AARHUS UNIVERSITY

Rasch analysis

Advanced course in questionnaire technique and clinimetrics

Eva Ørnbøl, Statistician

1

AARHUS UNIVERSITY

Outline

- IRT and RMT
- Why use a Rasch model?
- The Rasch model: definition and properties
- What to look for in a Rasch analysis
  - Model choice, fit, item threshold ordering, uni-dimensionality, DIF, local dependence, targeting, reliability
  - Summary of criteria
- Software and example

2

AARHUS UNIVERSITY

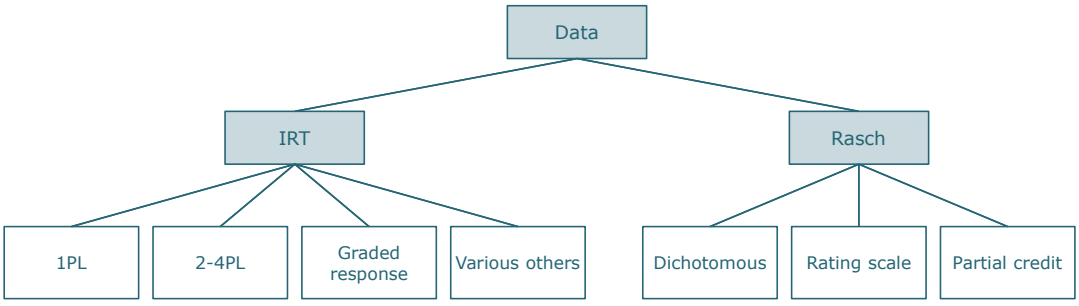
Paradigms of IRT and RMT

- IRT: stochastic models that estimate person and item locations on a latent continuum.
  - The model is chosen to describe the data as well as possible: parameters (discrimination, guessing) are added if they improve fit.
- RMT: the model expresses the requirements for measurement (invariance; the raw score as a sufficient statistic).
  - Data are tested against the model, and misfit is investigated rather than modelled.

3

AARHUS UNIVERSITY

IRT vs. RMT



Focuses on explaining variance

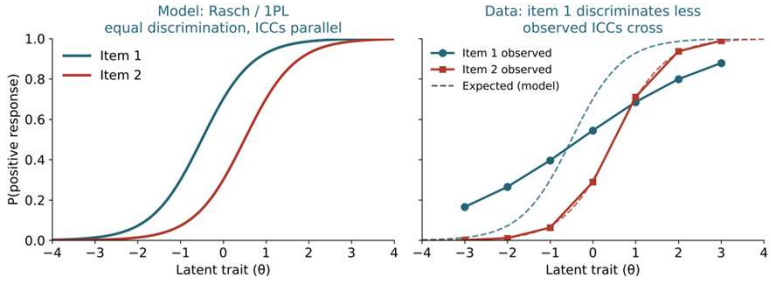
Emphasises the primacy of the model

4

AARHUS UNIVERSITY

Is the Rasch model just a 1PL model?

Mathematically, yes – the same equation. The difference lies in what you do when data do not fit.



IRT: add a discrimination parameter (2PL), so the model follows the data. The item order then depends on the person's location.

Rasch: keep the model. Item 1 misfits (under-fit, positive fit residual) and is investigated, revised or removed.

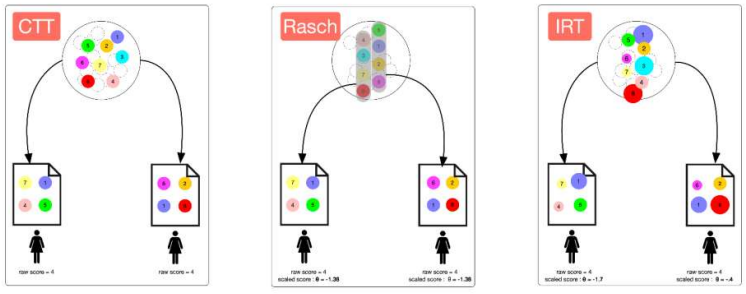
Dichotomous items. Illustration inspired by Stemler & Naples (2021).

5

AARHUS UNIVERSITY

CTT, Rasch and IRT: person estimation

Figure 5. Comparison of approach to person knowledge estimation between CTT, Rasch, and IRT

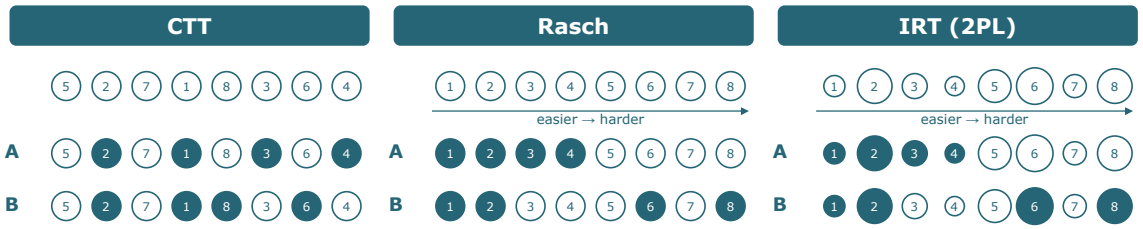


Stemler, S.E. & Naples, A. (2021). Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. Practical Assessment, Research, and Evaluation, 26, Article 11.

