AARHUS UNIVERSITY

Introduction to IRT

Advanced course in questionnaire technique and clinimetrics

Eva Ørnbøl, Statistician

1

AARHUS UNIVERSITY

Outline

- What we have used so far – CTT
- Introduction and origin of IRT
- From the Guttman scale to IRT
- Assumptions of IRT
- IRT models and graphs
  - Item characteristic curves: 1PL, 2PL, 3PL/4PL and non-parametric IRT
  - Items with more than two categories: category probability curves, GRM vs. PCM/RSM
- Test characteristic curve, item and test information
- Checking the assumptions
  - Unidimensionality and monotonicity
  - Differential item functioning (DIF)
  - Local independence

2

AARHUS UNIVERSITY

What we have used so far - CTT

- Most of the statistical methods have belonged to the framework of Classical Test Theory (CTT).
  - Builds on the true score model

$$X = T + E, \quad \rho_{TE} = 0, \quad \bar{E} = 0, \quad \rho(E_1, E_2) = 0$$

  - Correlations, Cronbach's alpha, factor analysis, intraclass correlation (ICC), LoA etc.
  - Items are summed to produce a total score being the focus of attention.
  - Bias towards average difficulties in test construction.
  - One common standard error of measurement (SEM) for all score levels.
  - Item and person parameters are sample dependent.
    - Item difficulty (% positive answers) and item–total correlations depend on the trait level and spread of the sample.
    - A person's total score depends on which items are included.
- A different framework is Item Response Theory (IRT) including Rasch Model Theory (RMT).

3

AARHUS UNIVERSITY

Intro to IRT

- A framework of models linking a person's level of an unobservable latent variable (θ) to the probability of a given response to each item.
- Each item is described by its own parameters, e.g. difficulty (location) and discrimination.
- IRT assumes items to be reflective – the latent variable causes the item responses.

4

AARHUS UNIVERSITY

Origin of IRT

- IRT stems from educational and psychological testing (Lord, 1952; Rasch, 1960; Birnbaum, 1968).
- Aim: to estimate a person's location (θ) on the latent continuum from the person's responses to a set of items.
- The initial focus was on items with two response options (correct/incorrect, present/not present) – later extended to items with ordered response categories (e.g. Likert scales).
- Items are placed on the same latent continuum through their parameters (location/difficulty) – this introduces a *ranking* of items, and persons and items can be compared directly.

5

AARHUS UNIVERSITY

Guttman scale

- The Guttman scale (Guttman, 1944) is an important precursor of IRT.
- Items are ranked by difficulty: a person with a positive response to one item also has positive responses to all easier items.
- The Guttman scale is deterministic: the probability of a positive response jumps from 0 to 1 at the item's location.

6

AARHUS UNIVERSITY

Guttman scale - example

	Patients						
Walking skill (easiest → hardest)	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Stand up	1	1	1	1	1	1	0
Walk indoor with help	1	1	1	1	1	0	0
Walk indoor without help	1	1	1	1	0	0	0
Walk outdoor 5 min	1	1	1	0	0	0	0
Walk outdoor 20 min	1	1	0	0	0	0	0
Run 5 min	1	0	0	0	0	0	0
Total score	**6**	**5**	**4**	**3**	**2**	**1**	**0**

7

AARHUS UNIVERSITY

Guttman error - example

	Patients						
Walking skill (easiest → hardest)	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Stand up	1	1	1	1	1	1	0
Walk indoor with help	1	1	1	1	1	0	0
Walk indoor without help	1	1	1	**0**	0	0	0
Walk outdoor 5 min	1	1	1	**1**	0	0	0
Walk outdoor 20 min	1	1	0	0	0	0	0
Run 5 min	1	0	0	0	0	0	0
Total score	**6**	**5**	**4**	**3**	**2**	**1**	**0**

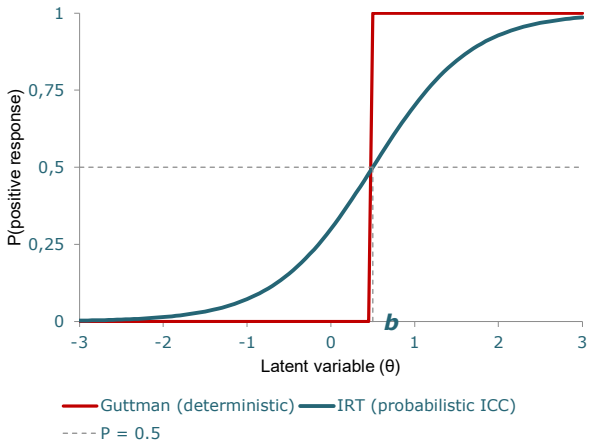
Patient D has a Guttman error (red) – but the same total score (3).
If you allow a few errors, you talk about a probabilistic Guttman scale.

8

AARHUS UNIVERSITY

From Guttman to IRT

- Guttman: the probability of a positive response jumps from 0 to 1 at the item's location – a Guttman error is impossible.
- IRT: the probability increases smoothly with θ – the item characteristic curve (ICC).
- An "error" is then simply a less likely response pattern.
- The item location (b) is the value of θ where the probability is 0.5.



