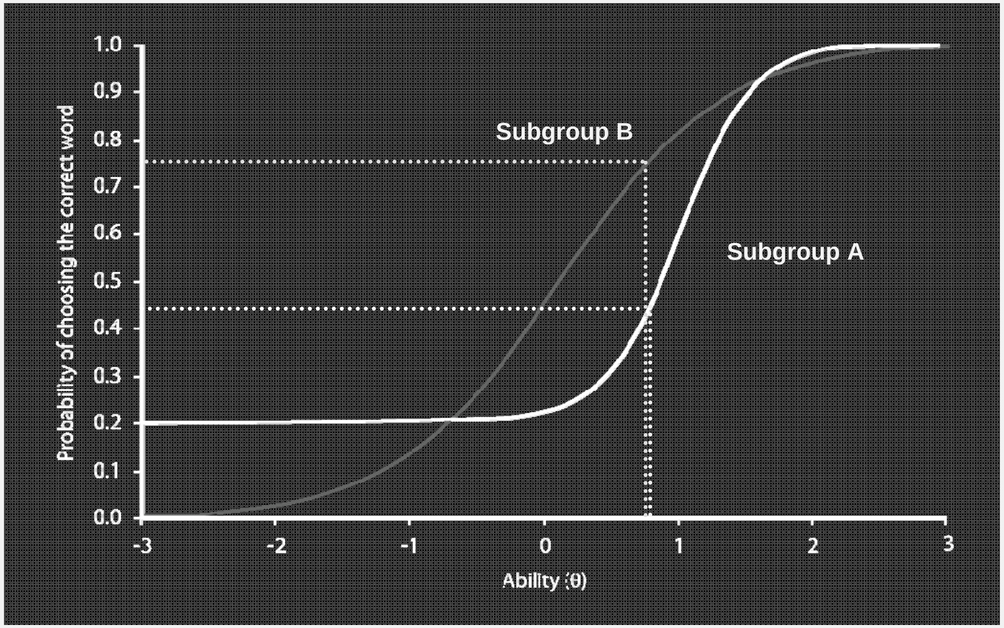
- **Location alone is an incomplete investigation of measurement properties of a set of items**
  - **particularly in the early stages of development**
- **Fit indices may flag items for removal that discriminate the dimension better than average**
- **Removing or retaining items on model/statistical grounds alone may alter the construct being measured**
- **All claims about measurement properties are 'internal' – Rasch scales do not yield more substantively objective measures of psychological constructs**

Comparability of responses/tests

- Scales are used to compare groups and to compare the same group over time
- This implies –
  - scores in different groups must ‘mean’ the same thing – groups must use or respond to the test in the same way
  - scores taken from an individual on different occasion must ‘mean’ the same thing – individuals must use or respond to the test in the same way over time
- This is ~~a~~ the big measurement issue for any psychometrically-based research
  - comparing ‘culturally different’ groups
  - longitudinal studies, particularly of groups at substantial life stage/ developmental change

‘Bias’ in IRT terms

- Responses to items should reflect only the location of the individual on the dimension (θ)
- Where differences in the probability of item endorsement depend on respondent characteristics other the θ, the item is said to be biased exhibit differential item functioning (DIF)



Investigating DIF

- **What you cannot do:**
  - **examine item means/endorsement rates - confounds differences in item (*a*,*b*,*c*) and group (θ) characteristics**
- **IRT – comparison of item slope and threshold parameters can reveal the nature and extent of item bias. Threshold change is particularly important.**
- **Confirmatory factor analysis – examination of factorial invariance is equivalent to testing DIF (and more)**
- **'Observed score' methods – based on Mantel-Haenszel statistic and generalizations**
- **DIF is *differential* – if all items change, DIF will not be detected – the problem is differential test functioning (DTF) or bias.**