6

AARHUS UNIVERSITY

Same sum score – same person estimate?



A = 4, B = 4 → same ability
Which items – and how difficult they are – plays no role.

A = 4, B = 4 → same θ
But B's pattern is unexpected (hard items endorsed, easy ones not) → poor person fit.

A = 4, B = 4 → different θ
Items with high discrimination (large circles) weigh more. B endorsed the more discriminating items → higher θ than A.

Filled circle = item endorsed. Both persons endorse 4 of 8 items. Based on Stemler & Naples (2021), Fig. 5.

7

AARHUS UNIVERSITY

Why use a Rasch model?

- **The sum score holds all the information:** the raw score is a sufficient statistic for θ, so persons with the same sum score get the same θ estimate. Under 2PL/GRM, θ depends on which items were endorsed (an implicitly weighted score).
- **Invariant item order:** the items are ordered the same way for all persons. A given θ (sum score) therefore tells which items a person at that level is likely to endorse – and the score means the same for everyone.
- **Misfit is information:** deviations point to problems with items or the construct, instead of being absorbed by extra parameters.
- **The price:** the requirements are strict, and items often have to be revised or removed. If an existing instrument must be kept as it is, an IRT model may describe it better.

"If what you want to do … is create a truly linear, equal interval measurement scale that works the same way for all test takers …, then only the Rasch model will suffice."
Stemler & Naples (2021), p. 15

8

AARHUS UNIVERSITY

The dichotomous Rasch model

$$P(X_{vi} = 1 \mid \Theta_v = \theta_v) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}$$

or, equivalently (logit form):

$$\log\left[\frac{P(X_{vi} = 1)}{1 - P(X_{vi} = 1)}\right] = \theta_v - \beta_i$$

For a dichotomous item (X = 0/1), the probability of a positive response depends only on the difference between the person's location θ_v and the item's location (difficulty) β_i.

θ_v and β_i are on the same logit scale.

Polytomous items: partial credit and rating scale models (see "Most used Rasch models").

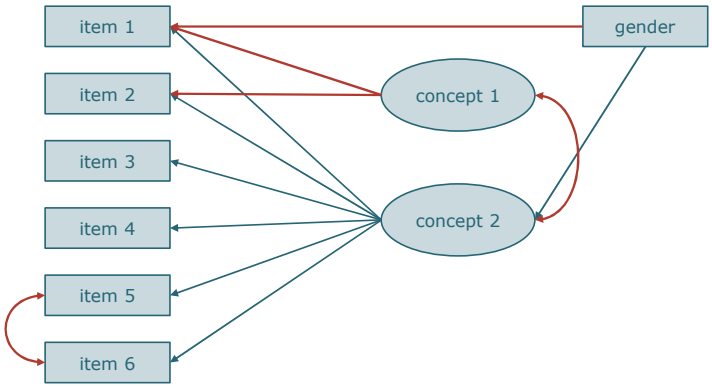
AARHUS UNIVERSITY

Rasch model – assumptions

- Same as the assumptions for IRT models
  - Uni-dimensionality of the concept (latent variable)
  - Local independence
  - No DIF
  - Monotonicity of the ICC
- Additional assumption
  - Equal discrimination for all items (parallel ICCs, no guessing parameter)
- Properties that follow from this
  - Sufficiency of the raw score
  - Specific objectivity

AARHUS UNIVERSITY

Rasch model – violations of the assumptions



As required

One latent variable: concept 2 → all items

Group difference, not DIF: gender → concept 2 (groups may differ on the latent variable)

Violations

Multidimensionality: concept 1 → items 1–2; concept 1 ↔ concept 2 are correlated latent variables

DIF: gender → item 1

Local dependence: item 5 ↔ item 6

11

AARHUS UNIVERSITY

Rasch model – sufficiency

- **Persons:** the sum score contains all the information about θ in the response pattern. Persons with the same sum score get the same θ (A and B, slide 7).
- **Items:** the item total (the number of persons endorsing the item) contains all the information about β.
- **Fit, not estimation:** the response pattern is used to check fit – A and B get the same θ, but B has poor person fit.
- **Link to specific objectivity:** conditioning on the sum score removes θ from the likelihood (CML) → item estimates that do not depend on the person distribution.
- **Practical message:** among IRT models, only the Rasch model makes the unweighted sum score sufficient (Andersen, 1977). Summing questionnaire items implicitly assumes a Rasch model.
- **But sufficient ≠ interval:** the relation between sum score and θ is non-linear (S-shaped); θ is the interval-scale measure.

12

AARHUS UNIVERSITY

Rasch model – specific objectivity

Persons: item-independent ("test-free")

The difference between two persons does not depend on which items are used.

$$\text{logit}\, P_{vi} - \text{logit}\, P_{wi} = (\theta_v - \beta_i) - (\theta_w - \beta_i) = \theta_v - \theta_w$$

→ Person estimates are on the same scale, whichever items were answered.