9

AARHUS UNIVERSITY

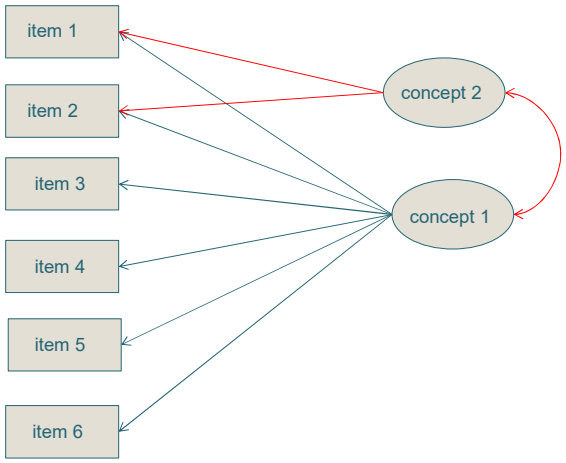
Assumptions of IRT

- **Unidimensionality:** one latent variable (θ) explains the item responses.
- **Local independence:** given θ, the responses to the items are independent.
- **Monotonicity:** the probability of a positive response increases with θ (cf. the ICC on the previous slide).
- **No differential item functioning (DIF):** the ICC is the same in all subgroups (e.g. sex, age) – measurement invariance.

10

AARHUS UNIVERSITY

Violation of unidimensionality

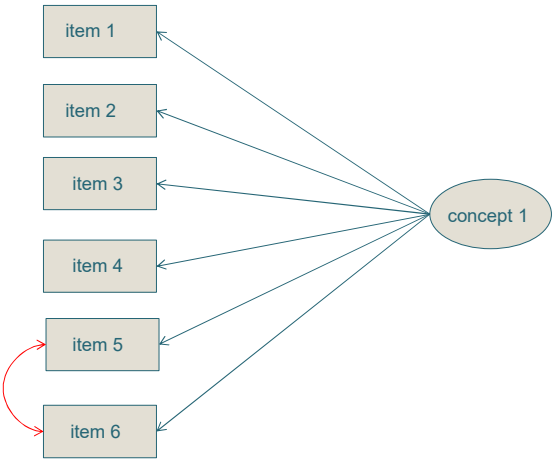


Violation: a different latent variable influences the responses of a subset of items.

11

AARHUS UNIVERSITY

Violation of local independence



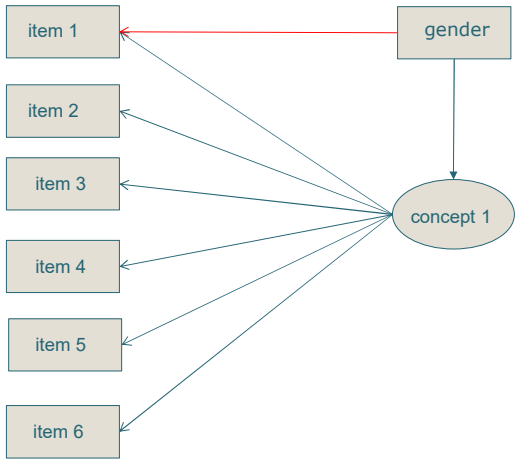
Violation: the latent variable does not explain all correlation between pairs of items.

Link to unidimensionality: if θ explains all correlations between the items, the items are uncorrelated when θ is held fixed – they are locally independent. A remaining correlation (here items 5 and 6) means that the items share something other than θ.

12

AARHUS UNIVERSITY

Violation of measurement invariance (DIF)



Violation (DIF): gender affects the response to item 1 directly – not only through the latent variable. A group difference in the latent variable itself (gender → concept 1) is allowed.

13

AARHUS UNIVERSITY

IRT - graphs

- IRT results are mainly presented and interpreted through graphs, e.g.:
  - ICC – item characteristic curve (also item response function, IRF)
  - CPC – category probability curve (also category characteristic curve, CCC)
  - TCC – test characteristic curve
  - IIF/TIF – item and test information function

14

AARHUS UNIVERSITY

Item characteristic curve (ICC)

- The ICC describes the *probability* of a specific response – usually a positive/correct answer – to item i as a function of the latent variable θ:

$$P_i(\theta) = P(X_i = 1 \mid \theta)$$

- Two approaches to the form of the ICC:

Parametric IRT
- The ICC follows a specified function (logistic)
- Each item is described by parameters, e.g. location (b) and discrimination (a)
- Examples: 1PL, 2PL, 3PL (next slides)

Hambleton et al. (1991); van der Linden & Hambleton (1997)

Non-parametric IRT
- Only assumes that the ICC is increasing (monotone)
- No parameters – items are ordered by their observed popularity
- Example: Mokken scaling – a probabilistic Guttman scale

Mokken (1971); Sijtsma & Molenaar (2002)

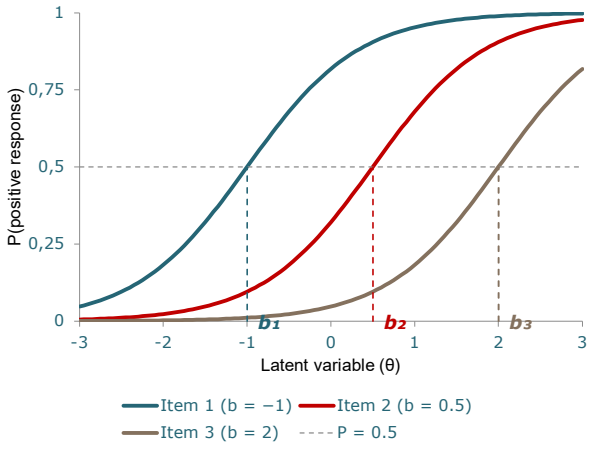
15

AARHUS UNIVERSITY

ICC: 1-parameter logistic model (1PL)

$$P_i(\theta) = \exp(a(\theta - b_i)) / [1 + \exp(a(\theta - b_i))]$$

- Only one parameter characterizes an item – the location/difficulty b_i (θ where P = 0.5)
- All items share the same discrimination a – the ICCs are parallel and never cross
- Rasch model: a = 1 (more in the Rasch session)



