Item Response Theory

Introductory Topics

Patrick Mair
Department of Statistics and Mathematics
WU Vienna University of Economics and Business

October 20, 2009

The Concept of Measurement

Problems of measurement occur in many scientific fields:

Physics (manifest variables)

- Time: year, month, hour, minute, second
- Measure of length: foot, cubit, meter
- Temperature: Celsius, Fahrenheit, Kelvin

Psychology (latent variables)

- Intelligence (IQ), cognitive abilities
- Personality traits

The Concept of Measurement

Recently IRT expanded into other fields such as

Medicine

- Pain scales
- Quality of life

Business, Marketing

- Customer satisfaction
- Product image, perception

IRT aims for measuring subject and item traits: Establish scale and map subjects/items on this scale.

Review: Classical Test Theory

What is usually done in CTT? $X = T + E$

- Selection of items due to reliability and discriminatory power
- Compare subjects by means of sum scores on an interval scale

Problems in CTT:

- Reliability cannot be computed in a direct way
  $\rho^2 = \sigma^2(T)/\sigma^2(X)$ or $\rho = r(X, T)$
- Measures are based on correlations ($\rightarrow$ sample dependent)
- No question about how the raw scores are achieved ("latent trait")
- Item/subject raw score for measuring item/subject difficulty/ability ($\rightarrow$ problem in measurement theory)

Definitions

Wikipedia: "Item response theory (IRT) is a body of theory describing the application of mathematical models to data from questionnaires and tests as a basis for measuring abilities, attitudes or other variables."

Basically, IRT allows for the analysis of responses on items by means of 2 parts:

- Item part: How difficult are the items?
- Subject part: How able are the persons?

Data structure

Starting point of each analysis: Person $\times$ Item matrix.

Data matrix $\mathbf{X}$ (5-item math ability test):

	I1	I2	I3	I4	I5	$\mathbf{R_v}$
Kurt	1	1	1	1	1	**5**
John	1	1	1	0	1	**4**
Achim	0	0	1	1	1	**3**
Daniel	1	0	0	0	0	**1**
Patrick	0	0	0	0	0	**0**
Christian	0	1	0	0	0	**1**
Michael H.	1	0	0	0	1	**2**
Stefan	1	0	0	0	0	**1**
Michael S.	1	0	1	0	0	**2**
David	1	1	1	0	0	**3**
$\mathbf{R_i}$	**7**	**4**	**5**	**2**	**4**	

Scaling items and persons

Aim (1) of IRT analysis (measurement):