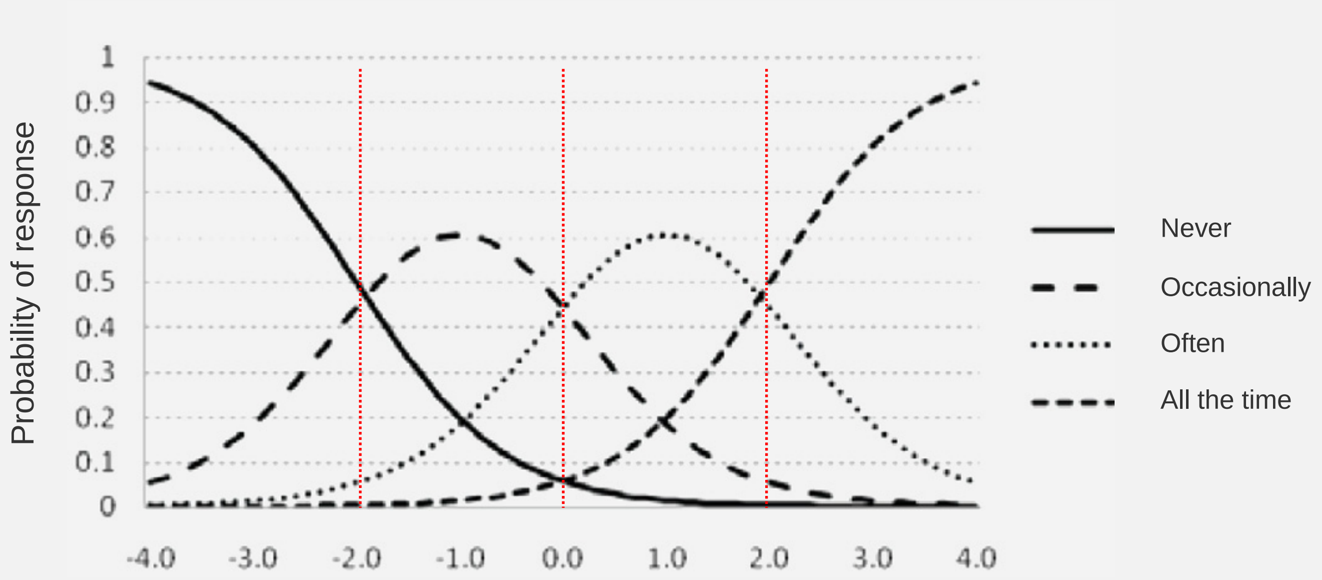
Responses to DIF and DTF

- Limited DIF
  - remove affected items
  - evaluate the impact on scores of retaining items
- For extensive DIF or DTF
  - use equating methods (if construct is known to be the same across groups or ages)
  - consider multidimensional approach – subscale of common, comparable items, subscale of ‘differential’ items
  - abandon hope!
    - the construct may not be comparable or not exist in different groups or at different ages
    - the ‘conduit’ for measurement – words and participant response – may be inadequate to capture the same underlying construct across groups or over time

Poly(cho)tomous responses

Getting up and going to school is a big hassle for me…

Never | Occasionally | Often | All the time



$a=1.4$; $b_{nev/occ}=-2.0$; $b_{occ/oft}=0.0$; $b_{oft/all}=2.0$

Common polytomous IRT models

Model	**Details**
Nominal Response Model	Accommodates unordered multiple response alternatives
Graded Response Model	**Response categories are assumed/forced to be ordered. Discrimination and thresholds may vary between items. Poly equivalent to 2 par model**
Partial Credit Model	The Rasch ‘equivalent’ to the Graded Response Model (but parameterization is different)
Rating Scale Model	Partial Credit Model constrained so that thresholds for response alternatives are the same for all items.

- Models with asymptotes are rare for polychotomous responses
- Many variations are possible by adding constraints etc.

