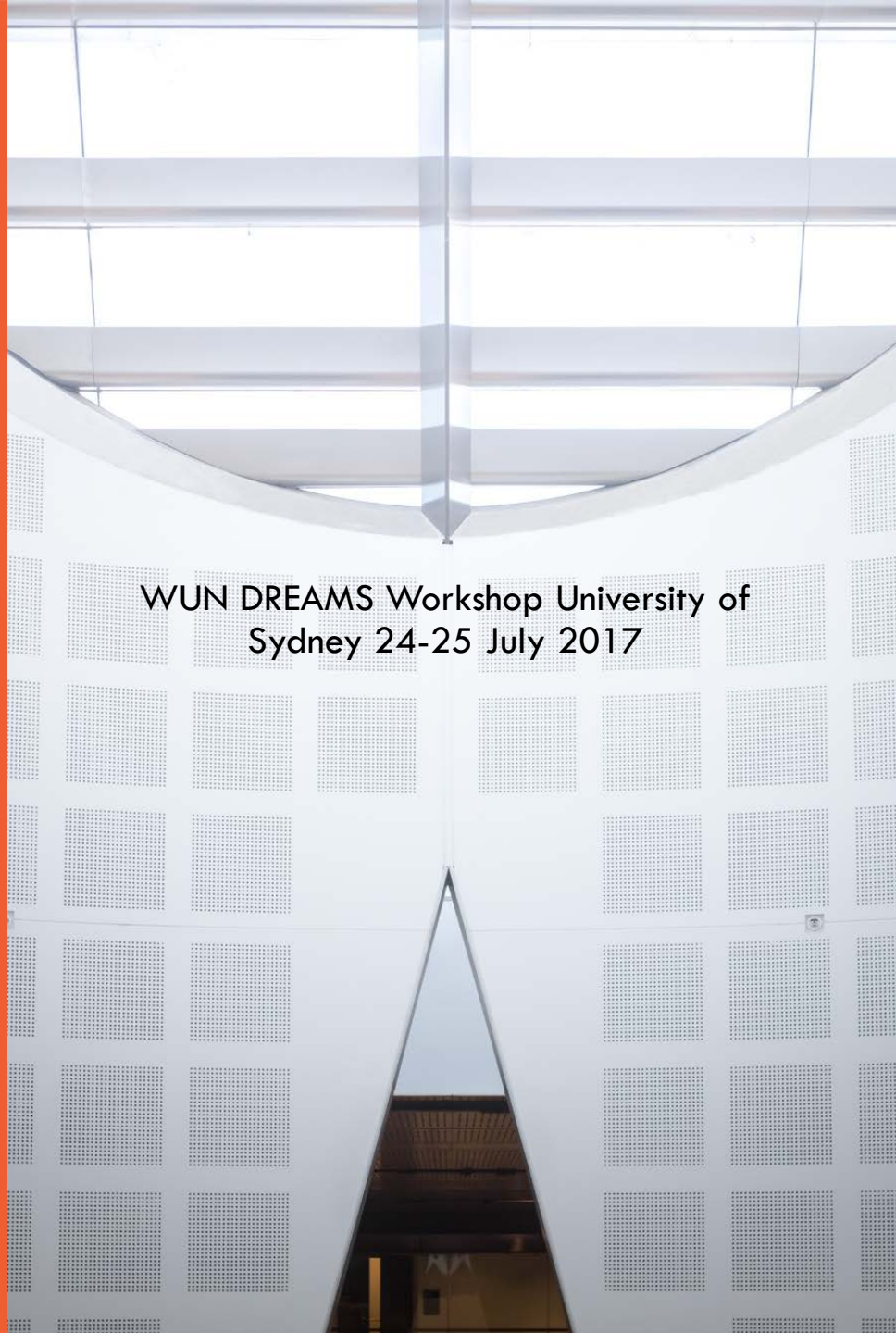
Modelling of SJTs for medical school selection and measuring in course professional development

Associate Professor Deborah O'Mara
and Imogene Rothnie
Sydney Medical School

THE UNIVERSITY OF SYDNEY



Overview

- SJT development for selection
- SJT application for in-course professional and personal assessment
- The scoring challenge before evaluation with Rasch measurement theory
- Empirical results of using the polytomous Rasch model to evaluate an SJT domain.
- The scoring and review challenge after evaluation with Rasch measurement theory.

SJT development for a selection trial in 2016

1. Test specification – 4 target attributes selected
   - Communication
   - Ethical
   - Professionalism
   - Personal and intellectual autonomy
2. Item writing
   - 6th May workshop with 8 medicine and 2 dentistry subject matter experts
   - 54 scenarios written with 6-9 options
3. Item review
   - 20th May workshop with 7 medicine and 1 dentistry subject matter experts
4. Concordance
   - June 2016 conducted electronically with 13 subject matter experts 12 medicine and 1 dentistry (30 invited)
   - 49 scenarios tested for up to 8 options and 18 not piloted due to a lack of concordance

Example of rating style SJT with 5 point appropriateness likert scale

A student, Mia, completed a dental procedure on her patient earlier today, where she received very positive feedback from her clinical tutor. Mia therefore decided to showcase her patient's x-rays, along with the positive feedback she received from her clinical tutor, on the student social media page. All current dental students at the University can access the page. The x-rays include the patient's name. Jaiswal is one of Mia's fellow students, and sees her social media post.

Response Instructions: How appropriate are each of the following responses by Jaiswal in this situation? (1 Very appropriate to 5 Very Inappropriate)

S5	Response	Key	Target Attribute (not seen)
Option A	Inform Mia that she should not have shared the patient's x-ray on social media	1	**Ethical**
Option B	Suggest to Mia that she should have obtained consent from the patient before posting the x-ray on social media	5	**Ethical**
Option C	Advise Mia to remove the social media post	1	**Ethical**
Option D	Congratulate Mia on her positive feedback	3	**Prof**
Option E	Inform his clinical tutor of Mia's social media post	2	**Ethical**
Option F	Report the post as inappropriate through the social media website	**3**	**Ethical**

SJT administration

- 2 Versions of the SJT test each with 20 scenarios was developed
- Each scenario had 6 options giving 120 questions in all
- Version A question target attributes

