Item response theory

Joachim Vandekerckhove

Spring 2025

Item response theory

A set of methods and models to think about testing and to develop exams and questionnaires.

Item response theory

A set of methods and models to think about testing and to develop exams and questionnaires.

Particularly useful when

- multiple respondents are presented with multiple items

Item response theory

A set of methods and models to think about testing and to develop exams and questionnaires.

Particularly useful when

- multiple respondents are presented with multiple items
- respondents may differ in their ability (or some other underlying construct)

Item response theory

A set of methods and models to think about testing and to develop exams and questionnaires.

Particularly useful when

- multiple respondents are presented with multiple items
- respondents may differ in their ability (or some other underlying construct)
- items may differ in their difficulty (or their correspondence to the construct)

Representations in IRT

Most commonly, the data are *dichotomous*: $X_{ip} \in \{0, 1\}$

Representations in IRT

Most commonly, the data are *dichotomous*: $X_{ip} \in \{0, 1\}$

A respondent p has an ability, $\theta_p \in \mathbb{R}$

Representations in IRT

Most commonly, the data are *dichotomous*: $X_{ip} \in \{0, 1\}$

A respondent p has an ability, $\theta_p \in \mathbb{R}$

An item i has a difficulty, $\beta_i \in \mathbb{R}$

Representations in IRT

Most commonly, the data are *dichotomous*: $X_{ip} \in \{0, 1\}$

A respondent p has an ability, $\theta_p \in \mathbb{R}$

An item i has a difficulty, $\beta_i \in \mathbb{R}$

Ability is not directly observable – we call it a *latent trait* (Birnbaum, 1968)

Representations in IRT

Most commonly, the data are *dichotomous*: $X_{ip} \in \{0, 1\}$

A respondent p has an ability, $\theta_p \in \mathbb{R}$

An item i has a difficulty, $\beta_i \in \mathbb{R}$

Ability is not directly observable – we call it a *latent trait* (Birnbaum, 1968)

Critically, in IRT we assume that the ability θ and the difficulty β are directly comparable

Representations in IRT

Most commonly, the data are *dichotomous*: $X_{ip} \in \{0, 1\}$

A respondent p has an ability, $\theta_p \in \mathbb{R}$

An item i has a difficulty, $\beta_i \in \mathbb{R}$

Ability is not directly observable – we call it a *latent trait* (Birnbaum, 1968)

Critically, in IRT we assume that the ability θ and the difficulty β are directly comparable

We can quantify the dominance, $\eta_{ip} \in \mathbb{R}$, a respondent has over an item

$$\eta_{ip} = \theta_p - \beta_i$$

Linking functions

We're now interested in using that dominance to express the probability that the respondent will answer an item 'correctly'

Linking functions

We're now interested in using that dominance to express the probability that the respondent will answer an item 'correctly'

Note that $\eta_{ip} \in \mathbb{R}$: the dominance parameter lives in the domain of the real numbers – it is not a probability

Linking functions

We're now interested in using that dominance to express the probability that the respondent will answer an item 'correctly'

Note that $\eta_{ip} \in \mathbb{R}$: the dominance parameter lives in the domain of the real numbers – it is not a probability

In order to express the dominance as a probability, we need to *map* the parameter to the $(0, 1)$ domain, where probabilities live

Linking functions

We're now interested in using that dominance to express the probability that the respondent will answer an item 'correctly'

Note that $\eta_{ip} \in \mathbb{R}$: the dominance parameter lives in the domain of the real numbers – it is not a probability

In order to express the dominance as a probability, we need to *map* the parameter to the $(0, 1)$ domain, where probabilities live

For this, we will use a *linking* function – a function that takes as input $\eta_{ip} \in \mathbb{R}$ and gives as output a value $(0, 1)$

$$\mathbb{R} \xrightarrow{\text{link}} (0, 1)$$

Linking functions

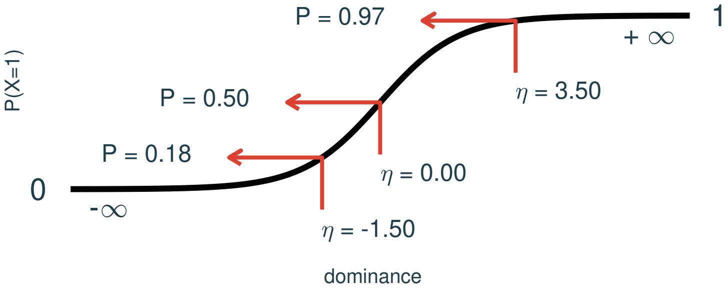
In IRT, we use the *logistic function* (also known as the *inverse logit*):