Graded Response Model

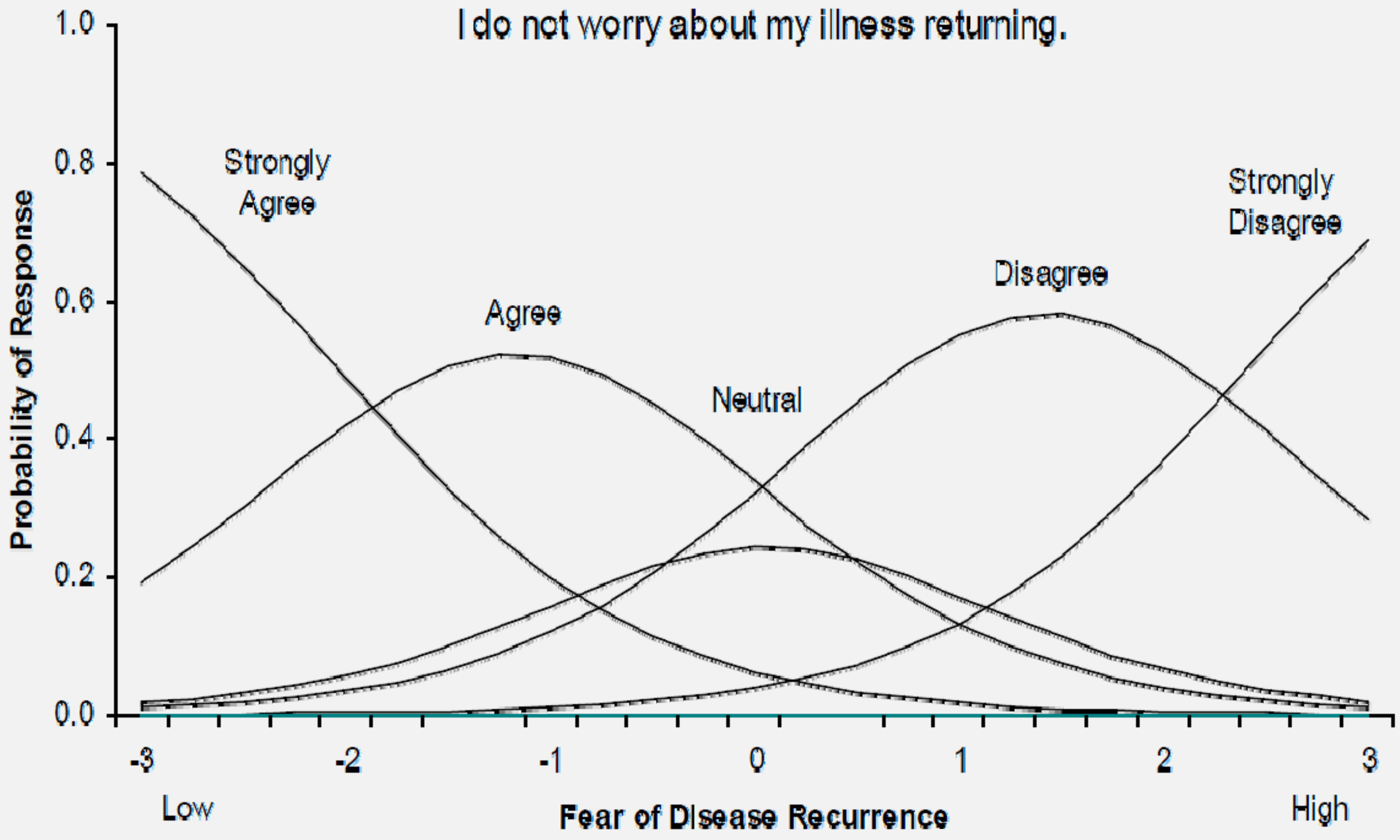
- Essentially same form as 2 parameter model
- Based on a series of binary models
- For R response categories *r*=1,2,3… there are R-1 thresholds
- Models the probability of any given response category or higher
  - for P_{i1} – b_{1i} ~ 1 *vs* 2,3,4; P_{i2} – b_{2i} ~ 1,2 *vs* 3,4; P_{3i} – b_{3i} ~ 1, 2,3 *vs* 4;