Example: two patients can be compared even if they answered different subsets of items, e.g. a short form, or an item bank used for computer adaptive testing (CAT).

Items: sample-independent ("sample-free")

The difference between two items does not depend on which persons are sampled.

$$\text{logit}\, P_{vi} - \text{logit}\, P_{vj} = (\theta_v - \beta_i) - (\theta_v - \beta_j) = \beta_j - \beta_i$$

→ Item parameters can be estimated conditional on the sum score (CML), without assumptions about the person distribution.

Example: "dressing" is easier than "lifting heavy objects", whether the items are calibrated in a mildly affected primary-care sample or a severely affected hospital sample.

Specific: objectivity holds only within the specific frame of reference – the items and population for which the data fit the model. It must be tested (fit, DIF) – it cannot be assumed.

AARHUS UNIVERSITY

Most used Rasch models

Model	Response format	Thresholds
Dichotomous Rasch model	0/1	One threshold per item = the item location β_i
Rating scale model (Andrich, 1978)	The same categories for all items	The same distances between thresholds for all items; only the item location shifts
Partial credit model (Masters, 1982)	Ordered categories; the number of categories may differ between items	Item-specific thresholds β_{ik}; the distances between thresholds may differ between items

All three models: equal discrimination for all items.

Threshold: the point on θ where two adjacent categories are equally likely (their category probability curves cross).

AARHUS UNIVERSITY

What to look for in a Rasch analysis?

Model	Fit	Assumptions	Scale properties
Model choice: dichotomous, rating scale or partial credit	Fit to the model (overall fit) Item and person fit Item threshold ordering	Uni-dimensionality Differential item functioning (DIF) Local dependence	Targeting Reliability

Action taken and re-analysis: rescoring of categories, deletion of items, subtests/testlets – then check everything again (an iterative process).

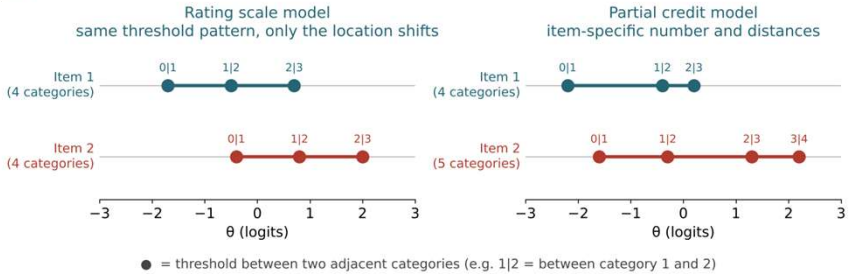
Based on Tennant & Conaghan (2007).

15

AARHUS UNIVERSITY

Model choice – rating scale or partial credit?

Dichotomous items → the dichotomous Rasch model. Polytomous items → rating scale or partial credit



Rating scale: $\beta_{ik} = \beta_i + \tau_k$ – one set of threshold distances τ_k for all items. Requires the same response categories for all items.

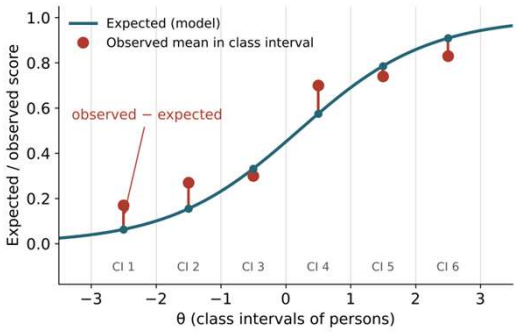
Partial credit: each item has its own thresholds β_{ik} – both the distances and the number of categories may differ between items.

Choosing the model: inspect the threshold map and/or use a likelihood-ratio test of RSM vs PCM. Report the model chosen and why.

16

AARHUS UNIVERSITY

Fit – overall fit



How the differences are computed

1. For each person and item: residual = observed score − expected score (from θ_v and the item parameters), standardised by its SD.

2. Persons are grouped into class intervals (CI) along θ. For each item and CI, the observed mean score is compared with the expected mean score.

3. Squared standardised differences summed over CIs → item χ^2 (df = CIs − 1); summed over all items → the overall test.

Large differences in the same direction across CIs → misfit. Here the item is flatter than expected (under-fit, cf. slide 5).

17

AARHUS UNIVERSITY

Fit – overall fit: tests and interpretation

Item–trait interaction χ^2

The item χ^2 summed over all items (reported by RUMM).

Significant → the item hierarchy is not the same along the trait: invariance is violated.

Criterion: non-significant.

Andersen's conditional likelihood-ratio test

Item parameters are estimated separately in low- and high-score groups and compared (Stata raschtest; R eRm, iarm).

A direct test of item invariance (specific objectivity, slide 13).

Criterion: non-significant.

