

Descriptive and explanatory IRT modeling of a new Quality of Life questionnaire specific for HIV patients

Christophe LALANNE

ch.lalanne@gmail.com

Hôpital Saint-Louis, Department of Clinical Research, Paris, France

July 22, 2009

Outline

1. Measuring PROs
2. Overview of the questionnaire
3. IRT modeling: descriptive vs. explanatory approaches
4. Conclusions



Patient reported outcomes (PROs)

- PROs covers a broad range of patients' self-assessment, e.g. the number and intensity of symptoms experienced in the past 15 days, treatment adherence, (health-related) quality of life (HRQL).
- Until recently, therapeutic strategies have mainly relied on the analysis of effect size (non-inferiority, placebo, etc.), based on standardized raw scores.
- Impact on HRQL has become a key factor when releasing a new drug and measuring HRQL is now current practice in Clinical Trials (Phases II and III).

The question is: Does IRT provide additional evidence on HRQL sensibility to treatment impact or intervening factors (e.g., during follow-up)?



Why evaluate HRQL of PLWHA?

- Patients living with HIV or AIDS suffer from many side-effects: Lipoatrophy, daily intake and number of tablets.
- Comorbidities (depressive or psychiatric disorders) add up to viral load and immunodepression consequences.
- Existing HRQL instruments (MOS-HIV, FAHI) do not cope with such specific endpoints.
- There is a growing need for a HRQL questionnaire specific for PLWHA, which takes into account the new HAART era as well as skin- or mental-related consequences.



How to evaluate HRQL with PLWHA?

- A generic questionnaire (e.g. SF-36) lacks sensibility and responsiveness for that kind of patients.
- We need to address a large range of constructs, including: physical state (including body change self-perception), mental state, social relationships, sexual relationships, treatment bothering, coping.
- The FDA or EMEA clearly state what is expected from a generic or specific HRQL questionnaire. Usually, one has to demonstrate that the new questionnaire suits the conceptual (*construct validity*) and endpoint (*criterion validity* and *sensibility*) model.
- Ideally, the questionnaire should be adjusted or adjustable to any culture.



Validation of the PROQOL-HIV questionnaire

We follow “classical” steps for the validation of a new questionnaire:

1. Items generation, forward/backward translation (+ cultural adaptation);
2. Statistical analysis and questionnaire reduction:
 - item level: saturation effect, redundancy, % of missing responses, item-test correlation (polychoric and linear),
 - questionnaire level: convergent/discriminant validity, scale reliability;
3. Item cluster analysis, exploratory and confirmatory factor analyses, multi-trait scaling.

The questionnaire was reduced to 40 items and we considered two factorial structures that explained about 60% of the total variance. For scoring purpose, we retained the 4 factor-solution (≥ 5 items per dimension) for reliability considerations.



Items

Example of Items:

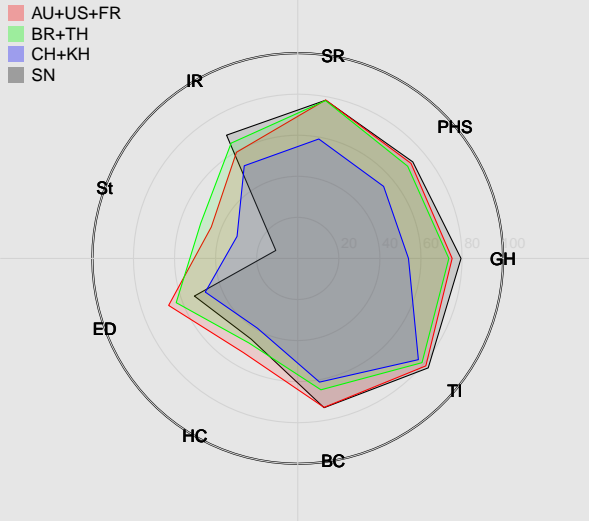
- ♣ I have felt tired
- ♣ I have had difficulty sleeping
- ♣ I have had difficulty concentrating
- or paying attention
- ♣ I have had difficulty with daily activities

Dimensions:

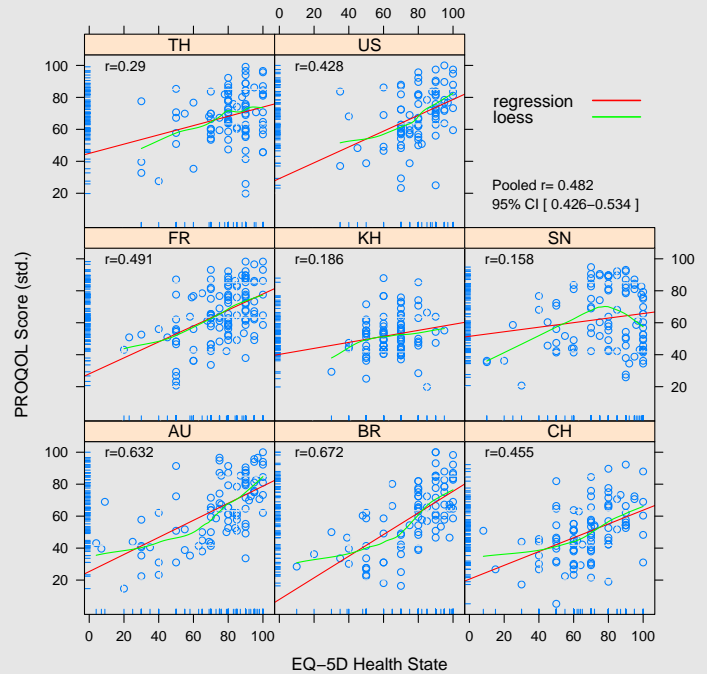
1. Physical health and symptoms (14 items + general health state)
2. Social and intimate relationships (5 items)
3. Treatment impact (10 items)
4. Emotional distress and health concern (10 items)



Sum scores analysis



Standardized scores on eight dimensions.



Total score against a generic instrument (EQ-5D).



Methodology

Our methodological strategy follows that of Wilson & De Boeck¹, in that we examine models for dichotomous responses of increasing complexity (hence explanatory power).

Person predictors		
Items predictors	none	person properties
none	doubly descriptive	person explanatory
Item properties	item explanatory	doubly explanatory

In the above, we exploited the fact that many IRT models, like the Rasch model, may be expressed as mixed-effects models:

$$\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{b}, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Sigma(\theta))$$

We also fitted the 2-PL model and models for ordinal responses with R^2 .

¹P De Boeck and M Wilson (eds.), *Explanatory Items Response Models*, Springer, 2004

²P Mair and R Hatzinger, Extended Rasch Modeling: The `eRm` package for the application of IRT models in R, *Journal of Statistical Software*, **20(9)**, 2007



Results for the Rasch Model

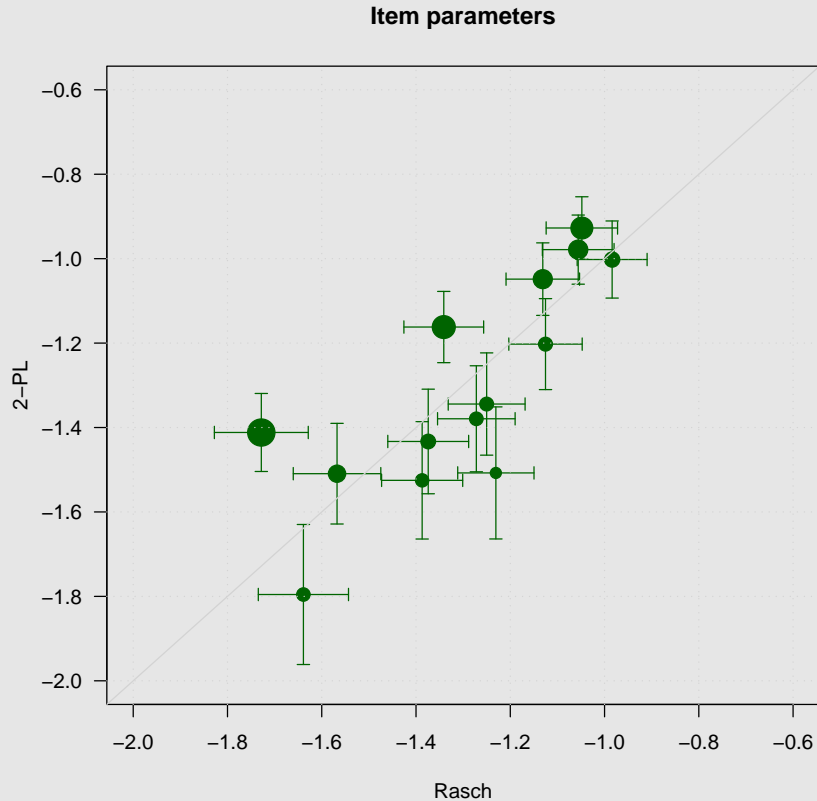
In what follows, we are considering a subset of 14 items ($> 30\%$ of explained variance in EFA). Estimation of items parameters was done using MML in R (ltm package³).

