Assessing the psychometric properties of a questionnaire

Christophe Lalanne

Fall 2011

Outline

Patients reported outcomes

Dimensionality of a scale

Item analysis

Rasch and 2-PL models

IRT model for polytomous items

Differential item functioning

"That the model is not true is certainly correct, no models are–not even the Newtonian laws. (...) Models should not be true, but it is important that they are applicable." —Rasch, 1960

Patient-reported outcomes

" Any outcome based on a patient's perception of a disease and its treatment(s) scored by the patient himself is called a Patient-Reported Outcome (PRO). PROs are a large set of patient-assessed measures ranging from single item (e.g., pain VAS, overall treatment evaluation, and clinical global improvement) to multi-item tools."

— EMA (2005)

" Any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else."

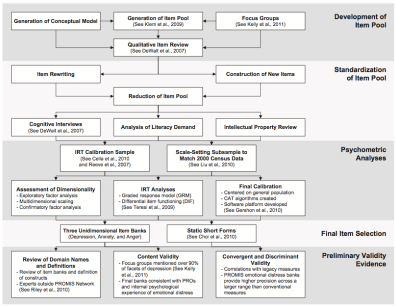
— FDA (2009)

A case study

Data comes from a large-scale US study that aims to build a calibrated item bank on PROs. In this example, we'll be using a questionnaire composed of 29 Likert-type items ($1 =$'Never', $2 =$'Rarely', $3 =$'Sometimes', $4 =$'Often', and $5 =$'Always') on anxiety administered to $N = 766$ individuals sampled from the general population (Pilkonis et al., 2011 and Choi et al., 2011).

PROMIS Cooperative Group. Unpublished Manual for the Patient-Reported Outcomes Measurement Information System (PROMIS) Version 1.1. October, 2008: http://www.nihpromis.org

The PROMIS methodology (Pilkonis et al., 2011)

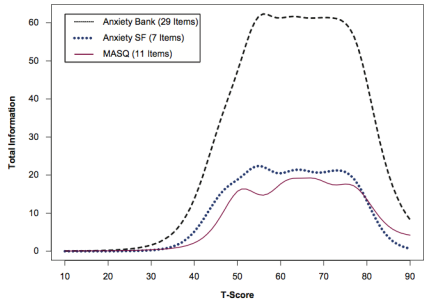


The Anxiety scale

Following a review of more than 140 existing instruments, scale reduction and short form validation were done with CFA and IRT.

1. I felt fearful
2. I felt frightened
3. It scared me when I felt nervous
4. I felt anxious
5. I felt like I needed help for my anxiety
6. I was concerned about my mental health
7. I felt upset
8. I had a racing or pounding heart
9. I was anxious if my normal routine was disturbed
10. I had sudden feelings of panic
11. I was easily startled
12. I had trouble paying attention
13. I avoided public places or activities
14. I felt fidgety
15. I felt something awful would happen
16. I felt worried
17. I felt terrified
18. I worried about other people's reactions to me
19. I found it hard to focus on anything other than my anxiety
20. My worries overwhelmed me
21. I had twitching or trembling muscles
22. I felt nervous
23. I felt indecisive
24. Many situations made me worry
25. I had difficulty sleeping
26. I had trouble relaxing
27. I felt uneasy
28. I felt tense
29. I had difficulty calming down

Test information curves for anxiety (Pilkonis et al., 2011)

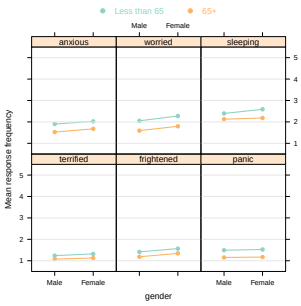


Patients demographic data

We only have data on participants' age, gender and education. They are summaried in the next table. These are important covariates that might be used in descriptive or explanatory models. Ethnicity would also be of interest.

		N	Female	Male
age	Less than 65	555	304	251
	65+	211	93	118
education	College or higher	596	310	286
	High school or lower	170	87	83
Overall		766	397	369

Frequency of responses by gender and age

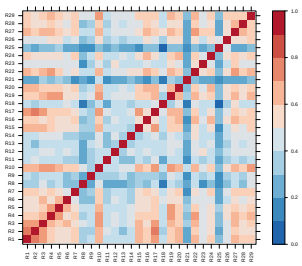


Assesing dimensionality

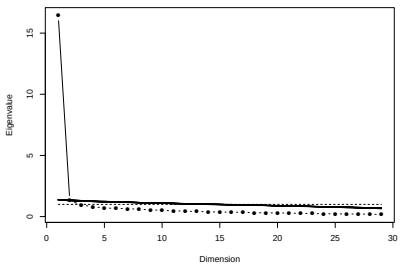
Factor Analysis can be used to study the dimensionality of the scale. Basically, the idea is to check whether the assumption of common factor(s) (running through all items) is reasonable enough so that we can aggregate individual responses to different items, i.e. map all individual scores on a common metric (factor scores).