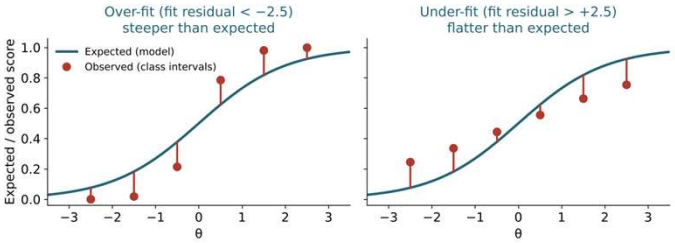
Caveat – χ^2 depends on sample size: with large N even trivial misfit is significant; with small N real misfit may go undetected. Look at the size of the misfit, not just the p-value, and consider checking robustness in random subsamples.

Statistics and labels differ between programs – check which one is reported.

18

AARHUS UNIVERSITY

Fit – item fit



Fit residual per item: the standardised sum of residuals over persons. Criterion: within ±2.5.

< −2.5: responses too predictable – redundancy or local dependence.

> +2.5: noisy responses – low discrimination, another construct, or DIF.

Item χ^2 (slide 17) for each item, with **Bonferroni correction**: many tests are made, so with 10 items use $p < 0.05/10 = 0.005$ for each item.

Winsteps reports **infit/outfit mean squares** instead (expected value 1; often 0.7–1.3 is accepted).

Caveat: ±2.5 is a convention, and like χ^2 the fit residual depends on sample size.

19

AARHUS UNIVERSITY

Fit – person fit and summary fit statistics

Person fit residual: the same principle, computed for each person across items.

> +2.5: aberrant response pattern – e.g. careless responding, unrecorded comorbidity, cognitive problems (cf. person B, slide 7).

< −2.5: response pattern too deterministic.

A few misfitting persons can distort item fit. They may be removed, but report it, check whether the conclusions change – and ask for whom the scale does not work.

Summary statistics (items and persons)

	Ideal	Problem
Mean fit residual	≈ 0	far from 0
SD of fit residuals	≈ 1	> 1.4

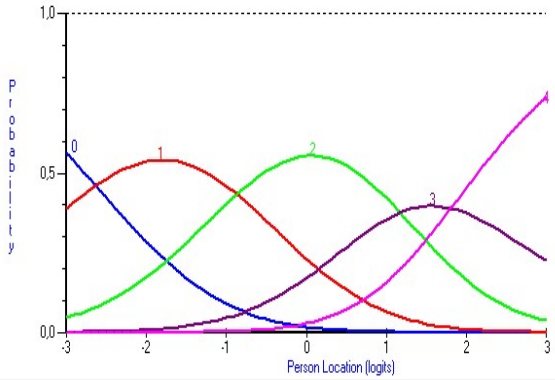
Usually reported as mean (SD) for items and for persons in the summary table of a Rasch paper.

20

AARHUS UNIVERSITY

Fit – item threshold ordering: ordered

Each category is the most probable response somewhere on θ.



PSS item 3: "In the last month, how often have you felt nervous and stressed?"

How to read a CPC: x-axis = person location (θ); one curve per response category; thresholds are where adjacent curves cross (slide 14).

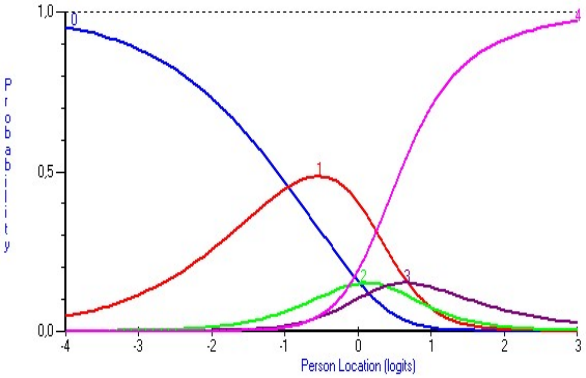
Category probability curves from RUMM2030. Nielsen, M.G., Ørnbøl, E. et al. (2016). The construct validity of the Perceived Stress Scale. J Psychosom Res, 84, 22–30.

21

AARHUS UNIVERSITY

Fit – item threshold ordering: disordered

Categories 2 and 3 are never the most probable response.



PSS item 4, before rescoring: "In the last month, how often have you felt confident about your ability to handle personal problems?"

Note: item 4 is positively worded and reverse-scored before analysis. Reversed items often show disordered thresholds.

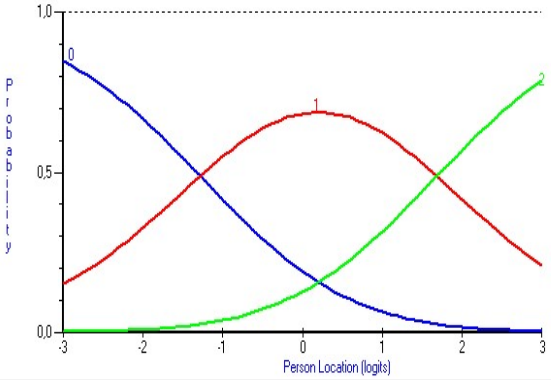
Category probability curves from RUMM2030. Nielsen, M.G., Ørnbøl, E. et al. (2016). The construct validity of the Perceived Stress Scale. J Psychosom Res, 84, 22–30.