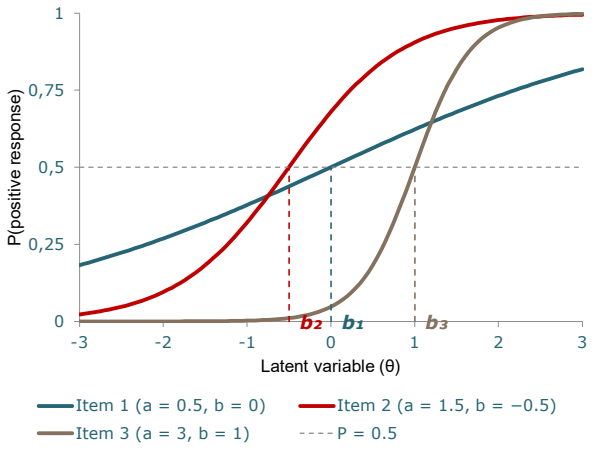
16

AARHUS UNIVERSITY

ICC: 2-parameter logistic model (2PL)

$$P_i(\theta) = \exp(a_i(\theta - b_i)) / [1 + \exp(a_i(\theta - b_i))]$$

- Each item has its own discrimination a_i – the slope of the ICC at b_i
- b_i is still the value of θ where P = 0.5
- The ICCs can cross: which item is "easiest" depends on θ



17

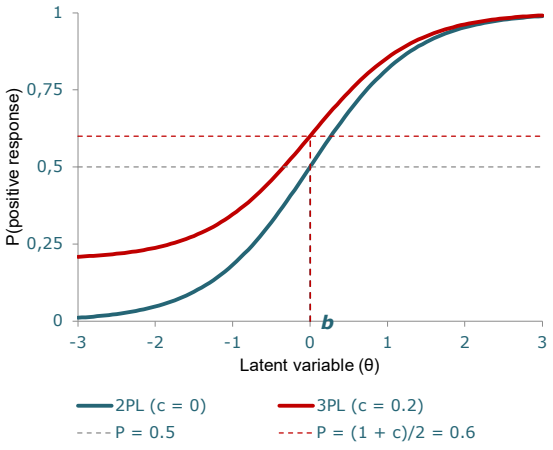
AARHUS UNIVERSITY

ICC: 3- and 4-parameter logistic models (3PL, 4PL)

- 3PL: a lower asymptote c_i ("guessing")

$$P_i(\theta) = c_i + (1 - c_i) \cdot [2PL\ ICC]$$

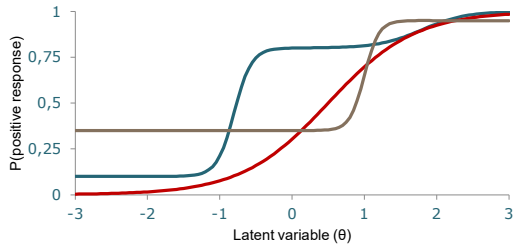
- 4PL: also an upper asymptote $d_i < 1$ ("carelessness")
- **Note:** with $c_i > 0$, b_i is no longer at P = 0.5 but at P = (1 + c_i)/2
- **PROMs:** guessing is rarely relevant – 3PL/4PL are mainly used for multiple-choice tests. In clinical research 1PL/Rasch, 2PL and models for ordinal items are most used.

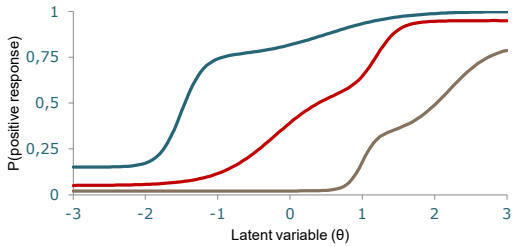


18

AARHUS UNIVERSITY

Non-parametric IRT: two models





Monotone homogeneity model (MHM)

- ICCs only have to be non-decreasing – any shape
- ICCs may cross
- Non-parametric counterpart of the 2PL

Double monotonicity model (DMM)