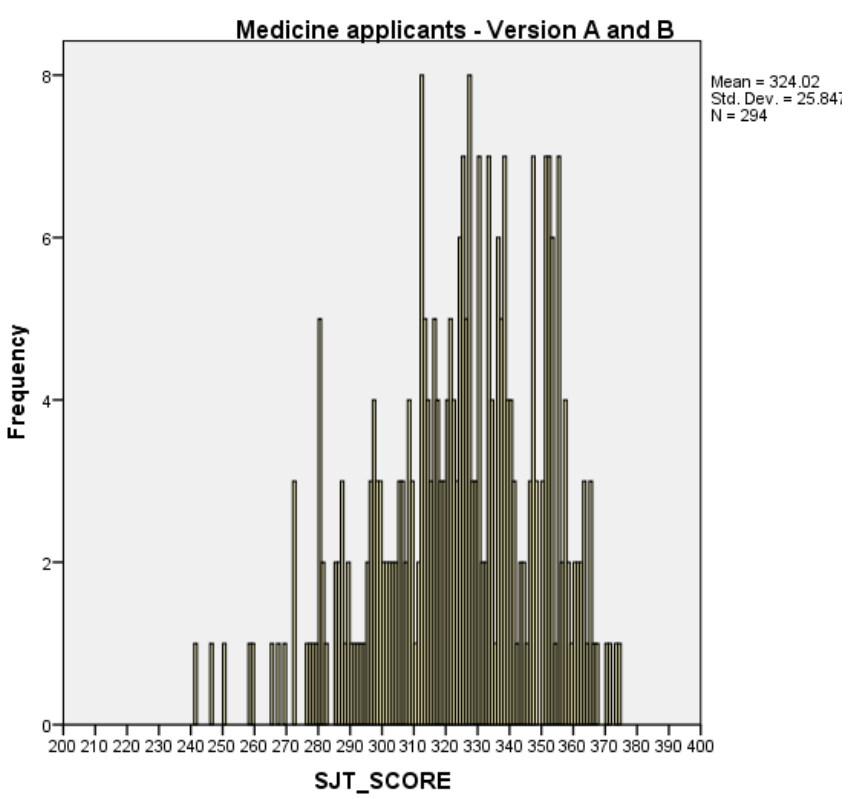
Communication	14
Ethical	50
Personal and intellectual autonomy	44
Professionalism	12
Total	120

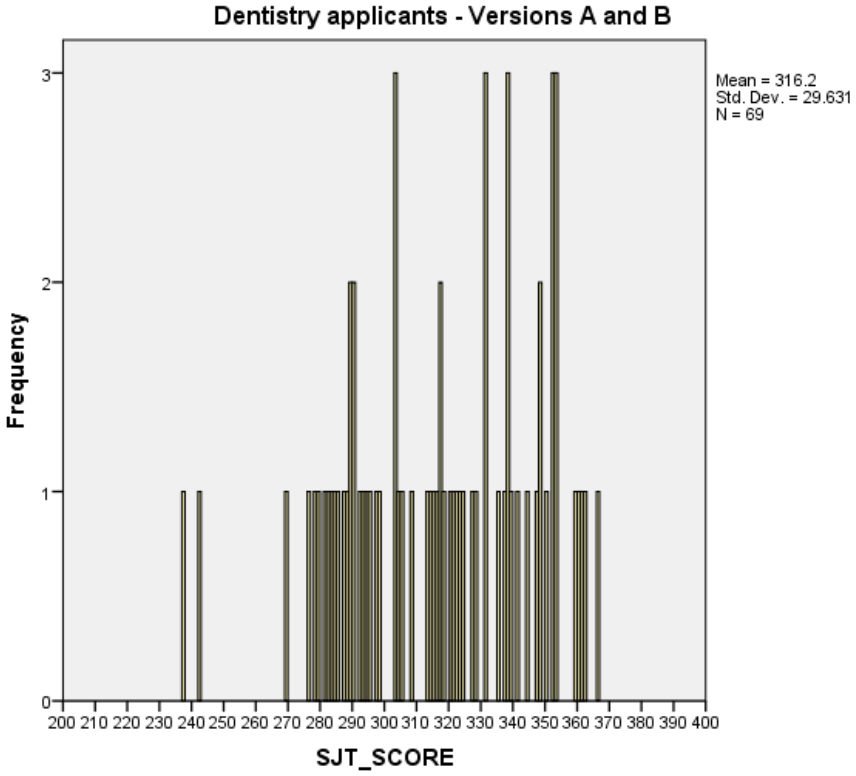
- Version B question target attributes

Communication	15
Ethical	53
Personal and intellectual autonomy	40
Professionalism	12
Total	120

- 40 minutes was allowed to complete the SJT with 5 minutes additional time for a short demographic and feedback survey
- Limesurvey was used to deliver the SJT.

SJT scores


Medicine applicants - Version A and B
Mean = 324.02
Std. Dev. = 25.847
N = 294
Frequency
0
2
4
6
8
200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400
SJT_SCORE


Dentistry applicants - Versions A and B
Mean = 316.2
Std. Dev. = 29.631
N = 69
Frequency
0
1
2
3
200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400
SJT_SCORE