$$P\left(X_{ip}=1\right)=\pi_{ip}=\frac{1}{1+\exp(-\eta_{ip})}$$

Linking functions

In IRT, we use the *logistic function* (also known as the *inverse logit*):

$$P\left(X_{ip}=1\right)=\pi_{ip}=\frac{1}{1+\exp(-\eta_{ip})}$$



The Rasch model

A much more convenient way of writing that is:

$$
\begin{aligned}
P\left(X_{ip}=1\right) &= \text{ilogit}\left(\eta_{ip}\right) \\
&= \text{ilogit}\left(\theta_p-\beta_i\right) \\
\Leftrightarrow \quad \text{logit}\left(P\left(X_{ip}=1\right)\right) &= \theta_p-\beta_i
\end{aligned}
$$

The Rasch model

A much more convenient way of writing that is:

$$
\begin{aligned}
P\left(X_{ip}=1\right) &= \text{ilogit}\left(\eta_{ip}\right) \\
&= \text{ilogit}\left(\theta_p-\beta_i\right) \\
\Leftrightarrow \quad \text{logit}\left(P\left(X_{ip}=1\right)\right) &= \theta_p-\beta_i
\end{aligned}
$$

This exact formulation is one of the most common IRT models (there are a few). This is called the *Rasch* model, after Danish psychometrician Georg Rasch (Rasch, 1961)

The Rasch model

A much more convenient way of writing that is:

$$\begin{aligned} P\left(X_{ip}=1\right) &= \text{ilogit}\left(\eta_{ip}\right) \\ &= \text{ilogit}\left(\theta_p - \beta_i\right) \\ \Leftrightarrow \quad \text{logit}\left(P\left(X_{ip}=1\right)\right) &= \theta_p - \beta_i \end{aligned}$$

This exact formulation is one of the most common IRT models (there are a few). This is called the *Rasch* model, after Danish psychometrician Georg Rasch (Rasch, 1961)

The attraction of this model is that it lets us take the data matrix $\mathbf{X}$ and infer person-specific abilities and item-specific difficulties

Item response data

The data matrix $\mathbf{X}$ is simply the person-by-item matrix of 0s and 1s

person	item 1	item 2	item 3	...
1	1	1	1	...
2	1	1	0	...
3	0	1	0	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
67	1	1	1	...

Guttman scales

Note that the Rasch model makes some strong assumptions about the data.

Guttman scales

Note that the Rasch model makes some strong assumptions about the data.

Non-exhaustively, it assumes that

- Only the person's ability and the item difficulty affect the proability of responding correctly

Guttman scales

Note that the Rasch model makes some strong assumptions about the data.

Non-exhaustively, it assumes that

- Only the person's ability and the item difficulty affect the proability of responding correctly
- The response on one item does not affect the response on any other item

Guttman scales

Note that the Rasch model makes some strong assumptions about the data.

Non-exhaustively, it assumes that

- Only the person's ability and the item difficulty affect the proability of responding correctly
- The response on one item does not affect the response on any other item
- Items can be rank ordered in difficulty

Guttman scales

Note that the Rasch model makes some strong assumptions about the data.

Non-exhaustively, it assumes that

- Only the person's ability and the item difficulty affect the proability of responding correctly
- The response on one item does not affect the response on any other item
- Items can be rank ordered in difficulty
- Participants can be rank ordered in ability

Guttman scales

Those last two are part of the *unidimensionality* assumption, which has a peculiar consequence

Guttman scales

Those last two are part of the *unidimensionality* assumption, which has a peculiar consequence

Consider this segment of the data matrix:

person	I_1	I_2	I_3
1	1	1	1
2	1	1	0
3	0	1	0
67	1	1	1

Guttman scales

Those last two are part of the *unidimensionality* assumption, which has a peculiar consequence

Consider this segment of the data matrix:

person	I_1	I_2	I_3
1	1	1	1
2	1	1	0
3	0	1	0
67	1	1	1

$\rightarrow$

person	I_2	I_1	I_3
3	1	0	0
2	1	1	0
1	1	1	1
67	1	1	1

If the unidimensionality assumption holds, the data matrix can be rearranged so that all its zeros are in a triangle