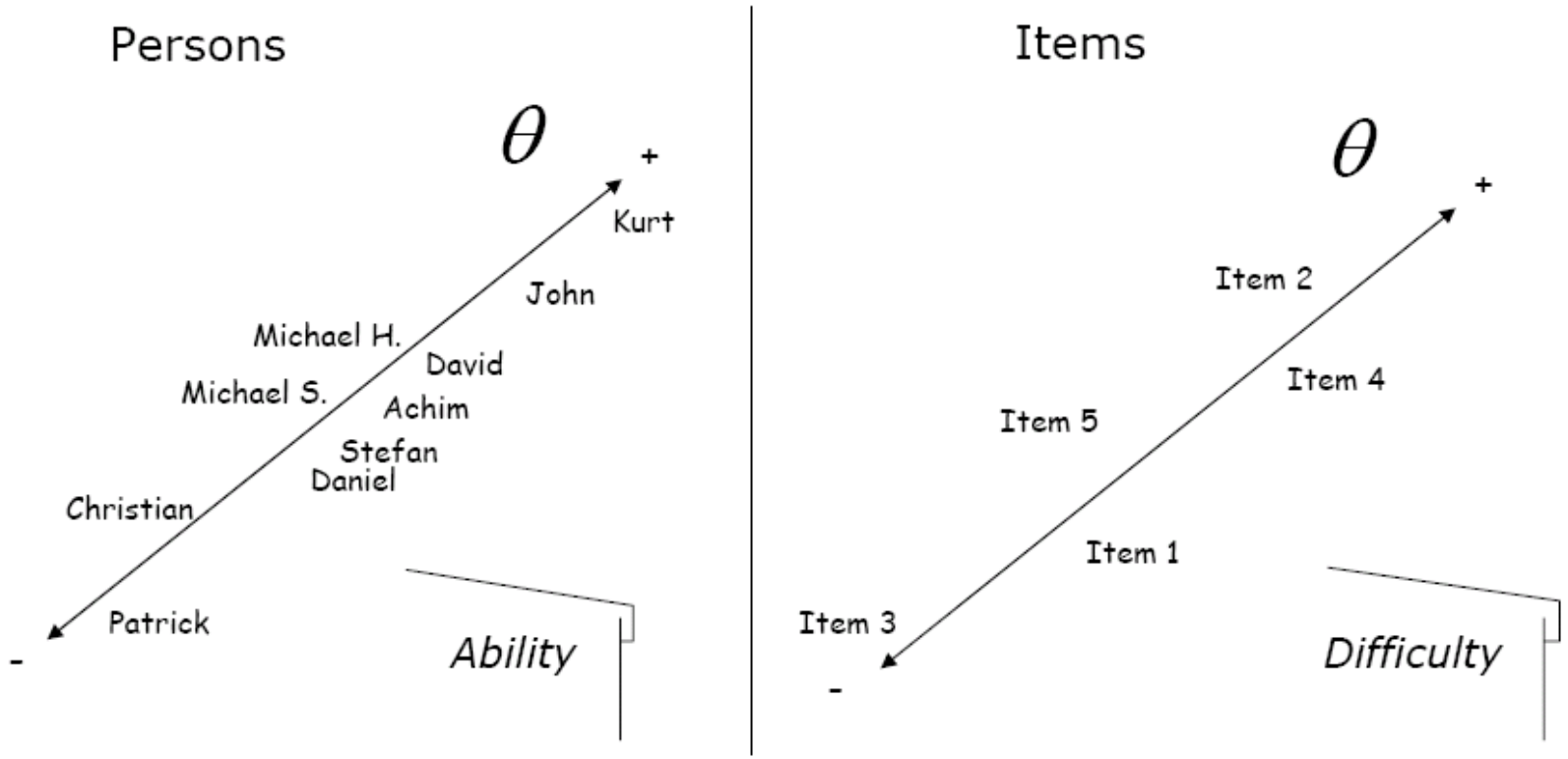
- Estimate a parameter ("difficulty" β_i) for each item i
- Estimate a parameter ("ability" θ_v) for each person v

...on an interval scale!

Raw scores: Sum of ordinal measures (allowed?) are still ordinal!
$\rightarrow$ **IRT provides interval scaled measures based on ordinal data**