22

AARHUS UNIVERSITY

Action taken – collapsing categories

After rescoring (0 → 0; 1, 2, 3 → 1; 4 → 2) the thresholds are ordered.



PSS item 4, after rescoring: the middle categories ("often", "sometimes", "rarely") were collapsed.

Caveat: collapsing changes the scoring of the instrument. Base it on the category labels (do they make sense?), report it, and confirm it in new data. Ordered thresholds do not mean that the item fits.

Category probability curves from RUMM2030. Nielsen, M.G., Ørnbøl, E. et al. (2016). The construct validity of the Perceived Stress Scale. J Psychosom Res, 84, 22–30.

23

AARHUS UNIVERSITY

Check assumptions – uni-dimensionality (Smith's t-test)

1. PCA of the residuals → 2. Split the items by the sign of their loading on the 1st component → 3. Estimate θ for each person from each item subset → 4. t-test of the two estimates for each person → 5. % of significant t-tests (|t| > 1.96), with binomial 95% CI

Criterion: less than 5% of the t-tests should be significant. Some significant tests are expected by chance, so compute a binomial 95% CI for the proportion: the lower bound of the CI should be ≤ 5%.

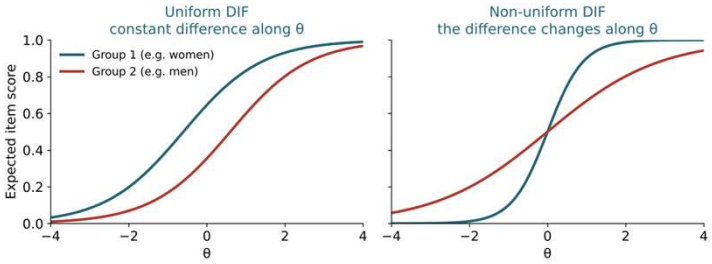
Example – PSS-10 (Nielsen et al., 2016): Rasch analysis and CFA supported two dimensions – the positively and the negatively worded items – whereas Mokken analysis suggested one. Different methods can give different answers.

24

AARHUS UNIVERSITY

Check assumptions – no DIF

DIF: persons from different groups with the same θ respond differently to an item. A difference between the groups on θ itself is not DIF (slide 11).



Uniform DIF: a constant difference between the groups along θ.

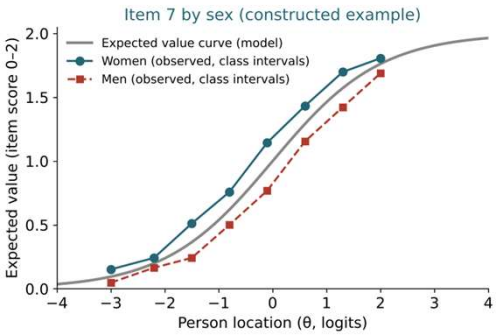
Non-uniform DIF: the difference changes along θ.

Which variables? At least age and sex (Tennant & Conaghan, 2007) – plus those relevant to the research question, e.g. language/country, diagnosis, or recruitment source.

25

AARHUS UNIVERSITY

Check assumptions – no DIF: methods and an example



Women score consistently higher than men at the same θ → uniform DIF.

Methods

ANOVA of the residuals, group × class interval (RUMM): main effect = uniform DIF; interaction = non-uniform DIF. Use Bonferroni (slide 19).

Conditional likelihood-ratio test by group (e.g. Stata raschtest – see the example at the end).

Partial gamma / Mantel–Haenszel (e.g. R iarm, DIGRAM) – also used for local independence.

What to do

Split the item by group, delete it, or accept it if the DIF is small. **Judge the size of the DIF, not only the p-value** (cf. DIF magnitude and multiple testing in the IRT introduction).

26

AARHUS UNIVERSITY

Check assumptions – local independence

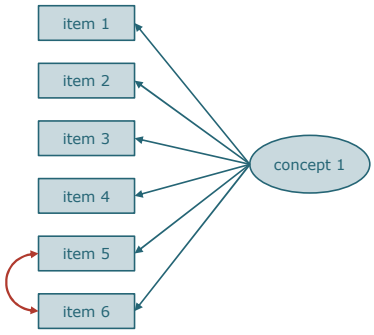
- It can be difficult to determine whether problems are due to local dependence or to lack of uni-dimensionality – both break local independence.
  - **Response dependence:** the answer to one item determines the answer to another – e.g. if you can walk several miles without difficulty, you can also walk one mile (Tennant & Conaghan, 2007).
- Examine the residual correlations between pairs of items:
  - Judge them relative to the average residual correlation, e.g. flag pairs above the average + 0.2 (Christensen et al., 2017).
  - Fixed cut-offs (0.2, 0.3) depend on the number of items, because the average residual correlation is negative.
- Alternatively, use the trick of turning local independence into a question of DIF (next slides).
- **Consequence:** local dependence inflates the reliability (PSI).
- **What to do:** combine dependent items into a subtest/testlet – e.g. one walking item with distance as the response categories.

27

AARHUS UNIVERSITY

Check assumptions – local independence

- The trick is to treat an item as a possible grouping variable you want to investigate for DIF.
- For each pair of items, you need to regard each one as the grouping variable and both should elicit DIF for local dependence to take place.
- You can apply the same range of methods as mentioned earlier for DIF.