SJT sub-group analyses

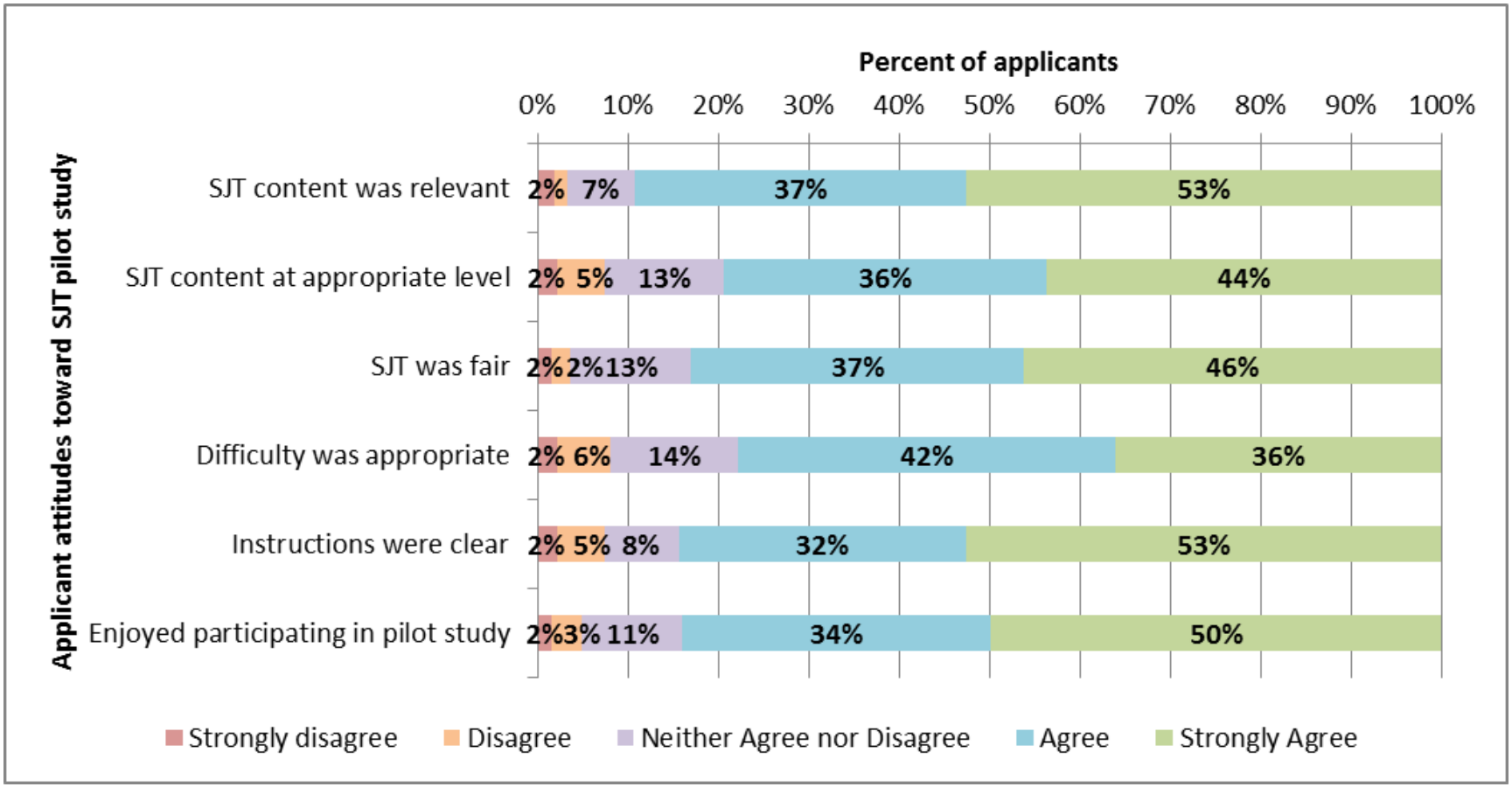
SIGNIFICANT DIFFERENCES

- Both parents born in Australia scored higher in Versions B
- Neither parent had a degree scored higher in Version A
- MD only applicants had higher scores only for version B

NON-SIGNIFICANT DIFFERENCES

- Gender
- English first language
- Country of birth Australia or not
- High school type
- Prior degree type
- No correlation with age

Acceptability of the SJT



SJT voluntary sample has possible bias

- Compared the demographics of the MMI evaluation survey for
  - 184 who did not wish to participate in the SJT and did an MMI
  - 142 who did the SJT and MMI
  - 114 who registered their interest in the SJT but did not complete it

Applicants more likely to volunteer and completed the SJT were

  - Female
  - Less likely to be born in Australia
  - English not first language Neither parent has a degree

- Applicants who did not participate in the SJT more often were
  - From Melbourne or completed their first degree there (40%)
  - Parent is a doctor (17%) or dentist (2%) net 19%

2017 extension of the SJT Project

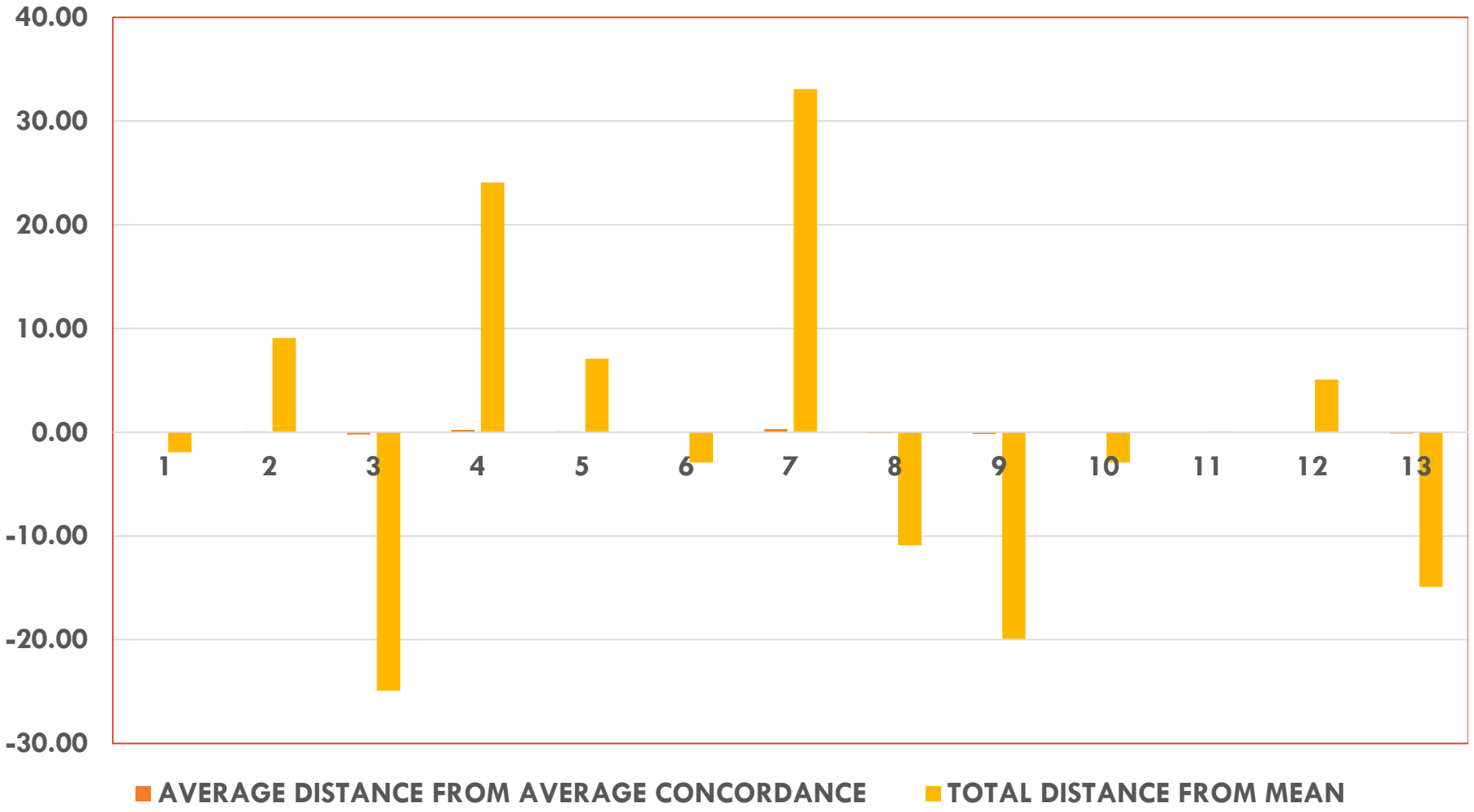
- Item analysis conducted by the WPG was used to identify the best 20 items from each Version of the 2016 selection SJT
- Best 5/6 options for the best scenarios were selected based on item analysis
- Scenarios that were too similar were dropped
- Resulting SJT test had 22 scenarios with 5 options each

