Assessing the psychometric properties of a questionnaire

Christophe Lalanne

Fall 2011
“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).

The PROMIS methodology
(Pilkonis et al., 2011)

[Diagram of the PROMIS methodology process]

- Generation of Conceptual Model
  - Generation of Item Pool
    - Focus Groups
  - Qualitative Item Review
  - Item Rewriting
  - Cognitive Interviews
  - Analysis of Literacy Demand
  - Intellectual Property Review

- Development of Item Pool
  - Construction of New Items
  - Reduction of Item Pool

- Standardization of Item Pool
  - IRT Calibration Sample
    - Scale-Setting Subsample to Match 2000 Census Data
      - Final Calibration
      - Three Unidimensional Item Banks
      - IRT Analyses
      - Assessment of Dimensionality
      - Scale-Setting Subsample to Match 2000 Census Data
      - Final Calibration
      - Three Unidimensional Item Banks

- Psychometric Analyses
  - Content Validity
    - Convergent and Discriminant Validity
      - Final Item Selection
      - Preliminary Validity Evidence

- Review of Domain Names and Definitions
  - Final item banks and definitions of constructs
  - Experts outside PROMIS Network
  - Convergent and Discriminant Validity
  - Final item banks and definitions of constructs
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 summarised in the next table. These are important covariates that might be used in descriptive or explanatory models. Ethnicity would also be of interest.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 65</td>
<td>555</td>
<td>304</td>
<td>251</td>
</tr>
<tr>
<td>65+</td>
<td>211</td>
<td>93</td>
<td>118</td>
</tr>
<tr>
<td>education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>College or higher</td>
<td>596</td>
<td>310</td>
<td>286</td>
</tr>
<tr>
<td>High school or lower</td>
<td>170</td>
<td>87</td>
<td>83</td>
</tr>
<tr>
<td>Overall</td>
<td>766</td>
<td>397</td>
<td>369</td>
</tr>
</tbody>
</table>
Frequency of responses by gender and age

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxious</td>
<td></td>
<td></td>
</tr>
<tr>
<td>worried</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sleeping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>terrified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>frightened</td>
<td></td>
<td></td>
</tr>
<tr>
<td>panic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anxious</td>
<td></td>
<td></td>
</tr>
<tr>
<td>worried</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sleeping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>terrified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>frightened</td>
<td></td>
<td></td>
</tr>
<tr>
<td>panic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Less than 65**
- **65+**
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

![Factor vs. raw scores](image-url)
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

Biserial correlation

Item position

- Terrified
- Frightened
- Panic
- Anxious
- Worried
- Sleeping
Items parameters with the Rasch model

- Terrified
- Frightened
- Panic
- Anxious
- Worried
- Sleeping

Probability of a positive response

θ
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 sufficient statistic for the Rasch model, any pattern of responses having the same total will get the same factor score.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-0.9210140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-0.9210140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-0.9210140</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-0.5829504</td>
<td></td>
</tr>
<tr>
<td>442</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>27</td>
<td>3.6233943</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>443</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>3.8282668</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>444</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>27</td>
<td>3.6233943</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>445</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>3.8282668</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>446</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>4.0537294</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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:
Items parameters with varying discrimination

Probability of a positive response

-6 -4 -2 0 2 4 6

0.0 0.2 0.4 0.6 0.8 1.0

θ

Probability of a positive response

terrified frightened panic
anxious worried sleeping
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.)

- Sleeping
- Worried
- Panic
- Terrified
Item threshold parameters (GRM uncons.)

- Sleeping
- Worried
- Panic
- Terrified
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?’.
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).