Guttman scales

If this rearrangement is possible, then we can combine the item scores into a single sum score, which can then be interpreted as a unidimensional measure of the ability of the respondents.

person	I_2	I_1	I_3	Sum Score
3	1	0	0	1
2	1	1	0	2
1	1	1	1	3
67	1	1	1	3

Guttman scales

If this rearrangement is possible, then we can combine the item scores into a single sum score, which can then be interpreted as a unidimensional measure of the ability of the respondents.

person	I_2	I_1	I_3	Sum Score
3	1	0	0	1
2	1	1	0	2
1	1	1	1	3
67	1	1	1	3

A sum score that meets this condition is called a *Guttman scale*

Guttman scales

In practice, any data set large enough will violate these assumptions somewhat, and so the sum score is not a ‘perfect’ Guttman scale

Guttman scales

In practice, any data set large enough will violate these assumptions somewhat, and so the sum score is not a ‘perfect’ Guttman scale

The most common reasons for violation of these assumptions are:

- There may be randomness in the data (e.g., respondents ‘lapse’ and choose the wrong answer due to distraction), or

Guttman scales

In practice, any data set large enough will violate these assumptions somewhat, and so the sum score is not a ‘perfect’ Guttman scale

The most common reasons for violation of these assumptions are:

- There may be randomness in the data (e.g., respondents ‘lapse’ and choose the wrong answer due to distraction), or
- The underlying construct may really be multidimensional (e.g., a math test that uses difficult phrases is really also a test of English comprehension), or

Guttman scales

In practice, any data set large enough will violate these assumptions somewhat, and so the sum score is not a 'perfect' Guttman scale

The most common reasons for violation of these assumptions are:

- There may be randomness in the data (e.g., respondents 'lapse' and choose the wrong answer due to distraction), or
- The underlying construct may really be multidimensional (e.g., a math test that uses difficult phrases is really also a test of English comprehension), or
- The order of the items causes sequential effects (some examples in Moore, 2002)

Item response theory

The Rasch model, $P(X_{ip}) = \text{ilogit}(\theta_p - \beta_i)$, is a relatively simple IRT model. In psychometric circles, it is often called the *one-parameter logistic model* or simply 1PL

Item response theory

The Rasch model, $P(X_{ip}) = \text{ilogit}(\theta_p - \beta_i)$, is a relatively simple IRT model. In psychometric circles, it is often called the *one-parameter logistic model* or simply 1PL

Some of its relatives are:

- 2PL: $P(X_{ip}) = \text{ilogit}(\alpha_i\,(\theta_p - \beta_i))$
  Here, α_i is a discriminability index that captures how important the item i is for measuring the construct

Item response theory

The Rasch model, $P(X_{ip}) = \text{ilogit}(\theta_p - \beta_i)$, is a relatively simple IRT model. In psychometric circles, it is often called the *one-parameter logistic model* or simply 1PL

Some of its relatives are:

- 2PL: $P(X_{ip}) = \text{ilogit}(\alpha_i\,(\theta_p - \beta_i))$
  Here, α_i is a discriminability index that captures how important the item i is for measuring the construct
- 3PL: $P(X_{ip}) = \gamma_i + (1 - \gamma_i) \times \text{ilogit}(\alpha_i\,(\theta_p - \beta_i))$
  Here, γ_i is a guessing parameter that captures how well a participant could do by guessing the answer

Item response theory

The Rasch model, $P(X_{ip}) = \text{ilogit}(\theta_p - \beta_i)$, is a relatively simple IRT model. In psychometric circles, it is often called the *one-parameter logistic model* or simply 1PL

Some of its relatives are:

- 2PL: $P(X_{ip}) = \text{ilogit}(\alpha_i\,(\theta_p - \beta_i))$
  Here, α_i is a discriminability index that captures how important the item i is for measuring the construct
- 3PL: $P(X_{ip}) = \gamma_i + (1 - \gamma_i) \times \text{ilogit}(\alpha_i\,(\theta_p - \beta_i))$
  Here, γ_i is a guessing parameter that captures how well a participant could do by guessing the answer
- Explanatory item response models: $P(X_{ip}) = \text{ilogit}\,(\theta_p - \gamma W_i)$
  Here, β_i is replaced by an expression based on some predictor W

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading: Addison-Wesley.

Moore, D. W. (2002). Measuring new types of question-order effects: Additive and subtractive. *The Public Opinion Quarterly*, *66*(1), 80–91.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333).

Item response theory

Joachim Vandekerckhove

Spring 2025