Domain	Number of Questions
Communication	14
Ethical	39
Personal Autonomy and Integrity	45
Professionalism	12
Total Number of SJTs	**110**

Concordance Sample

- Small base of 13
- 3-4 “rogue” assessors
- Original concordance with 49 scenarios
  - alpha=0.949 for 334 items
  - average intra-class correlation =0.834
- 2017 22 scenarios sub-set concordance
  - alpha=0.833 for 98 items
  - average intra-class correlation =0.055
  - 12 items with zero variance excluded

13 SJT Concordance Rater by average and total distance from average concordance over 110 items



2017 application of SJTs in the SMP

- All Stage 1 and Stage 2 students were required to do the 110 item SJT test as part of the Personal and Professional Development Theme in late March 2017.
- This provided a sample base of 618 students.
- Will correlate the SJT score with clinical assessment performance later in the year.
- Options for scoring and modelling were explored using IRT.

The Rasch Model

- Probability of correct response is a function of candidate ability – item difficulty.
- Measures of candidate ability and item difficulty (parameters) are estimated using total scores (sufficient statistics) and logarithmic transformation.
- Iterative process.
- Results in parameter separation ie:
- When relationships between observed data meet the requirements of the Rasch model there is:
  - *local objectivity/measurement invariance*
    - The distance between 2 items in terms of difficulty is **independent** of the sample of persons who attempted them.
    - The distance between to people in terms of proficiency is **independent** of the sample of items they attempted.

Outputs of Rasch Analysis

- Logits: the unit of measurement used in RMT. The $\log_e$ transformation of the probability of a correct response.
- Element measures are presented on an additive, linear scale.
- 0 is specified as the *mean* value (ie negative and positive values)
- Conjoint : ability and difficulty are compared on the same scale.
- Measures of fit, precision and reliability.
- A calibrated instrument for the measurement of a unidimensional construct (one attribute at a time).

The Polytomous Rasch Model (PRM): Rating Scale and Partial Credit Model Parameterisations

- Rating scale (Andrich, 1978), partial credit (Masters, 1982)
- Continuum of latent construct may be divided into contiguous categories, at threshold locations.
- Dichotomous (simple) model – one threshold, Polytomous- 2,3,etc
- PRM : an empirical test of hypothesis that increasing scores/higher ratings on an item represent more of construct.
- Rating scale: equal number categories, uniform scale format.
- Partial credit: unique number categories, unique categories.
- No assumptions are made about the distance between categories.

Outputs of PRM

- Overall and individual item fit. Construct validity
- Threshold person-targeting "map", person separation index (reliability). Standard setting
- Threshold probability curves.
- Category probability curves.
  - An evaluation of how well categories are working in each scale/partial credit item. Item and scoring review
- Item and Threshold locations. Item banking, test construction

Scoring and evaluating the SJT with the PRM

- SJT assumed to be multidimensional
- Restricted analysis to one "domain" – Personal Integrity and Autonomy.
- 41 items, 618 candidates.
- Instructions to consider each item independently.
- Fractional scoring from concordance method cannot be directly entered into Rasch software (RUMM 2030, UWA 2010).
  - Need zero
  - Consecutive integer scoring
  - Part of the research question?

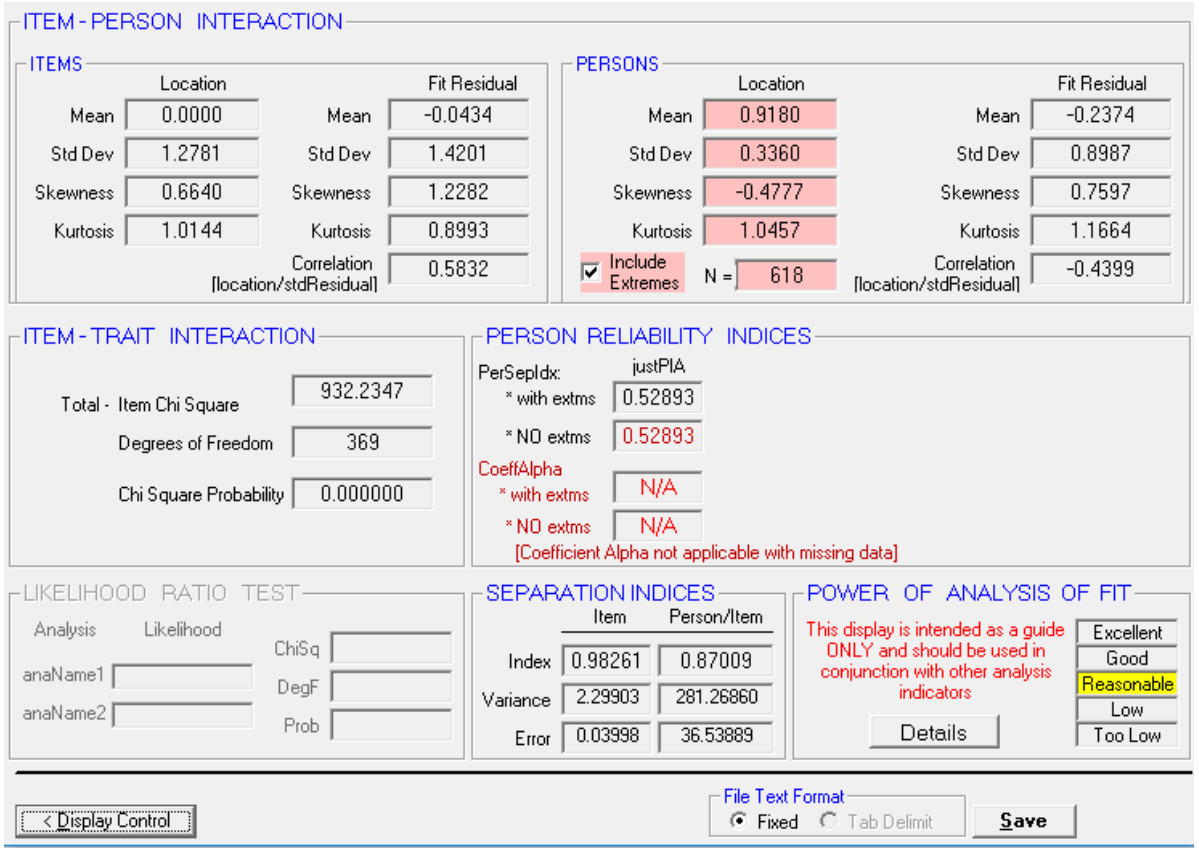
Rescoring concordance values to integer scale (and I can guess what you're thinking)