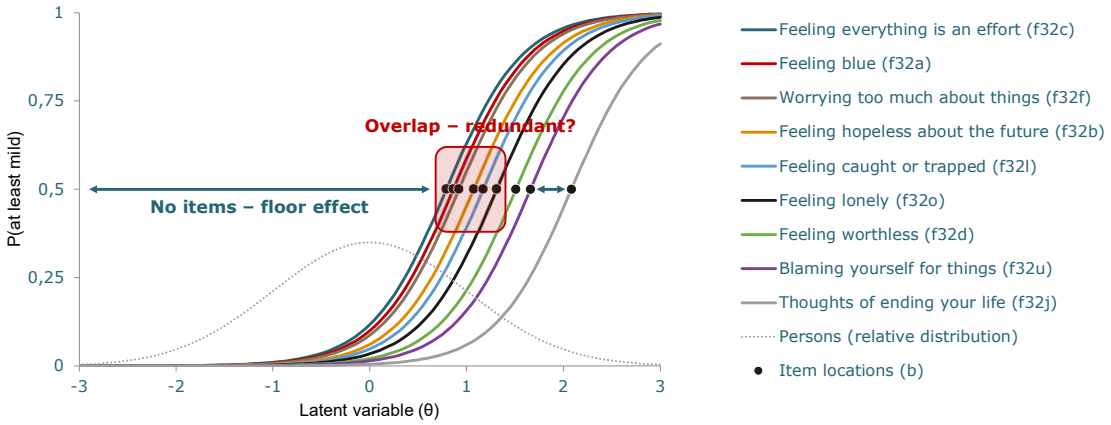
- In addition: ICCs do not cross
- Same item ordering for all persons (invariant item ordering)
- Non-parametric counterpart of the 1PL/Rasch model

Sijtsma & Molenaar (2002)

19

AARHUS UNIVERSITY

ICC example: gaps and overlap – depression items



1PL model fitted to all 482 persons (common discrimination a = 2.55). Positive response = at least mild.

20

ICC example: gaps and overlap – SF-36 physical functioning

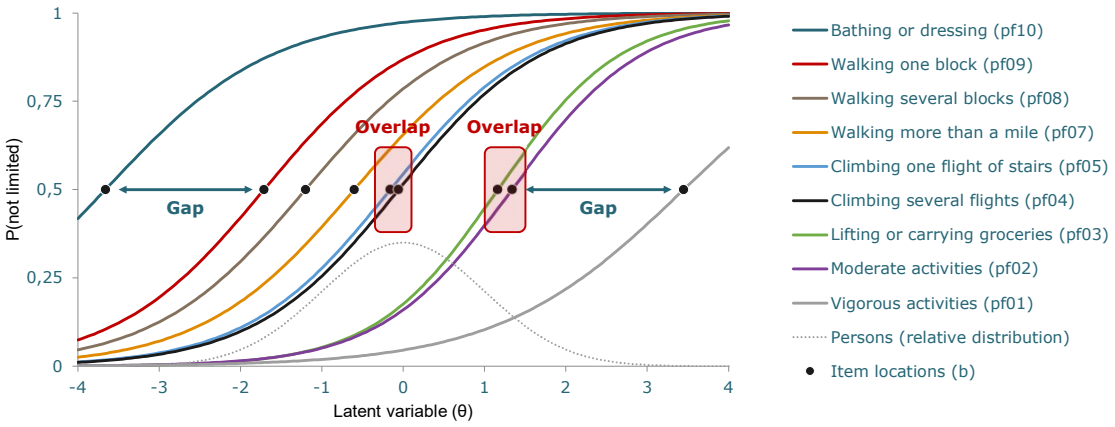


Illustration – item parameters approximated from an earlier analysis. Item pf06 (bending, kneeling or stooping) was not shown in the original graph.

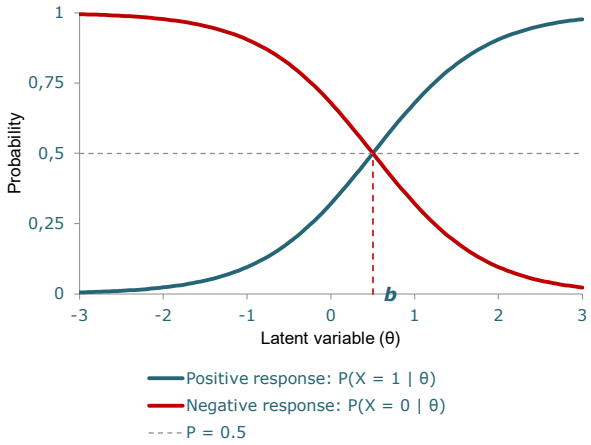
21

From ICC to category probability curves (CPC)

- With two response categories, the curve for the negative response is the mirror image of the ICC:

$$P(X_i = 0 \mid \theta) = 1 - P(X_i = 1 \mid \theta)$$

- The two curves cross at P = 0.5 – i.e. at the item location b_i (1PL/2PL)
- With more than two categories, there is one curve per category – the category probability curves (next slides)



22

AARHUS UNIVERSITY

Items with more than two ordered categories

- Each category has its own curve – the category probability curves (CPC)
- Two families of models:

Graded response model (GRM)
- Cumulative: $b_k = \theta$ where $P(X \geq k) = 0.5$
- Extension of the 2PL – one discrimination per item
- Thresholds are always ordered
- Most used in general IRT (e.g. PROMIS)

Samejima (1969)

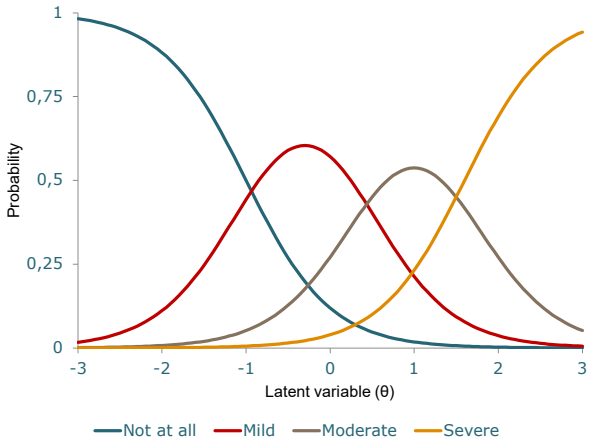
Partial credit / rating scale model (PCM/RSM)
- Adjacent categories: threshold = θ where two adjacent categories are equally likely
- Belongs to the Rasch family
- Thresholds can be disordered → Rasch session