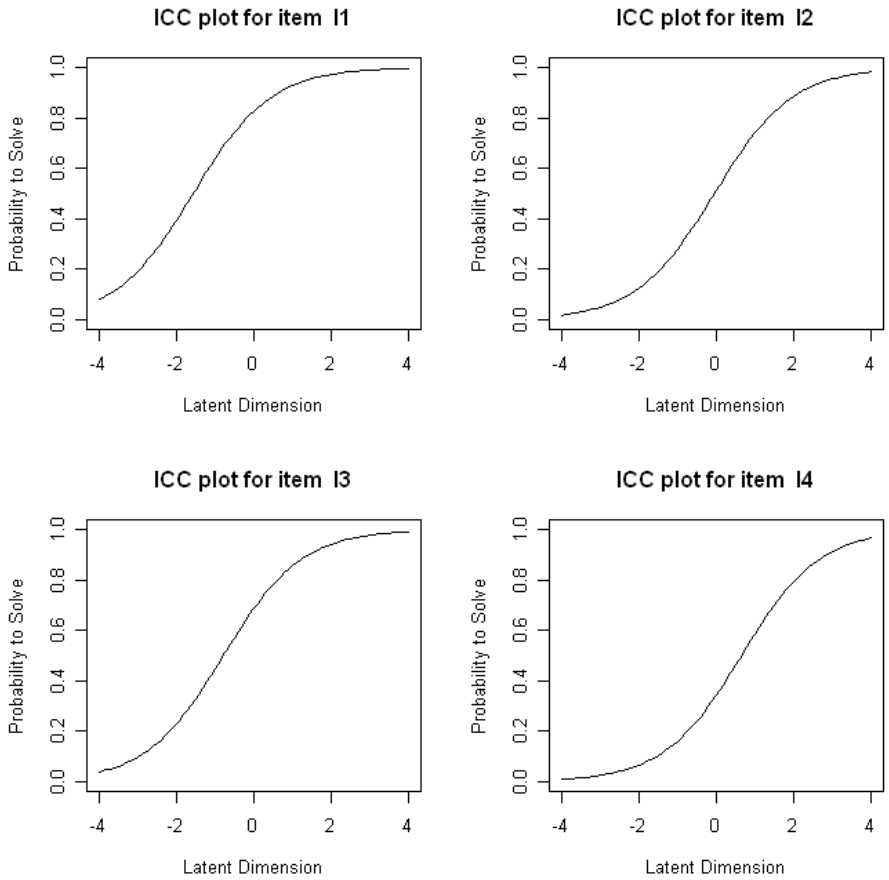
Scaling items and persons

Mapping items AND persons on unidimensional latent trait:



Item characteristic curves

Aim (2) in IRT: We estimate functions which allow to determine the probability for a correct answer of person v on item i.



Item subsets

Aim (3) in IRT: Item Selection

Suppose that we want to establish a test for math ability:

- We collect a "universe" of items
- We want to choose a "good" subset in terms of fairness, discriminatory power, etc.
- Apply IRT models to get such subsets

Basically, you search for data that fit a certain IRT model.

IRT approaches

There are basically two approaches for IRT modeling:

- Modeling approach: As usual in statistics, you have your observed data and you try to find a (highly parametrized) model which fits these data (US school), i.e. emphasis on aim (1) and aim (2). Main focus is on person parameters.
- Test construction approach: "Raschians", Rasch model is godlike and you have to find data (by eliminating items) which fit the Rasch model (European school), i.e. emphasis on aim (3). Main focus is on item parameter.

The Rasch model

Logistic model for dichotomous data proposed by Rasch (1960):

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

X_{vi} ... response of person v on item i
θ_v ... ability parameter of person v
β_i ... difficulty parameter of item i

Outstanding property: *Specific objectivity* (SO; mathematical and epistemological)

- Item and person parameters can be estimated independently from each other (*separability theorem*)
- Statements on persons/items independent from the item/person sample

Parameter Estimation

Basically, there are 3 ways for estimating the parameters:

- Joint ML: Not recommended; the larger the sample the more parameters I have to estimate.
- Conditional ML: Is the direct "translation" of SO into statistical theory. Very elegant but limited way for estimation.
- Marginal ML: I have to pose an assumption on the distribution of the θ's (usually normal). Very flexible and fast method.

Assumptions

Assumptions/implications of the Rasch model:

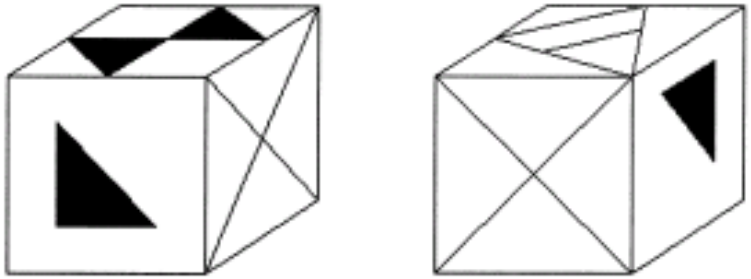
- Unidimensionality
- Sufficiency of the raw scores
- Parallel ICC's
- Local independence: The only source of item correlations is allowed to be θ_v which leads to $p(\mathbf{x}|\theta) = \prod_{i=1}^{K} p(x_i|\theta)$.

In general, the classical Rasch model is a very restrictive model (...too restrictive?).