	Original Concordance Value						Re-scored Integer Scale				
Response	*1*	*2*	*3*	*4*	*5*		*1*	*2*	*3*	*4*	*5*
S1A_1	0.000	0.154	0.538	0.231	0.077		0	2	4	3	1
S1A_2	0.000	0.000	0.000	0.462	0.538		0	0	0	1	2
S1A_3	0.000	0.000	0.077	0.538	0.385		0	0	1	3	2
S1A_4	0.000	0.077	0.077	0.615	0.231		0	1	1	3	2
S1A_5	0.923	0.077	0.000	0.000	0.000		2	1	0	0	0

1= Very appropriate
2= Slightly appropriate
3= Neither appropriate or inappropriate
4= Slightly inappropriate
5= Very inappropriate.

RESULTS: Overall Model Fit

– Summary Statistics: Distribution of standardised residuals



RESULTS: Unidimensionality

Principal components of residuals

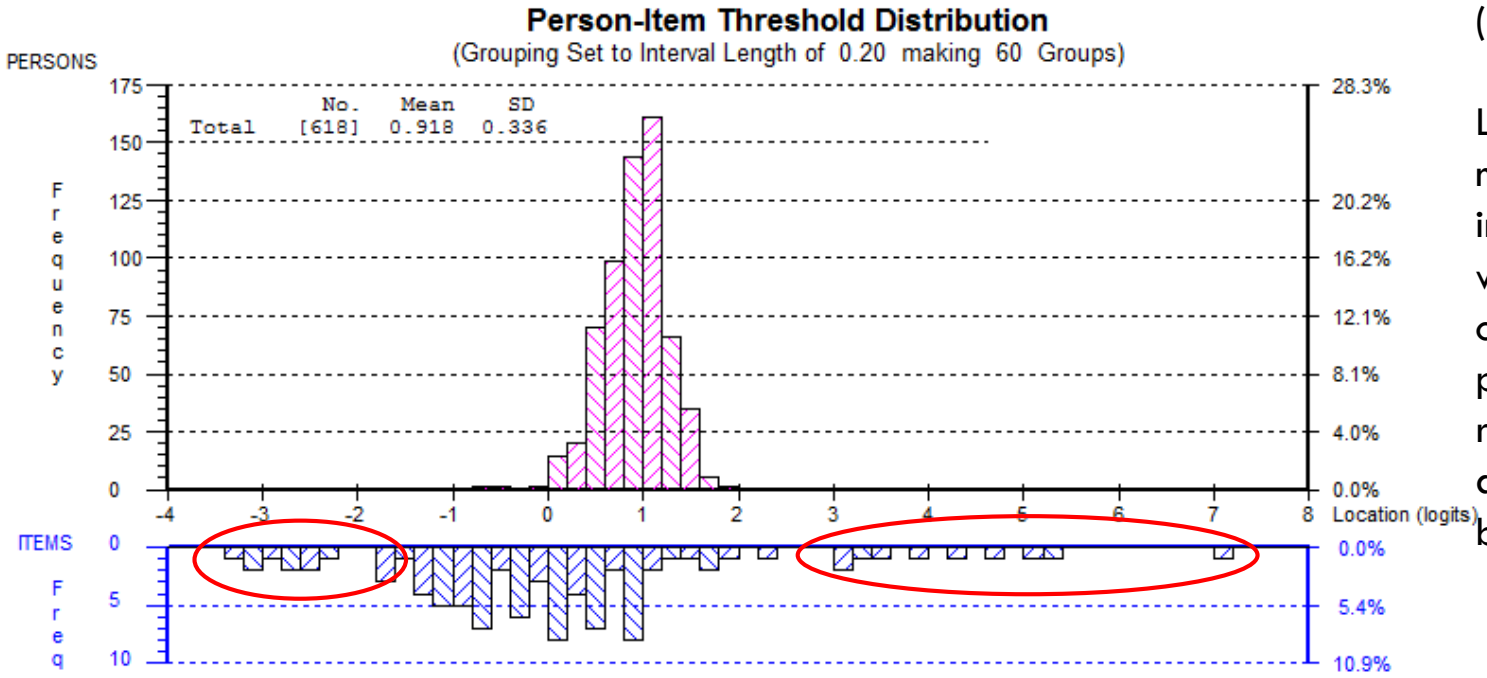
Display: PRINCIPAL COMPONENT SUMMARY

PC	Eigen	Percent	CPercent	StdErr
PC001	7.166	6.51%	6.51%	1.011
PC002	3.920	3.56%	10.08%	0.552
PC003	2.675	2.43%	12.51%	0.375

Item	PC1
S7B_3	0.56
S18B5	0.47
S17B5	0.45

- Small but possibly significant first contrast
- Some off-target items

Targeting. Person Separation.