Masters (1982); Andrich (1978)

AARHUS UNIVERSITY

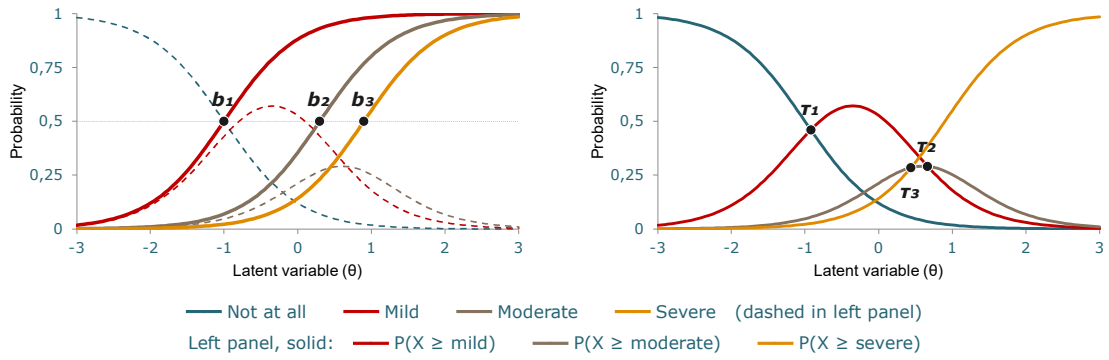
CPC: example with four response categories

- Hypothetical item with the four categories of the depression scale: not at all, mild, moderate, severe
- Each category is the most likely response in a range of θ
- The lowest and highest categories are monotone; the middle categories are bell-shaped
- Curves from a graded response model (a = 2; b = −1, 0.4, 1.6)



AARHUS UNIVERSITY

Thresholds: GRM vs PCM/RSM – same curves, two definitions



GRM (general IRT)

- b_k = θ where P(X ≥ k) = 0.5 (solid cumulative curves)
- $b_1 < b_2 < b_3$ – always ordered

PCM/RSM (Rasch family)

- Threshold τ_k = θ where categories k−1 and k are equally likely
- Here $\tau_3 < \tau_2$: disordered – "moderate" is never the most likely response → Rasch session

25

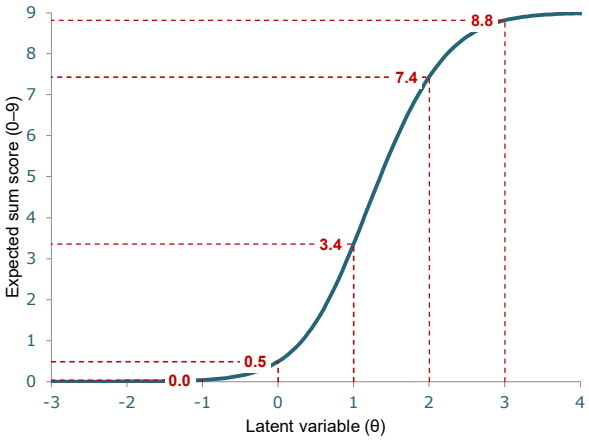
AARHUS UNIVERSITY

Test characteristic curve (TCC)

- The TCC is the sum of the ICCs:

$$TCC(\theta) = \Sigma_i P_i(\theta)$$

- It gives the expected sum score for a person at a given θ – a translation between the IRT scale and the familiar sum score
- The TCC is not linear: equal steps in θ do not give equal steps in the sum score
- The sum score is compressed at the ends of the scale (floor/ceiling)

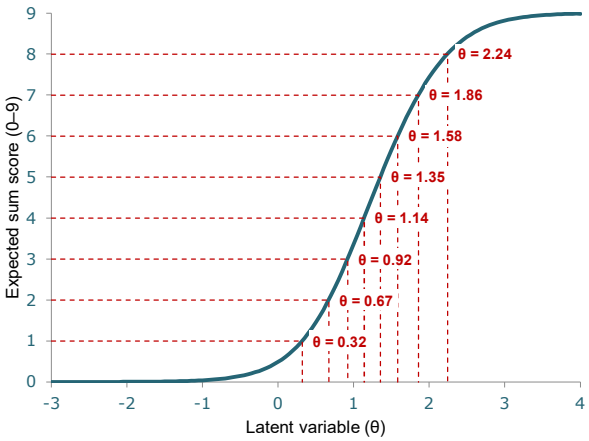


26

AARHUS UNIVERSITY

TCC: equal steps in the sum score

- Read the TCC the other way: from the sum score to θ
- Equal steps of one point in the sum score correspond to unequal steps in θ – small in the middle, large at the ends
- Sum score 0 covers everything below θ ≈ 0.3 – here 52% of the persons
- A change of one point therefore does not mean the same change in the latent variable everywhere on the scale (cf. responsiveness and interpretation, day 2)