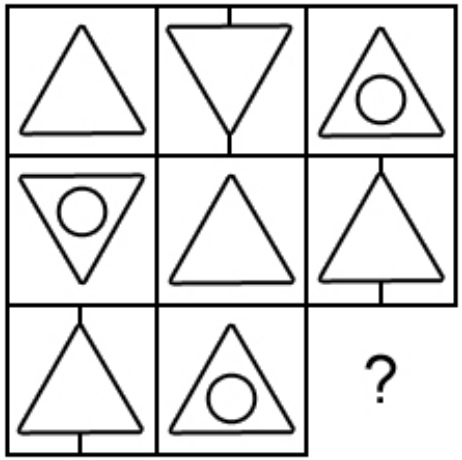
At the end the Rasch model is still a "seal of approval" for (psychological) tests which allows for a straightforward interpretation of the parameters. It can be tested extensively (goodness-of-fit, specific assumptions, item-wise tests, etc.)

Examples of Rasch homogeneous tests

3DW (3-dimensional cube test; Gittler, 1990)



WMT (Wiener Matrizen Test; Formann & Piswanger, 1979)



More general IRT models

- Relaxing the parallel ICC assumption: 2-parameter logistic models (2-PL)

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

  Each item gets a weight and thus the raw score is not sufficient anymore.

- Introducing a guessing parameter (lower asymptote): 3-PL

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

- Introducing an upper asymptote parameter: 4-PL

More general IRT models

- Relaxing the unidimensionality assumption: Multidimensional IRT models, e.g., PISA study. Math results from 2003:

1. Hong Kong
2. Finland
...
18. Austria
19. Germany
...
28. USA
...

- Modern approach: Put these models into the framework of GLMM (*generalized linear mixed models*) and allowing for item and person covariates, longitudinal and multilevel data.

Polytomous models

- Rating scale model (RSM): Polytomous extension of the Rasch model; adjacent category logits.
- Partial credit model (PCM): Extension of the RSM in terms of different number of categories per item.
- Graded response model (GRM): Extension of the 2-PL to polytomous data; cumulativ category logits.
- Nominal response model (NRM): Category discrimination parameter; baseline category logits.

Applications

In psychology IRT is classically used for testing cognitive abilities, but also for personality scales, in economic psychology, education, etc. Whenever your aim is to measure latent traits (with respect to the issues explained above), you can take into account IRT-models. At the moment IRT is expanding rapidly to other research fields than psychology:

- Marketing: Scales for affective responses to consumptions, cognitive dissonance in consumer behavior, leisure activities, cross-cultural measurements for advertisements, etc.
- Health Sciences: Scales for general health status (SF-36), rehabilitation programs, radiographic scales, etc. (see Bezruzko, 2005)
- Sports, ...

Future of IRT

Modern IRT research:

- GLMM (covariates)
- Multidimensional models
- Close gap to SEM

Will the questionnaire survive?

- Text data (text mining, qualitative research)
- Neurological approaches (Neuromarketing)

Software

A couple of programs for IRT estimation are available: Some of them are limited to certain model families, some of them not very easy to handle.

- MULTILOG, BILOG, PARSCALE (SSI group)
- WINSTEPS (Linacre, 2006)
- RUMM (Andrich et al., 2000)
- WINMIRA (von Davier, 2000)
- Packages `ltm` (Rizopolous, 2006) and `eRm` in `R` (open source)

Additional Readings

- Bond, T.G., & Fox, C.M. (2007). Applying the Rasch model. (*very introductory, describes concepts completely non-mathematical, detailed examples and basic plots*)
- Embretson, S., & Reise, S. (2000). Item response theory for psychologists. (*introductory, presents many extended models, many examples and plots*)
- Baker, F.B., & Kim, S. (2004). IRT: Parameter estimation techniques. (*only for programmers*)
- Fischer, G., & Molenaar, I. (1995). Rasch models: Foundations, recent developments and applications. (*very technical, highly structured, all about Rasch-type models, very few examples*)
- de Boeck, P., & Wilson, M. (2004). Explanatory item response models. (*GLMM framework, comprehensive, certainly a future direction in IRT*)