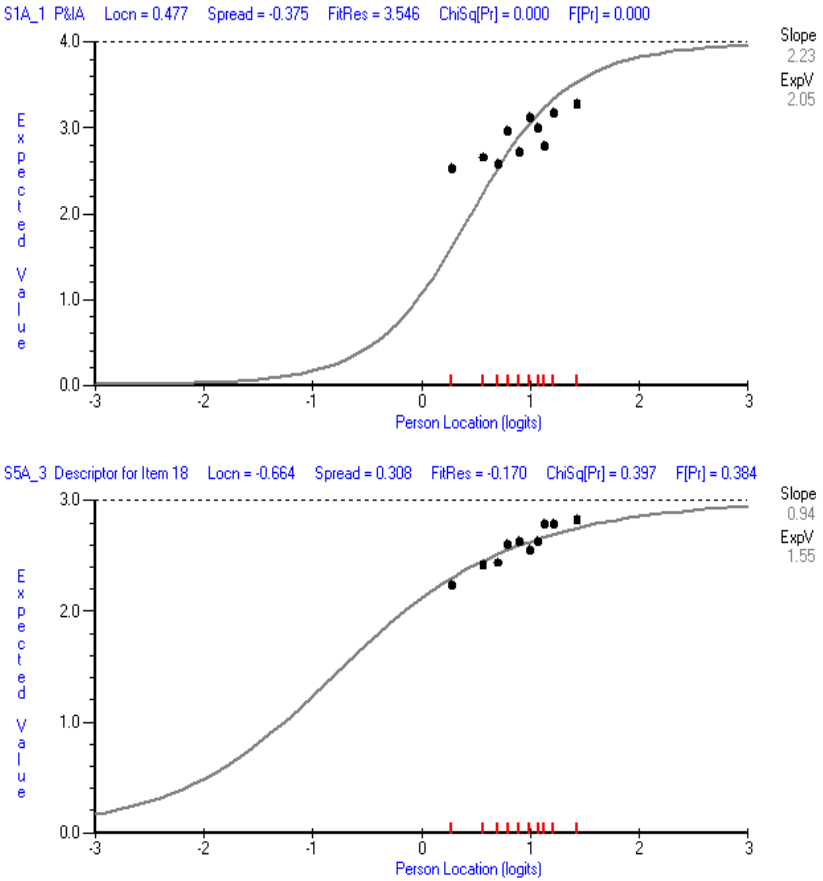
Person Separation Index (P.S.I) = 0.53

Low. Not much test information where most of the proficiency range appears to be.

Item Fit and difficulty locations

Item	Location	SE	Fit Resid	DF	ChiSq	Prob	F-stat	Prob	UnCThr 1	UnCThr 2	UnCThr 3	UnCThr 4
S1A_1	0.477	0.032	3.546	599.71	46.782	0.000	5.044	0.000	3.15321	-2.54549	2.24530	-0.94307
S7B_2	2.086	0.056	3.833	599.71	27.535	0.001	3.095	0.001	7.19672	-2.49898	1.56171	
S7B_3	1.64	0.037	3.201	594.85	121.381	0.000	14.076	0.000	3.21473	-0.7116	5.19506	-1.13638
S15B2	0.523	0.037	2.944	600.68	48.477	0.000	5.22	0.000	0.55215	1.25283	-0.2364	
S17B4	-0.867	0.073	-0.797	597.76	31.623	0.000	3.871	0.000	-1.21298	-1.23972	-0.14795	
S16A2	-1.293	0.083	-1.271	599.71	32.363	0.000	3.957	0.000	-2.73126	0.14490		
S15A5	-0.826	0.085	-1.49	600.68	43.095	0.000	5.846	0.000	-0.93540	-0.71661		
S7A_3	-1.096	0.124	-1.839	600.68	34.088	0.000	4.719	0.000	-1.09588			
S16A3	-1.723	0.095	-1.968	600.68	45.176	0.000	6.111	0.000	-3.03763	-0.4088		
S8B_1	0.908	0.051	0.238	599.71	9.516	0.391	1.026	0.417	0.85263	0.96273		
S3B_5	-0.309	0.068	0.168	600.68	11.216	0.261	1.251	0.261	-0.62306	0.00532		
S15A4	-0.313	0.051	0.006	599.71	10.161	0.338	1.115	0.350	-1.78082	0.31052	0.5298	
S5A_3	-0.664	0.066	-0.17	600.68	9.447	0.397	1.069	0.384	-1.10985	-1.00413	0.12214	
S3B_3	0.557	0.038	1.98	599.71	21.443	0.011	2.281	0.016	0.88000	0.46801	0.32345	

Item Characteristic Curve (ICC) Individual item fit

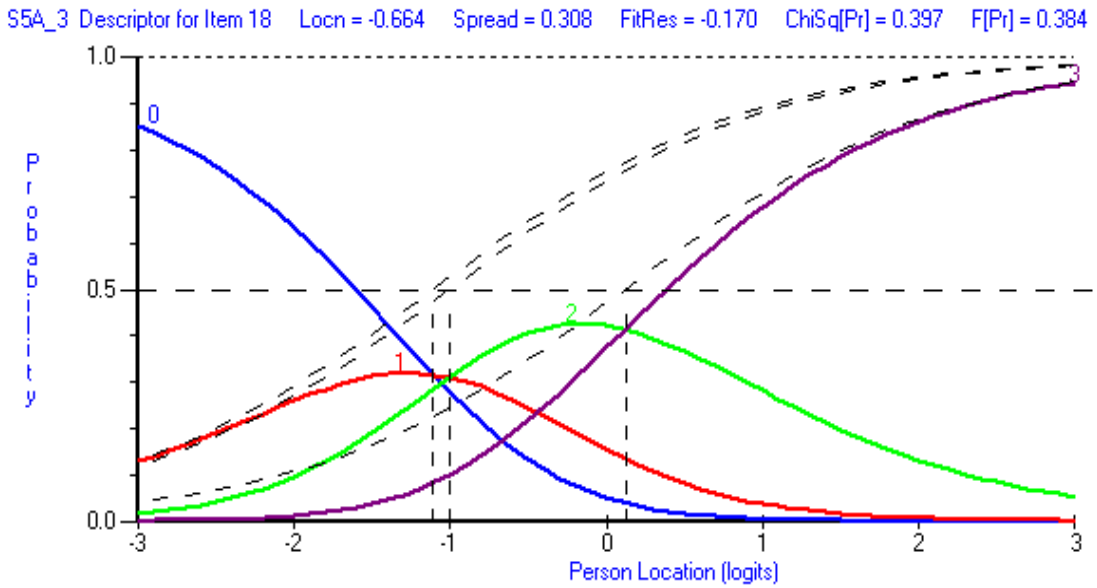


Poor fit, "under" discriminating.
More randomness than model predicts
Observed ('dots') do not follow expected 'line'.
Low proficiency – greater than expected value
High proficiency – less than expected value

Reasonable fit.
Observed values follow expected (theoretical).