27

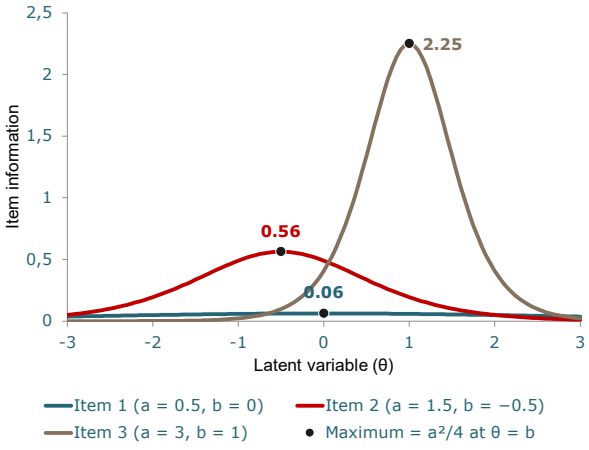
AARHUS UNIVERSITY

Item information

- Item information shows how precisely an item measures θ at different levels

$$I_i(\theta) = a_i^2 \cdot P_i(\theta) \cdot (1 - P_i(\theta)) \quad (1PL/2PL)$$

- Highest at $\theta = b_i$, with maximum $a_i^2/4$ – a more discriminating item gives more information, but over a narrower range
- Rasch model (a = 1): I = P(1 − P), maximum 0.25
- For items with more than two categories, information is computed from all category curves
- Non-parametric IRT has no information function – instead scalability coefficients (H) and a global reliability coefficient are used (Sijtsma & Molenaar, 2002)



28

AARHUS UNIVERSITY

Test information

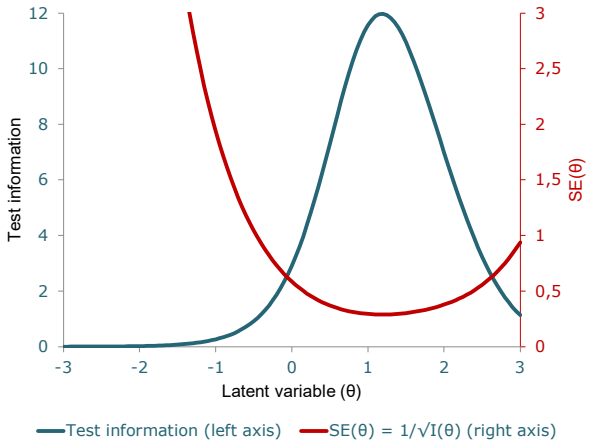
- Test information is the sum of the item information:

$$I(\theta) = \Sigma_i I_i(\theta)$$

- Standard error of the person estimate:

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

- Precision varies with θ – in contrast to the single SEM in CTT (slide 3)
- The test measures best where the items are located (θ ≈ 0.8–2) – and poorly at low levels, where most persons are (cf. slide 20)

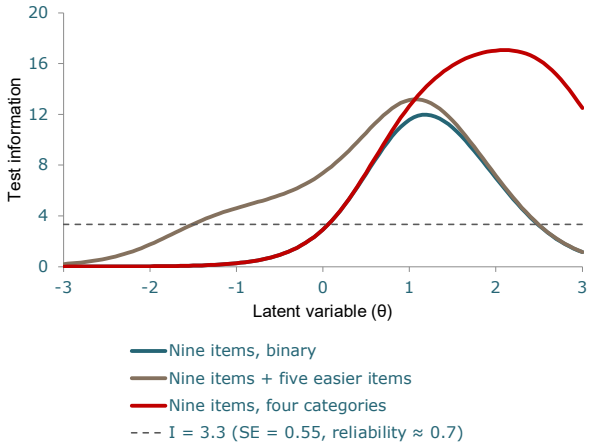


29

AARHUS UNIVERSITY

Test information: improving the test

- Adding easier items (milder symptoms) gives information where most persons are – reducing the floor effect
- Using all four response categories adds information where the extra thresholds are – here mainly above the existing items
- Rule of thumb (Var(θ) = 1): reliability ≈ $1 - SE^2$, so $I(\theta) \geq 3.3 \Leftrightarrow SE \leq 0.55 \Leftrightarrow$ reliability ≥ 0.7



30

AARHUS UNIVERSITY

Checking unidimensionality and monotonicity

Unidimensionality

- Factor analysis of the items – for ordinal items based on polychoric correlations (cf. EFA, day 1)
- Multidimensional IRT models exist but are beyond the scope of this course
- Rasch models have their own tests → Rasch session

Monotonicity

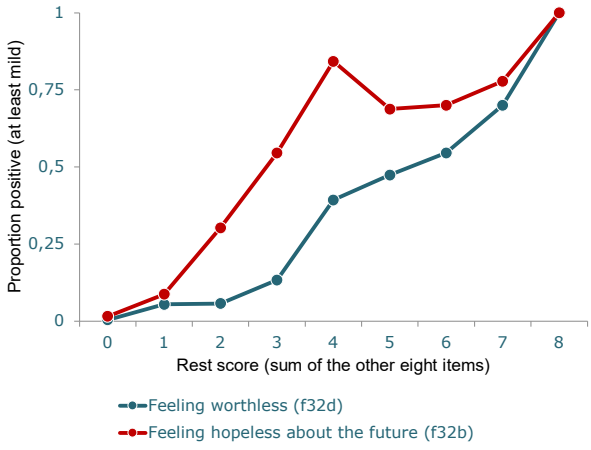
- Parametric IRT: built into the model – checked indirectly through item fit (observed vs. expected proportions by score group)
- Non-parametric IRT (Mokken): rest-score curves and scalability coefficients (H) – next slide