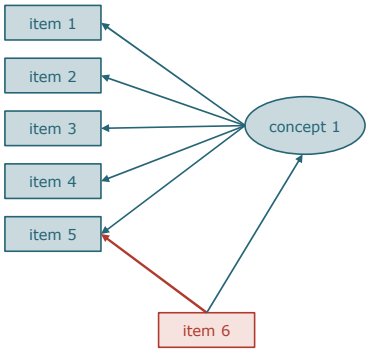


Item 5 ↔ item 6: local dependence (cf. slide 11)

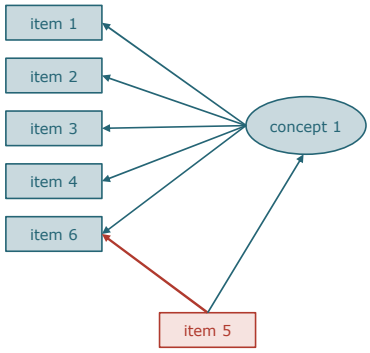
28

AARHUS UNIVERSITY

Check assumptions – local independence



Item 6 as grouping variable → DIF in item 5?



Item 5 as grouping variable → DIF in item 6?

Red arrow = DIF. Local dependence is indicated if both show DIF. The petrol arrow to concept 1 is allowed: the item may correlate with θ.

29

AARHUS UNIVERSITY

Check assumptions – local independence: partial gamma

Example: f32a "Feeling blue" by f32f "Worrying too much about things" (N = 482). Gamma = rank correlation for ordinal items; partial gamma = the same, conditional on the rest score.

	γ	95% CI	p
Unstratified	0.79	0.70 to 0.88	< 0.001
Partial (conditional on the rest score; 196 persons in informative strata)	−0.04	−0.38 to 0.29	0.81
Homogeneity of γ across strata	$\chi^2(4) = 2.91$		0.71

Interpretation: the two items are strongly associated overall – expected, as both measure depression. With f32f as grouping variable and conditional on the rest score, the association disappears (γ ≈ 0) → no DIF in f32a by f32f → no sign of local dependence.

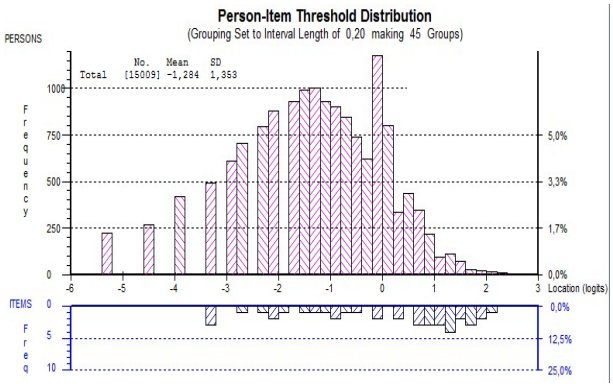
Remember: local dependence requires DIF in both directions. Here the first test shows no DIF, so the second direction (f32f by f32a) cannot change the conclusion – but if the first test shows DIF, the other direction must be tested.

Power: γ can only be computed in rest-score groups where both items vary. In groups 0 and 6–8 they do not, so only 196 of 482 contribute (floor effect – cf. targeting).

30

AARHUS UNIVERSITY

Scale properties – targeting



How to read: persons (top) and item thresholds (bottom) on the same logit scale; the items are centred at 0.

Well targeted: mean person location ≈ 0, and the item thresholds cover the range where the persons are.

Here (PSS-10, N = 15,009): mean person location −1.28 – the sample is less stressed than the items target. Few thresholds below about −3 → imprecise measurement (large SE) of persons with low stress: a floor effect.

Precision depends on targeting: estimates are sample-independent, but not equally precise (slide 13). Cf. the partial gamma example, where only 196 of 482 persons contributed.

Person–item threshold distribution from RUMM2030. Nielsen, M.G., Ørnbøl, E. et al. (2016). The construct validity of the Perceived Stress Scale. J Psychosom Res, 84, 22–30.

AARHUS UNIVERSITY

Scale properties – reliability

- **Person separation index (PSI):**
  - Computed with the same formula as Cronbach's alpha, but on the logit person estimates instead of the raw scores. Interpreted in the same way.
  - Minimum 0.70 for group comparisons; minimum 0.85 for individual use (Tennant & Conaghan, 2007).
  - Winsteps reports the person separation ratio G instead: G = 1.5 corresponds to a reliability of about 0.70, G = 2.5 to about 0.86.
- **Interpreting the PSI:**
  - High is not always good: local dependence inflates the PSI (slide 27).
  - The PSI depends on targeting: poor targeting (slide 31) gives large SEs and a lower PSI. It shows how well this sample can be separated, not only how good the items are.
  - Good reliability does not mean fit to the model: in PSS-10 the PSI was 0.85 and Cronbach's α 0.86, yet the data did not fit the Rasch model (Nielsen et al., 2016).