(S5A_3) Thank … for pointing out the areas that he needs to develop

Response Code	1 Very appropriate	2 Slightly appropriate	3 Neither appropriate nor inappropriate	4 Slightly inappropriate	5 Very inappropriate
Concordance value	0.615	0.308	0.077	0.000	0.000
Rasch scale rescore	3	2	1	0	0
Student response %	66	28	5	1	0



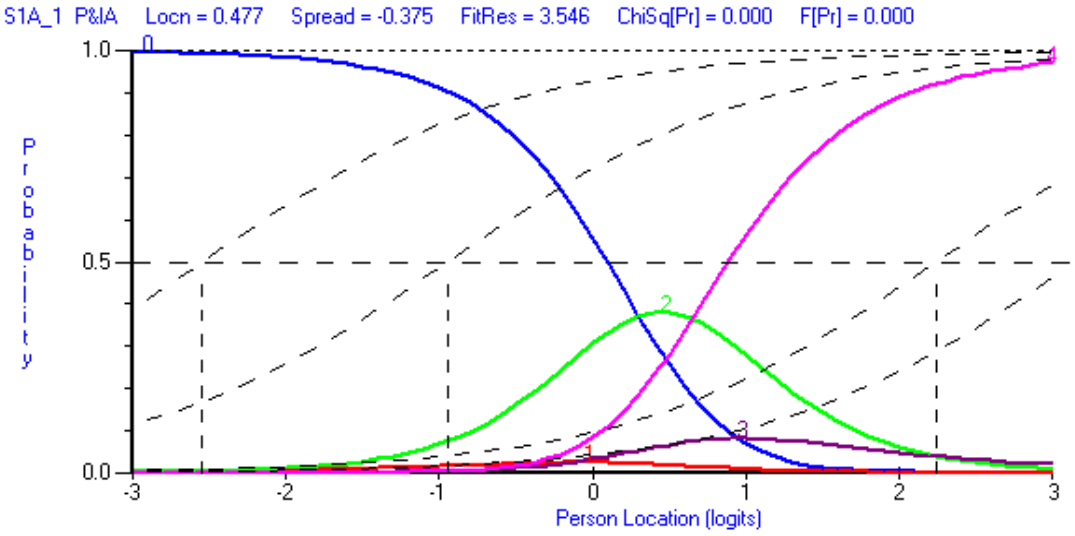
Fit OK, but:
Steep slope ICC due to close thresholds (1, 2).

Category 1, almost never max probability.

Collapse categories?

Poor fit - S1A_1. Take some vitamins

Response Code	1 Very appropriate	2 Slightly appropriate	3 Neither appropriate nor inappropriate	4 Slightly inappropriate	5 Very inappropriate
Concordance value	0.000	0.154	0.538	0.231	0.077
Rasch scale rescore	0	2	4	3	1
Student response %	9	32	50	8	1



Poor fit, disordered thresholds.
Loads strongly on the "non Rasch" contrast.

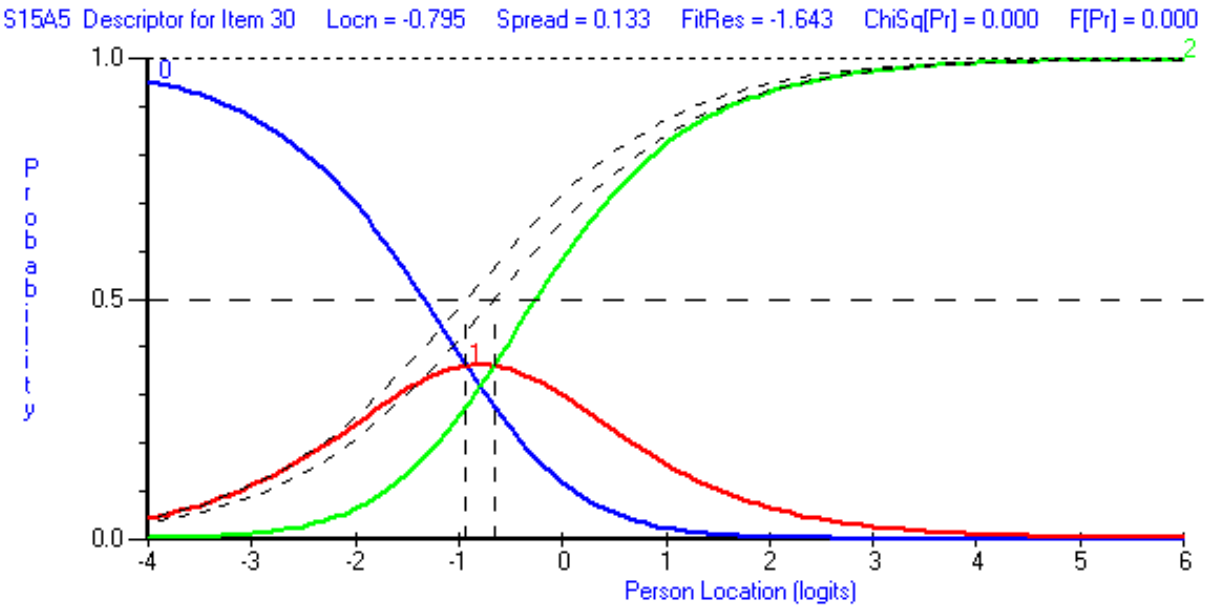
0-1 appears HIGH

1-2 appears LOW

What's the real question? Is there more than one? (vitamins per se, vitamins here)

Ideal responses according to concordance panel

Response Code	1 Very appropriate	2 Slightly appropriate	3 Neither appropriate nor inappropriate	4 Slightly inappropriate	5 Very inappropriate
Concordance value	0.923	0.077	0.000	0.000	0.000
Rasch scale rescore	2	1	0	0	0
Student response %	81	15	3	0	0