$$P(X_{ij} \geq r_j \mid \theta_i) = \frac{1}{(1+e^{-a_j(\theta_i - b_{rj})})}$$

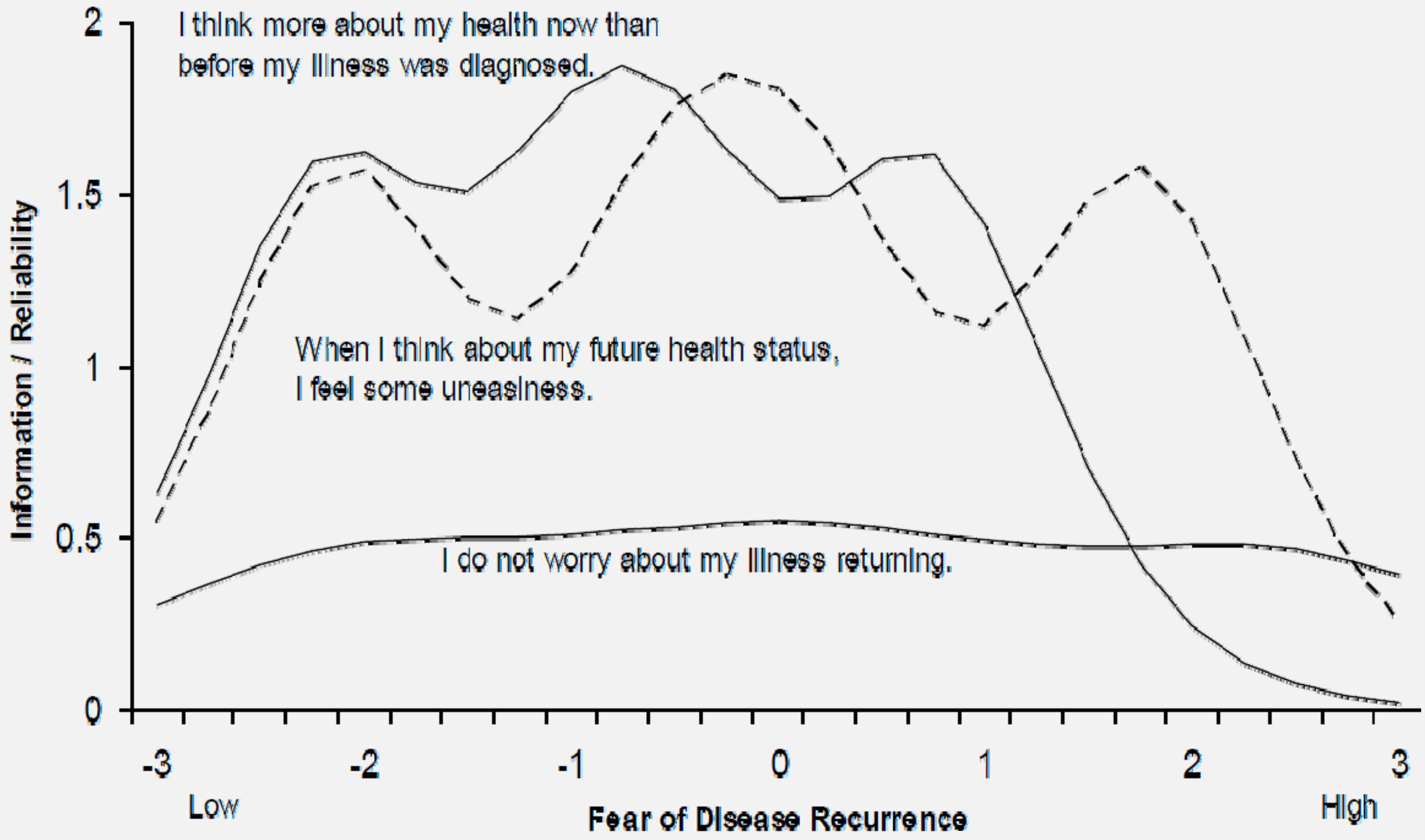
$$P(X_{ij} \geq 0_j \mid \theta_i) = 1 \qquad P(X_{ij} \geq r_j \mid \theta_i) = 0$$