AARHUS UNIVERSITY

Summary: what to look for

Aspect	Statistic	Criterion
Model choice	Threshold map; likelihood-ratio test RSM vs PCM	Report the model chosen and why
Overall fit	Item–trait interaction χ^2; Andersen's conditional LR test	Non-significant (power depends on N)
Item fit	Fit residuals; item χ^2 (Winsteps: infit/outfit)	Each item within ±2.5; χ^2 non-significant after Bonferroni correction
Item and person fit, summary	Mean (SD) of fit residuals	Mean ≈ 0, SD ≈ 1 (SD ≤ 1.4)
Fit residual sign	< −2.5 / > +2.5	Over-fit (redundancy, local dependence) / under-fit (low discrimination, other construct, DIF)
Thresholds	Category probability curves, threshold map	Ordered thresholds; otherwise consider collapsing categories
Uni-dimensionality	Smith's t-tests (PCA of residuals)	< 5% significant tests; lower bound of the binomial 95% CI ≤ 5%
DIF	ANOVA of residuals; conditional LR test; partial gamma	No DIF for age, sex and relevant groups; Bonferroni; judge the size
Local dependence	Residual correlations; item as DIF variable (partial gamma)	Flag pairs > average residual correlation + 0.2; local dependence if DIF in both directions
Targeting	Person–item threshold distribution	Mean person location close to 0; thresholds cover the persons
Reliability	Person separation index (PSI)	≥ 0.70 for groups, ≥ 0.85 for individuals; high PSI ≠ fit
Action taken	Rescoring, item deletion, subtests/testlets	Report what was done and the fit after each step

33

AARHUS UNIVERSITY

Rasch analysis – software

Dedicated programs

RUMM2030 (used in most of the examples)

Winsteps

DIGRAM

R

eRm, psychotools (CML)

iarm (fit, DIF, local dependence – partial gamma)

TAM, mirt (MML)

Stata

raschtestv7 (CML; SSC)

irt 1pl / pcm / rsm (built-in, MML – assumes normally distributed θ)

CML vs MML: conditional maximum likelihood uses the sum score as sufficient statistic and needs no assumption about the distribution of θ (slides 12–13). Marginal maximum likelihood assumes a distribution for θ.

Statistics and labels differ between programs – check which ones are reported (slide 18).

34

AARHUS UNIVERSITY

Example: nine depression items (f32) – model and overall fit

```
raschtestv7 f32a f32b f32c f32d f32f f32j f32l f32o f32u, dif(sex age55) icc genfit(outfit_p infit_p)
```

Data: nine depression items, N = 482, coded 0/1 (positive = at least mild).

Model: items coded 0/1 → the dichotomous Rasch model (slide 16), estimated by CML. Threshold ordering (slides 21–22) is not relevant: one threshold per item.

Who contributes? 256 of 482 have score 0 or 9 (52% score 0) and carry no information in CML – the analysis rests on 226 persons. A floor effect: a targeting problem (slide 31).

Overall fit: R1c = 82.3 (df 56), p = 0.013; Andersen LR = 81.9 (df 56), p = 0.014 → misfit at the 5% level (slides 17–18).

Caution: some score groups have fewer than 30 persons, so the tests can be unreliable. Next: which items and persons misfit?

AARHUS UNIVERSITY

Example: item and person fit

Item	Difficulty	R1c p	Outfit z	Infit z
f32a Feeling blue	−1.02	0.12	0.18	−1.03
f32b Hopeless about the future	**−0.49**	**0.046**	**−2.09**	**−2.14**
f32c Everything is an effort	**−1.20**	**0.050**	**1.91**	**2.64**
f32d Feeling worthless	0.62	0.25	−1.56	−1.23
f32f Worrying too much	−0.87	0.30	0.54	1.10
f32j Thoughts of ending your life	2.09	0.43	0.15	0.50
f32l Feeling caught or trapped	−0.24	0.55	−0.54	−0.19
f32o Feeling lonely	0.11	0.55	1.25	0.65
f32u Blaming yourself	1.00	0.57	−1.07	−0.91

Difficulties centred at 0. Bonferroni: 0.05/9 = 0.0056 – no single item χ^2 (R1c) is significant.

Item fit: f32b over-fits (z ≈ −2.1), f32c under-fits (infit z = 2.6).

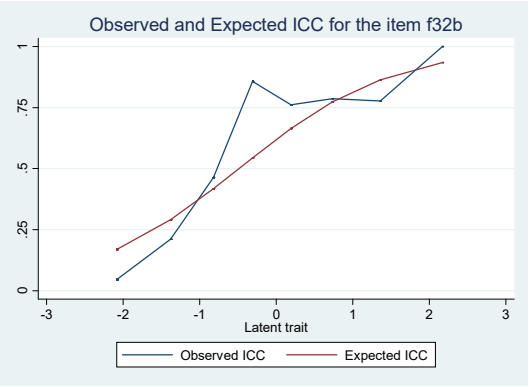
Item summary (z): infit −0.07 (SD 1.45), outfit −0.14 (SD 1.31) – SD around 1.4: some item misfit (slides 19–20).