- The very limited range of item difficulties ($[-1.73; -0.98]$) reflects the ceiling effect observed in the distribution of individual responses (few people are actually reporting a “bad” HRQL).
- However, this model stands on the assumption of equal discriminative power for each item, which is obviously questionable.
- A 2-PL model indicated that items 2, 4 and 6 have higher discrimination (> 2.3) whereas most of the other discrimination parameters lie in the range $[1.4; 1.8]$. An LRT test indicated a better fit of the 2-PL model ($\chi^2(13) = 58.16, p < 0.001$).

³D Rizopoulos, ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses, *Journal of Statistical Software*, 17(5), 2006



Results for the Rasch Model (Con't)



Modeling persons' characteristics

If we take into account country of origin, gender, as well as clinical markers, which all are known to influence one's reported HRQL, the residual person variance is lower, 1.80 (0.17), compared to its estimation under the Rasch model (refitted with Stata), 3.04 (0.27). The analysis of model fits favors the LRRM (lower BIC).

In both cases, individual differences are statistically significant with $p < 0.001$. Under the Rasch model, the odds increase by a factor ≈ 5 when θ increase by one SD.

	LRRM est.		Models comp.			
	β	SE		df	AIC	BIC
country	-0.148	0.084				
gender	-0.265	0.138	Rasch	15	7603.9	7711.5
depression	-0.644	0.150	2-PL	28	7575.9	7776.5
symptoms	-0.138	0.011	LRRM	19	7307.6	7443.8



Modeling persons' characteristics (Con't)

Items parameters under LRRM:

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
i1	3.172698	.3534861	8.98	0.000	2.479878	3.865518
i2	3.340889	.3548226	9.42	0.000	2.645449	4.036328
i3	3.353076	.3549239	9.45	0.000	2.657438	4.048714
i4	3.768734	.3588724	10.50	0.000	3.065357	4.472111
i5	3.36479	.3550478	9.48	0.000	2.668909	4.060671
i6	4.447698	.3678772	12.09	0.000	3.726672	5.168724
i7	3.273144	.3542793	9.24	0.000	2.578769	3.967518
i8	3.703637	.3581509	10.34	0.000	3.001674	4.4056
i9	4.089461	.3626754	11.28	0.000	3.37863	4.800291
i10	3.927709	.3606914	10.89	0.000	3.220767	4.634651
i11	4.281441	.3652302	11.72	0.000	3.565603	4.997279
i12	3.488756	.3561023	9.80	0.000	2.790809	4.186704
i13	3.476063	.3559493	9.77	0.000	2.778415	4.173711
i14	3.737269	.358539	10.42	0.000	3.034546	4.439993



Do we need to dichotomize the responses?

Choosing the right way to dichotomize items isn't an easy task. Other well-known models allow to directly cope with categorical responses, whether it be ordinal or not:

- Partial-credit model (PCM): response categories do not necessarily reflect a constant gradation on the latent trait.
- Rating-scale model (RSM): constant threshold distances, for *all* items.

Testing one model against the other can be used to verify whether the threshold distances, commonly encountered with Lickert-type items, are a characteristic of the response format (RSM) or of the particular item (PCM).



Results for PCM and RSM

Estimation of items parameters was done using CML in R (eRm package). Results for model comparison are shown below:

	\mathcal{L}	AIC	BIC	cAIC
PCM	-6112.7	12295.4	12458.9	12493.9
RSM	-6168.1	12358.1	12409.5	12420.5

Although we might be inclined to favor the RSM (fewer parameters, lower BIC), the increase in the number of parameters (compared to RM) and the imbalance in response categories (e.g. 6 times more “never” than “always” responses) render these models probably less suitable for HRQL studies.



Conclusions

1. Compared to a traditional analysis where sum score is the outcome variable and patients' characteristics are the explanatory variables, we reached quite the same conclusions when using the IRT framework:
Women generally report a lower HRQL, and HRQL level is correlated to the number of experienced symptoms and the existence of a depressive disorder. Other factors seem to influence HRQL as well, namely anti-retroviral therapy, diagnosis date, coinfection (hepC/B).
2. However, IRT allows to work on a probability scale and may be used to weight items.



Conclusions

3. Furthermore, such an explanatory approach allows to study (some of the) items properties in relation to the probability of a positive response (Linear logistic test model, not shown here). Items formulation is an important issue when developing a new HRQL questionnaire, and it is interesting to study if reverse coded items or specific sentences might behave as expected.
4. IRT may be used to uncover DIF at the item level; however, when items are translated *and* adapted to local culture, the meaning of DIF is not so clear.