  - common discrimination parameter (a_i) for item
  - calculate probability of particular response by subtraction:
    - Prob of response 1 → 1 – P_{i1} Prob of response 2 → P_{i1} – P_{i2}
    - Prob of response 3 → P_{i2} – P_{i3} Prob of response 4 → P_{i3} – 0

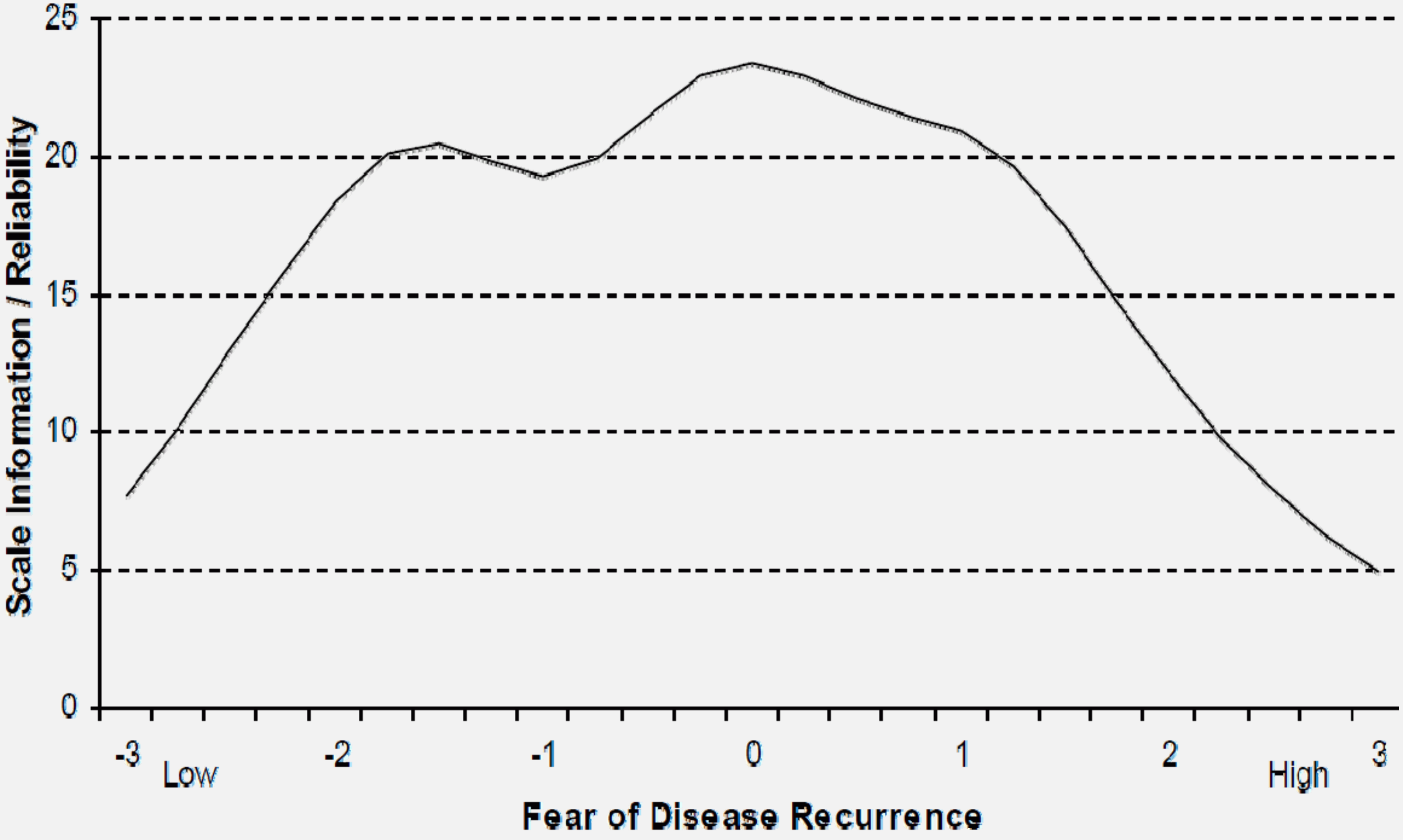
Evaluating polytomous items



Evaluating polytomous items



Evaluating polytomous scales



Estimating respondent location (θ)