Person fit (mean squares, expected 1; 226 persons): infit 0.99 (SD 0.30), outfit 0.99 (SD 1.01, max 12.0) – good on average, but a few very unexpected response patterns (cf. person B, slide 7).

Link to the IRT introduction: in the 2PL model, f32b had the highest and f32c the lowest discrimination.

AARHUS UNIVERSITY

Example: over-fit, uni-dimensionality and DIF



f32b: the observed ICC is steeper than expected → over-fit.

Observed vs expected ICC (slides 17 and 19): persons grouped by score, observed proportion positive compared with the model.

Uni-dimensionality (Smith, slide 24): items split by the sign of the first residual component (a, b, c, l vs d, f, j, o, u). 4 of 226 t-tests significant = 1.8% (95% CI 0.5–4.5%) → no evidence against uni-dimensionality, but low power with 4–5 items per subset.

Local dependence (slide 30): only the pair f32a "Feeling blue" / f32f "Worrying too much" was examined (partial gamma) – no sign of local dependence.

DIF – conditional LR test by group:

Sex: Z = 12.2 (df 8), p = 0.14 → no DIF.

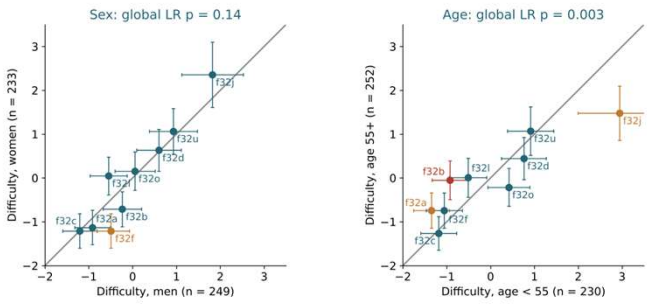
Age (< 55 vs 55+): Z = 23.2 (df 8), p = 0.003 → DIF.

Next step: the LR test is global – which item(s)? Next slide.

37

AARHUS UNIVERSITY

Example: which item? Difficulties by group



Method: estimate the difficulties separately in each group (centred at 0) and compare, $z = (b_1 - b_2) / \sqrt{(SE_1^2 + SE_2^2)}$. Items on the line have no DIF.

Sex: no item beyond Bonferroni; f32f "Worrying too much" is closest (z = 2.4) – consistent with the global test.

Age: f32b "Hopeless about the future" is easier for those under 55 (z = −2.8, beyond Bonferroni); f32j "Thoughts of ending your life" is easier for those 55+ (z = 2.5). Judge the size before acting (slide 26).

38

AARHUS UNIVERSITY

Example: the final scale

After several iterations (one item removed at a time, refitting after each step): **kept f32a, f32d, f32j, f32l, f32o, f32u; removed f32b, f32c, f32f.**

Why one at a time? Fit and DIF are relative to the other items – removing an item changes the sum score, so an item can stand out (or stop standing out) at the next step.

	9 items	6 items
Overall fit: R1c p / Andersen LR p	0.013 / 0.014	0.79 / 0.78
Item fit: largest |z|	2.6 (f32c)	1.6 (f32u)
Person fit: infit SD / outfit max	0.30 / 12.0	0.49 / 8.5
DIF sex: LR p	0.14	0.55
DIF age: LR p	0.003	**0.004**
Score 0 or max (no information)	256 (53%)	**315 (65%)**
Persons contributing	226	**167**
Reliability: PSI	0.53	**0.22**

Fit is good, but age DIF is not solved. Difficulties by age group (slide 38): f32a "Feeling blue" is easier for those under 55 (0.85 logits, z = −2.7, beyond Bonferroni 0.05/6); f32j is easier for those 55+ (1.34 logits, z = 2.5) – large differences (slide 26).

Targeting and reliability are worse: 65% score 0 or 6, and PSI falls from 0.53 to 0.22 – both far below 0.70 (slide 32).

Content and face validity: statistics alone do not decide. Do the six items look like a sensible depression scale to patients and clinicians – e.g. f32j "Thoughts of ending your life" fits, but does it belong in a short scale? And is the construct still covered?

Next: decide how to handle age DIF, then validate the scale in a new sample – the items were chosen on the same data.

AARHUS UNIVERSITY

References

- Rasch models in Health, Ed. Christensen, K. B., Kreiner, S. & Mesbah, M. 2013 Wiley
- Petrillo, J. et al (2015).Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. *Value in Health*, Vol. 18, p. 25-34.
- Tennant, A. & Conaghan, P. G. (2007). The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? Arthritis & Rheumatism, Vol. 57, No. 8, p. 1358–1362.
- 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(3), 178–194.
- Nielsen, M. G., Ørnbøl, E., Vestergaard, M., Bech, P., Larsen, F. B., Lasgaard, M. & Christensen, K. S. (2016). The construct validity of the Perceived Stress Scale. Journal of Psychosomatic Research, 84, 22–30.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. Psychometrika, 42, 69–81.
- Stemler, S. E. & Naples, A. (2021). Rasch measurement v. item response theory: knowing when to cross the line. Practical Assessment, Research, and Evaluation, 26, Article 11.