Close thresholds
Steep ICC

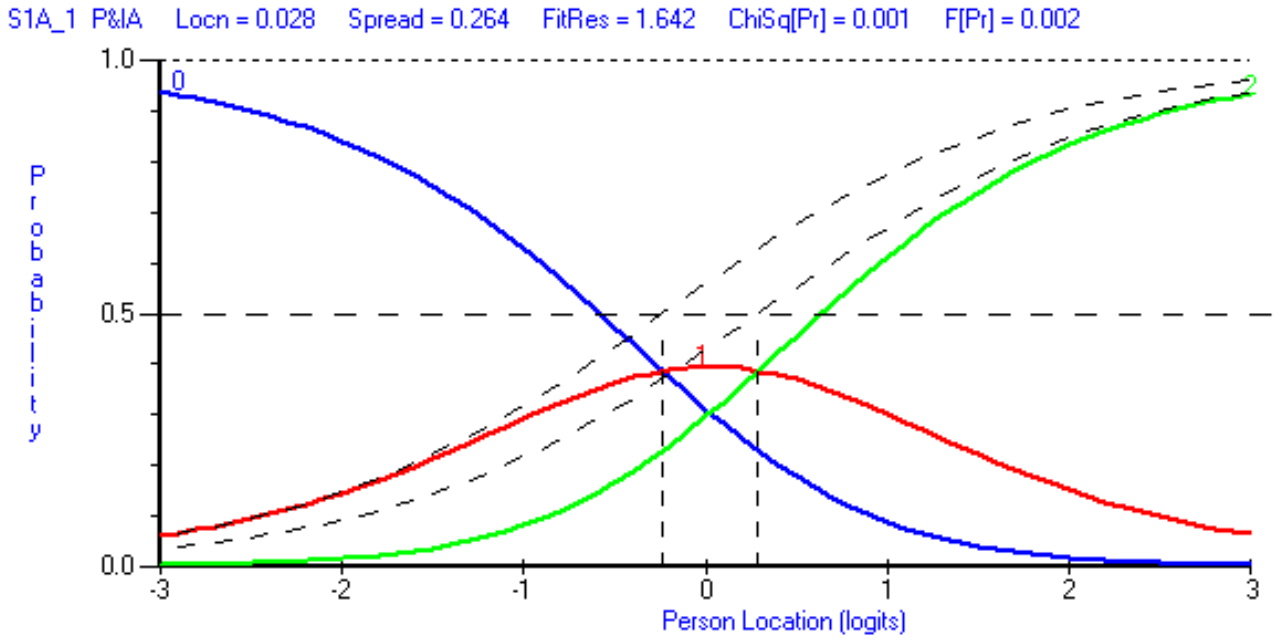
Graphical representation of threshold maps.

SOLUTIONS?

- Principal components: Create subtests?
- Re-score items – collapse categories – fix disordering?
- Item fit: Delete worst of the mis-fitting items.
- Back to test authors/content experts.
- Look again.
- If calibration is the purpose – bank stable for anchoring evaluation.

Rescored disordered thresholds (vitamins)

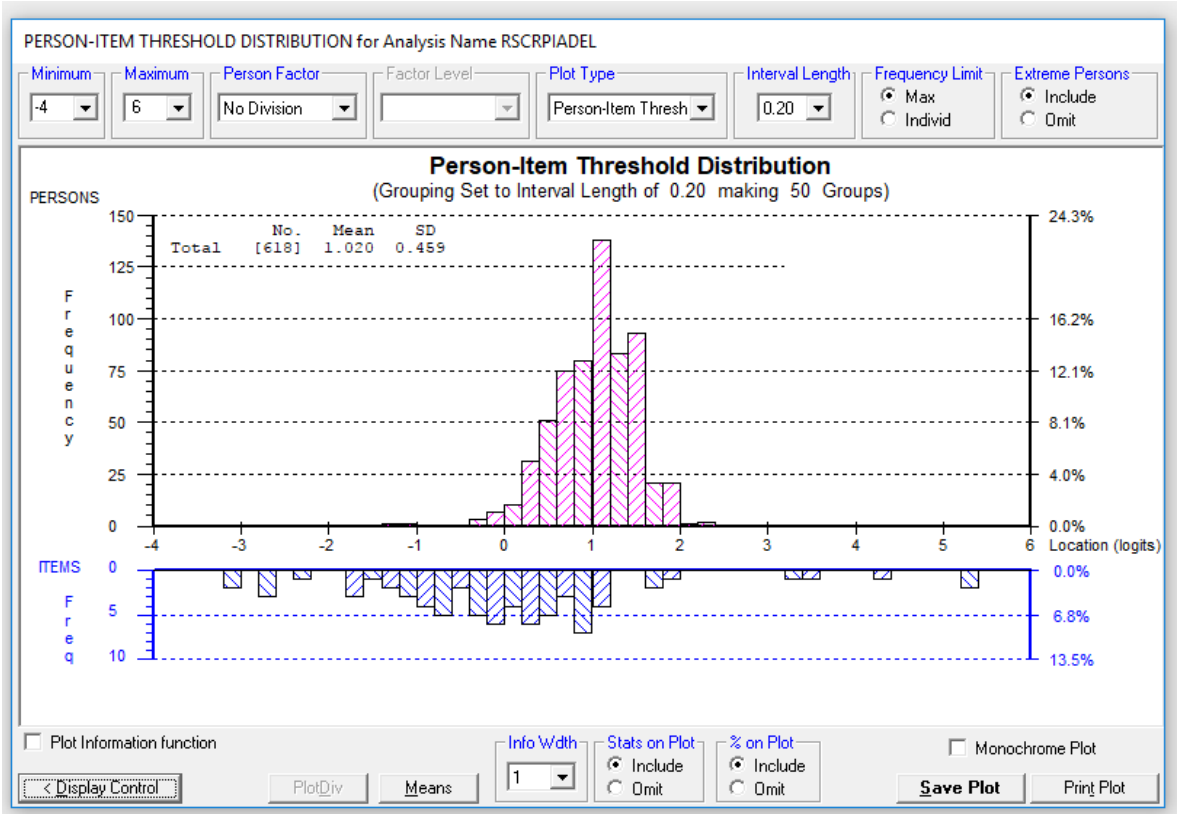
Response Code	1 Very appropriate	2 Slightly appropriate	3 Neither appropriate nor inappropriate	4 Slightly inappropriate	5 Very inappropriate
Concordance value	0.000	0.154	0.538	0.231	0.077
Rasch scale rescore	0	1	2	2	0
Student response %	9	32	50	8	1



Attempt to keep information (not dichotomise) .

My relatively arbitrary distinction between "slightly" options

Re-scored selection, deleted worst fit



P.S.I ↑ 0.63.

Greater spread of person distribution

Take home messages

- Nothing beats a good theory.
- To show less and more , a path between them needs describing.
- Partial agreement may not be a point on the path.
- Back to the theory for discussion and treatment
  - Greater specificity of proficiency hierarchy on a domain a priori (if possible??)
  - Psychometrics makes good diagnostic tools
- Small concordance panels include too much error

References

- Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561-574
- Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174

Image reproduced from Eckes, T (2015) Introduction to Many-Facet Rasch Measurement : Analyzing and Evaluating Rater-Mediated Assessments- 2nd Edition.