- Thetas (θ_j) are parameters to be estimated in the model just like other IRT parameters
- For Rasch models, number correct/number of thresholds passed is a sufficient statistic for θ
- Correlation between and simple sum of items is often very high (>>.9), except at ends of scale
- When item parameters have been estimated in large samples (calibration) they may be considered fixed and used to estimate θ for subsequent respondents
- θ may be estimated from different subsets of items
  - respondents need not answer the same items

CAT – Computerized Adaptive Testing

- **Computerized Adaptive Testing -**
  - **chooses subsequent items for presentation dependent on respondent's previous responses**
  - **items are selected from a bank of calibrated items**
  - **presentation stops when adequate precision of location has been determined**
  - **requires presenting fewer items but maintains/improves individual precision of measurement**

PROMIS

Patient Reported Outcomes Measurement Information System

- **NIH supported, IRT-based measurement system**
- **Instruments currently cover:**
  - **anger, anxiety, depression, alcohol, pain, fatigue, physical functioning, sleep, sexual function and satisfaction, social participation, social support, global health**
- **Online CAT testing (Assessment Center)**
- **Paper forms and short-forms**
- **Development ongoing**

http://www.nihpromis.org

Is it worth it?/Does it matter

NO

- Conventional methods often yield very similar results when …
  - scale/items have been developed using them
  - number of points on scale is largish and response distributions are ‘humped’
  - when interest lies in the ‘middle’ range

YES

- IRT may lead to better outcomes when …
  - development of a scale is in early phases
  - interest lies at extremes or specific locations (screening)
  - comparability between groups/over time is critical
  - measurement matters

Software

- Specialized Software
  - BILOG-MG, MULTILOG, PARSCALE (1, 2, 3par IRT binary, ordinal)
  - WINSTEPS, RUMM (Rasch Unidimensional Measurement Models) (Rasch modelling, binary, ordinal)
  - Mplus (CFA/EFA with special IRT features)
  - see list at http://www.rasch.org/software.htm

- SPSS – Nothing much!
  - Standard exploratory factor analysis
  - RELIABILITY procedure

Software

- **Stata**
  - **alpha (Cronbach's alpha)**
  - **OpenIRT (1, 2, 3 par IRT, binary only)**
  - **raschtest (Rasch modelling, binary only)**
  - **generalized SEM (from v.13)**
- **R**
  - **ltm (1, 2, 3 par IRT, binary, ordinal)**
  - **eRm (extended Rasch modelling)**
  - **MiscPsycho (Classical and Rasch models)**

References

van der Linden, W.J. & Hambleton, R.K. (1997) Handbook of Modern Item Response Theory. New York: Springer. [Particularly good section on polychotomous response models.]

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum: Hillsdale, N.J. [A noted text in this field – has a succinct but comprehensive chapter on classical test theory.]

Raykov, T. and Marcoulides, G.A. (2011) *Introduction to Psychometric Theory*. New York, NY: Routledge. [Recent and pretty rigorous text.]

Revelle, W. *eBook on Psychometric Theory* http://personality-project.org/r/book [Online text oriented towards using R for psychometric research.]

Wright, B.D. & Masters, G. (1982) *Rating Scale Analysis*. Mesa Press, Chicago. [Rasch models for ordered scales.]

Maydeu-Olivares, A., Drasgow, F. & Mead. A.D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement* 18(3): 245-256. [A good ‘directory’ of the different types of models for ordered response scales.]