31

AARHUS UNIVERSITY

Monotonicity: rest-score curves

- Rest score = sum score of the other items
- For each rest-score group: the proportion with a positive response to the item
- Monotonicity: the curve should not decrease – small decreases in small groups may be random (tested in Mokken analysis)
- The same idea is used in parametric IRT: observed proportions are compared with the model ICC (item fit)

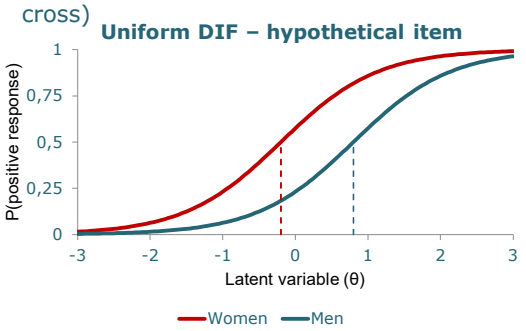


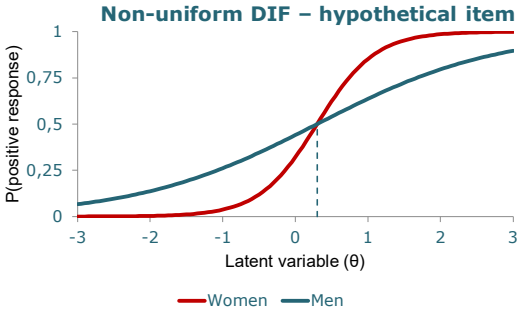
32

AARHUS UNIVERSITY

Check assumptions – no DIF

- DIF: persons with the same θ but from different groups have different probabilities of a positive response
- Uniform DIF: one group has a higher probability at all levels of θ (different b)
- Non-uniform DIF: which group has the higher probability depends on θ (different a – the ICCs cross)





AARHUS UNIVERSITY

Methods for detecting DIF

- Principle: compare the groups for persons at the same level of the latent variable – matched on the sum score or on θ

Observed-score methods
- Matched on the sum score
- Mantel–Haenszel (uniform DIF)
- Logistic regression (uniform and non-uniform DIF)

IRT methods
- Matched on θ
- Compare item parameters between the groups – after linking the scales
- Likelihood ratio test of equal parameters

Rasch methods
- Residual ANOVA, partial gamma, …
- → Rasch session

- Items with more than two categories: ordinal logistic regression, generalised Mantel–Haenszel, or comparison of GRM parameters between groups

Holland & Wainer (1993); Crane et al. (2007); Scott et al. (2010)

AARHUS UNIVERSITY

DIF: Mantel–Haenszel – example

- SF-36 item PF1 (vigorous activities), dichotomised: limited a little / not limited vs. limited a lot
- Persons are matched on the physical functioning score (0–100), here in two score groups

Score group 0–70

	Not limited a lot	Limited a lot	Odds ratio
Women	65	25	**0.74**
Men	70	20	

Score group 71–100

	Not limited a lot	Limited a lot	Odds ratio
Women	85	20	**0.15**
Men	116	4	

Test	H0	Result	Conclusion
Mantel–Haenszel test of the common OR	Common OR = 1 – no uniform DIF	OR(MH) = 0.44 (0.25–0.76); $\chi^2(1) = 8.4$, p = 0.004	H0 rejected → uniform DIF
Breslow–Day test	OR equal in all score groups – no non-uniform DIF	$\chi^2(1) = 6.4$, p = 0.011	H0 rejected → non-uniform DIF

- OR < 1: at the same score, women have lower odds of not being limited a lot. In practice: use finer score groups – two groups is very coarse

AARHUS UNIVERSITY

DIF: logistic regression

- Three nested models for item i (S = sum score, G = group):
  - Model 1: logit $P(X_i = 1) = \beta_0 + \beta_1 \cdot S$
  - Model 2: logit $P(X_i = 1) = \beta_0 + \beta_1 \cdot S + \beta_2 \cdot G$
  - Model 3: logit $P(X_i = 1) = \beta_0 + \beta_1 \cdot S + \beta_2 \cdot G + \beta_3 \cdot G \times S$
- Model 2 vs. 1 – H0: $\beta_2 = 0$ (no uniform DIF). If rejected: uniform DIF – the groups differ by the same amount (on the logit scale) at all levels of S
- Model 3 vs. 2 – H0: $\beta_3 = 0$ (no non-uniform DIF). If rejected: non-uniform DIF – the group difference depends on S
- Example – worrying too much about things (f32f), women vs. men:
  - β_2: OR = 2.05 (95% CI 1.10–3.80), p = 0.023 → H0 rejected: uniform DIF – at the same sum score, women have twice the odds of reporting worrying
  - β_3: p = 0.12 → H0 not rejected: no evidence of non-uniform DIF

AARHUS UNIVERSITY

DIF results for the nine depression items (women vs. men)