FA will output numerical value like loadings, communality and uniquenesses which reflect the weight of any single item on the dimension, how much of the variability is due to the common factor, and the magnitude of the error term. Although the scale has already been validated, we will use exploratory FA (as if we didn't know how many factors we should look for).

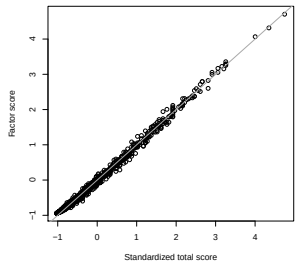
Interitem (polychoric) correlations



Scree plot of eigenvalues with parallel analysis



Factor vs. raw scores



Internal consistency

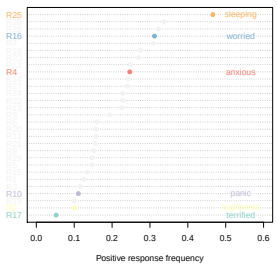
The Cronbach's alpha is a sample-dependent index used to ascertain a lower-bound of the reliability of an instrument. It is no more than an indicator of variance shared by all items considered in the computation of a scale score. The following assumptions are made: (a) no residual correlations, (b) items have identical loadings, and (c) the scale is unidimensional.
Here, on the whole set of items, it amounts to 0.971, with 95% CI $[0.967; 0.975]$ (BCA).

> " Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement. (Cronbach and Shavelson, 2004)"

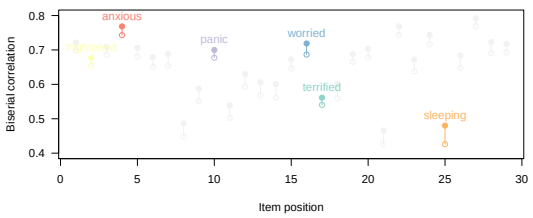
Modeling responses to dichotomous items

In what follows, we will restrict our analysis to binary-scored items (1/2=0, 3–5=1) and apply the Rasch model. This will allow to estimate item severity which will be used to locate each item along the construct's latent continuum. In this case, we will consider that each item equally discriminate among individuals.

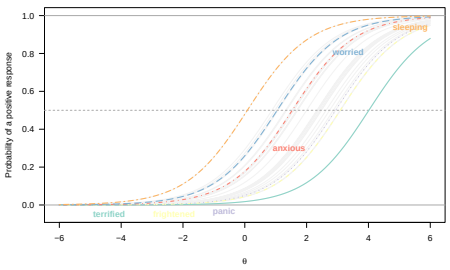
Frequency of responses not 'rarely' or 'never'



Correlation of items with total score



Items parameters with the Rasch model

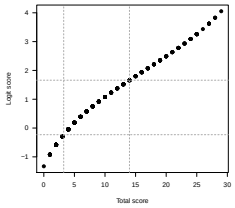


From raw scores to factor scores

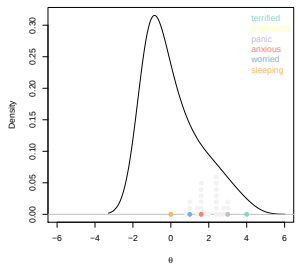
There were 446 patterns of responses observed, yielding 30 different total scores. As we know that the sum score is a <u>sufficient statistic</u> for the Rasch model, any pattern of responses having the same total will get the same factor score.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	Total	Logit
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1.3279079
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	-0.9210140
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	-0.9210140
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	-0.9210140
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2	-0.5829504
442	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	27	3.6233943
443	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	28	3.8282668
444	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	27	3.6233943
445	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	28	3.8282668
446	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	29	4.0537294

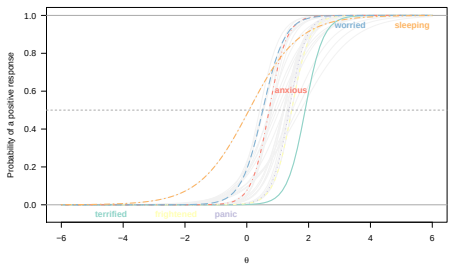
Another way to show the mapping between raw and factor scores is to graph the relationship between sum scores and factor scores (on the logit scale), as shown below:



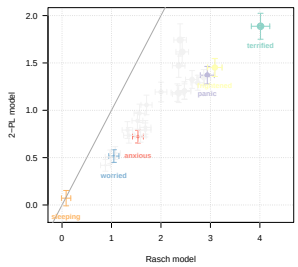
Item-person map



Items parameters with varying discrimination



Comparison of the 1- and 2-PL models

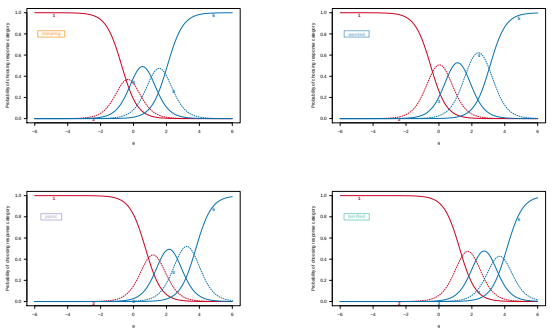


Which model perform best?

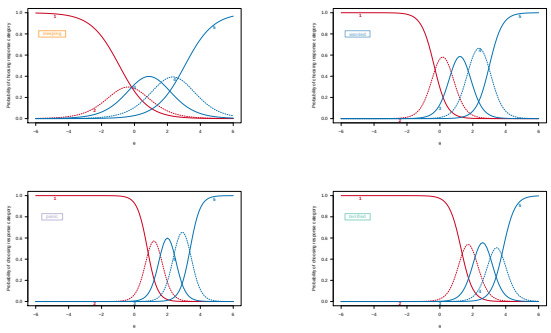
Of course, it seems that allowing discrimination to vary between items yields better results (from a statistical perspective). However, none of the above models are really satisfactory as they do not fully account for item response format.

In fact, what we need is a model that would allow to work at the level of response categories. Several models have been proposed to deal with polytomous items. We will use the Graded Response Model (Samejima, 1969).

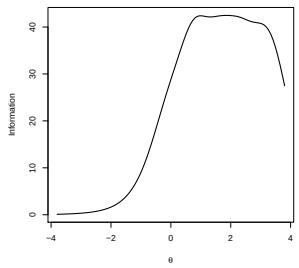
Item threshold parameters (GRM cons.)



Item threshold parameters (GRM uncons.)



Test information curve (GRM uncons.)



What's next?

Until now, we have concentrated on the mapping between individual raw responses and a standardized scale, reflecting an hypothesized construct of anxiety, that allows to locate individuals and items.

We might ask whether person parameters depend on external covariates, like gender or age. The establishment of measurement invariance across groups is a logical prerequisite to conducting substantive cross-group comparisons (Vandenberg and Lance, 2000).

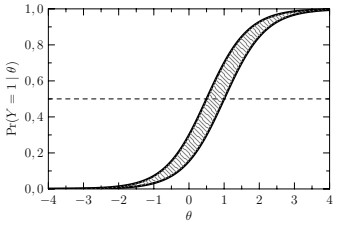
Differential item functioning

Differential Item Functioning (DIF) is said to occur when the probability of endorsing a particular item differs according to a subject-specific covariate (e.g., age, gender, country), holding subject trait constant. It has been studied in many different areas:

- Psychiatric research (Crane et al., 2007): gender biases for 'I feel sad' and 'Able to enjoy life'.
- Personality assessment (Kulas et al., 2008): age and gender-effect on the NEO-PI questionnaire.
- Health-related Quality of Life (Petersen et al., 2003): country biases for 'Did you worry?' or 'Did you feel depressed?'.

Illustration

People with the same level on the latent trait (e.g. moderate level of anxiety) have a different probability of endorsing the item depending on their group membership (e.g. gender).



Bibliography

1 Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

2 Pilkonis, P., Choi, S., Reise, S., Stover, A. and Riley, W. et al. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283. PMID: 21697139.

3 Choi, S., Gibbons, L. and Crane, P. (2011). `lordif`: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/Item Response Theory and monte carlo simulations. *Journal of Statistical Software, 39*(8).

4 Cronbach, L. and Shavelson, R. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418.

5 Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

6 Vandenberg, R. and Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

7 Crane, P., Cetin, K., Cook, K., Johnson, K. and Deyo, R. et al. (2007). Differential item functioning impact in a modified version of the Roland-Morris Disability Questionnaire. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 16*(6), 981–990.

8 Kulas, J., Merriam, J. and Onama, Y. (2008). Item-trait association, scale multidimensionality, and differential item functioning identification in personality assessment. *Journal of Research in Personality, 42*(4), 1102–1108.

9 Petersen, M., Groenvold, M., Bjorner, J., Aaronson, N. and Conroy, T. et al. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 12*(4), 373–385.

Made with ConTEXt version 2011.05.18 18:04 (pdfTEX version 140),
`questionnaires.tex`, 2038093 on 2011/10/20