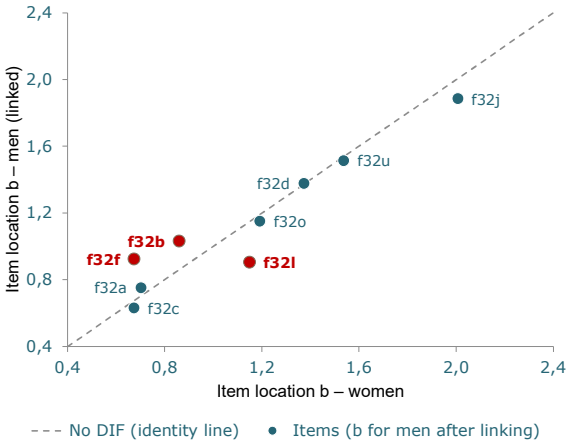
Item	Mantel–Haenszel OR (95% CI)	MH p	Log. reg. uniform p	Log. reg. non-uniform p
Feeling blue (f32a)	1.20 (0.62–2.30)	0.71	0.76	0.40
Feeling hopeless about the future (f32b)	1.35 (0.64–2.86)	0.56	0.19	0.87
Feeling everything is an effort (f32c)	0.90 (0.51–1.60)	0.84	0.90	0.23
Feeling worthless (f32d)	0.93 (0.39–2.24)	0.95	0.83	0.39
Worrying too much about things (f32f)	2.08 (1.12–3.88)	0.029	0.022	0.12
Thoughts of ending your life (f32j)	0.59 (0.19–1.81)	0.53	0.46	0.69
Feeling caught or trapped (f32l)	0.47 (0.23–0.96)	0.054	0.022	0.40
Feeling lonely (f32o)	0.87 (0.44–1.74)	0.83	0.76	0.81
Blaming yourself for things (f32u)	0.83 (0.34–2.03)	0.85	0.70	0.67

- Highlighted: $p < 0.05$ in at least one test. OR > 1: women endorse the item more often than men at the same sum score
- Nine items tested → Bonferroni: $0.05/9 \approx 0.006$ – none of the items shows statistically significant DIF after correction

AARHUS UNIVERSITY

IRT-based DIF: compare item parameters after linking

- Fit the model in each group and compare the item locations (b)
- Separate fits are not on the same scale: each group is standardised to mean 0, SD 1
- Here all b-values for men are on average 0.29 higher – a scale/impact difference, not DIF
- Link the scales, then compare: items far from the line show DIF
  - Anchor items: items assumed to be free of DIF define the common scale – the other items are compared on that scale
  - Multigroup model: all items constrained equal except the studied item; a likelihood ratio test compares the models (IRT-LR-DIF)



AARHUS UNIVERSITY

If DIF is found – then what?

- Statistical significance is not the same as importance – also assess the magnitude, e.g.:
  - Mantel–Haenszel OR outside 0.65–1.53
  - Change in Nagelkerke R^2 between the logistic models > 0.035 (Jodoin & Gierl, 2001)
  - Change in the coefficient of the sum score > 5–10% when the group is added (Crane et al., 2007)
  - No consensus on the criteria
- Possible actions:
  - Remove the item – but this may reduce reliability and content validity
  - Keep the item but model the DIF (group-specific item parameters)
  - Evaluate the impact: does DIF change the person scores or group comparisons?
  - Keep the purpose of the measurement in mind

Example: depression items (slide 36)

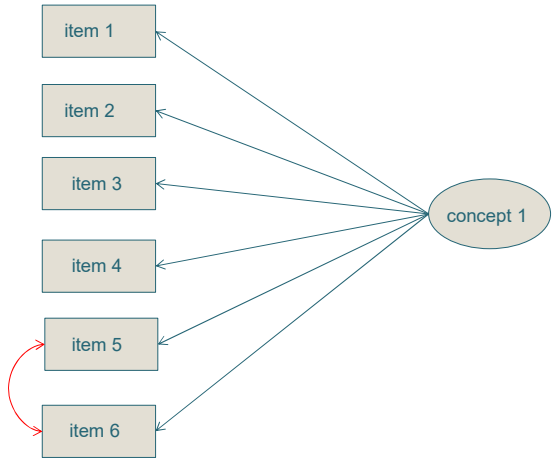
Worrying too much (f32f), OR 2.08 – feeling caught or trapped (f32l), OR 0.47

- Significance: p ≈ 0.02–0.05, not significant after Bonferroni
- Magnitude: both ORs outside 0.65–1.53 → not negligible
- Content: is there a plausible gender difference in how worry and feeling trapped are expressed?
- Impact: DIF in opposite directions may partly cancel in the sum score – compare scores with and without the two items
- Purpose: if men and women are to be compared, check the impact before concluding

39

AARHUS UNIVERSITY

Check assumptions – local independence



- After fitting the model: look at the correlation between the residuals of each pair of items
  - Yen's Q3 – a residual correlation clearly above the average (e.g. > 0.2 above) indicates local dependence
- Possible actions: combine the dependent items into one (testlet / super-item) or remove one of them
- Rasch approach (item as a DIF variable, partial gamma) → Rasch session

40

AARHUS UNIVERSITY

References: IRT models and textbooks

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Addison-Wesley.

Furr, R. M. (2022). *Psychometrics: An introduction* (4th ed.), chapter 14. Sage.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Lord, F. M. (1952). *A theory of test scores*. Psychometric Monographs, No. 7.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Pædagogiske Institut.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. Springer.

AARHUS UNIVERSITY

References: DIF and local dependence

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, 41, 178–194.

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: experience from the New Haven EPESE study. *Journal of Clinical Epidemiology*, 53, 285–289.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., et al. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, 16, 69–84.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Lawrence Erlbaum.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329–349.

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling*, 58, 79–98.

Scott, N. W., Fayers, P. M., Aaronson, N. K., et al. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes*, 8, 81.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum.