IRT (GMMSGE01): Parametric IRT (dichotomous data)

Jorge N. Tendeiro

14 December 2017

Contents

1 Main idea

Parametric item response theory (IRT) provides a theoretical framework that allows modeling the relationship

$$\text{item} \longleftrightarrow \text{person}$$

by means of a mathematical function:

$$P(X_i = c|\theta_n) = f(\theta_n)$$

- X_i is the random variable denoting the answer to item i, with discrete response categories;
- c is the observed response:
  - If X is dichotomous, $c = 0, 1$. Usually 0 denotes incorrect answers and 1 denotes correct answers.
  - If X is polytomous, $c = 0, 1, \ldots, m$ $(m > 1)$.
- $\theta_n = n^{\text{th}}$ person's trait parameter.

This is the *item response function (IRF)*. The IRF is therefore a function relating the latent trait to the probability of answering the item correctly.

In IRT, items and persons are located on the same latent scale. For example, see Figure 1 displaying the IRF of a dichotomous item. The item location or difficulty (to be defined shortly) is indicated by the value b in the x-axis for which the probability of a correct answer is .50. Thus, values in the x-axis denote item locations. Moreover, values in the x-axis can also be interpreted as person latent scores θ. For example, a

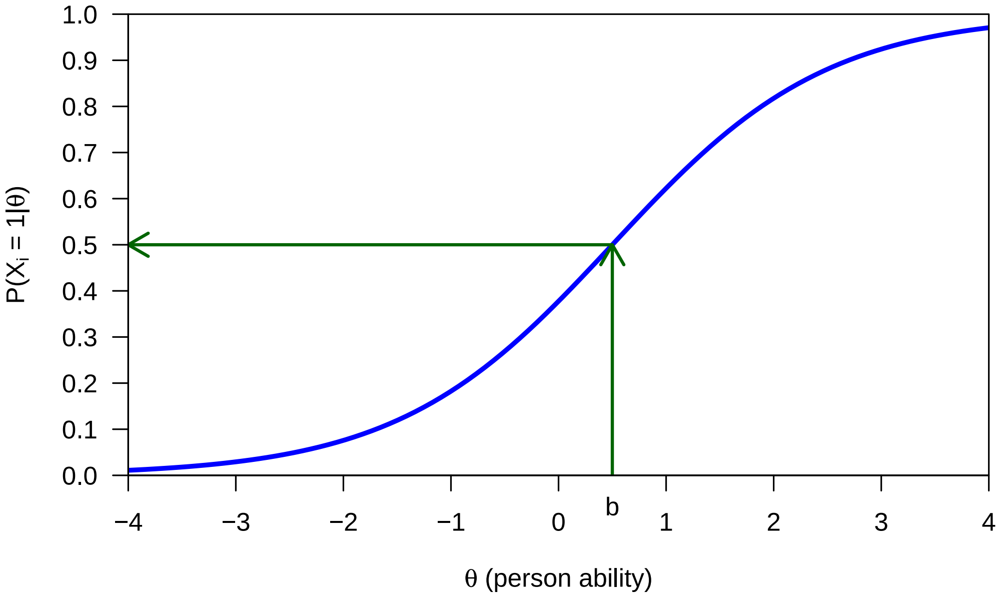

Figure 1: Items and persons located on a common scale

person with parameter $\theta = 2$ (i.e., $\theta > b$) has probability above .50 of answering this item correctly. Similarly, persons with trait $\theta < b$ are more likely to answering this item incorrectly.

The main assumptions of the mainstream parametric IRT models are the following:

- The IRF has a mathematically closed form that, when plotted, looks like a smooth S-shaped curve (see Figure 1).
- Responses to items are independent conditional of θ. This is known as *local independence*. In mathematical terms:

$$P(X_i = 1, X_j = 1|\theta) = P(X_i = 1|\theta)P(X_j = 1|\theta)$$

- Unidimensionality: One latent trait suffices to explain the relationship between the items.

In this lecture we will cover the mostly used IRT models for dichotomous data: The one-, two-, and three-parameter logistic models. We will also talk briefly about model estimation and model fit.

2 The Rasch model (1PLM)

The item response function (IRF) for item i is given by

$$P(X_i = 1|\theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)}, \tag{1}$$

where

- θ is the person's latent ability parameter;
- δ_i is the item's *difficulty* or popularity.

This IRF is known as the one-parameter logistic model because it involves only one item parameter (difficulty δ) and its functional form is based on the logistic function.

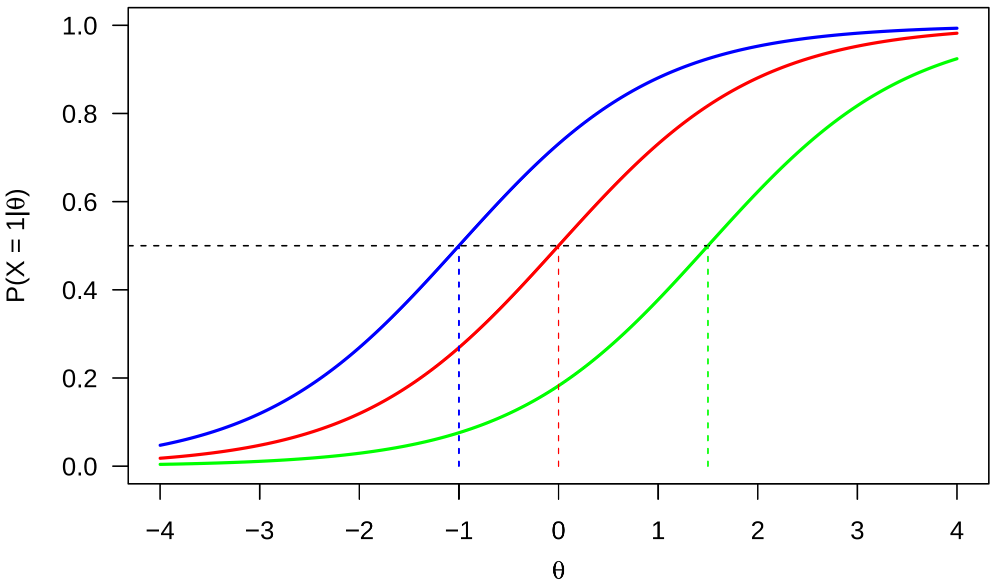

Figure 2: 1PLM (Rasch model)

Figure 2 shows three examples of IRFs under the Rasch model for three difficulty levels: $\delta = -1$ (blue), $\delta = 0$ (red), and $\delta = 1.5$ (green).

Properties of IRFs from the Rasch model:

- The probabilities increase with the latent trait θ for each item. So, the higher the latent trait, the higher the probability of answering the item correctly.
- The curves are 'parallel', that is, they do not intersect.
- Items differ only in difficulty: Easier items are on the left (low difficuly δ, like the blue item) and difficult items are on the right (high difficulty δ, like to green item).
- Parameter δ is the value of latent trait for which the probability of answering the item correctly is 50%: $P(X = 1|\theta = \delta) = .50$.

3 The 2PL model (2PLM)

The 2PLM generalizes the 1PLM by adding a new item parameter known as the *discrimination* parameter. The IRF for item i is given by

$$P(X_i = 1|\theta) = \frac{\exp[\alpha_i(\theta - \delta_i)]}{1 + \exp[\alpha_i(\theta - \delta_i)]}, \qquad (2)$$

where α_i is the discrimination parameter for item i. This parameter takes on positive real values, although values larger than, say, 4 or 5 are not common.

Figure 3 shows three examples of IRFs under the 2PLM. The difficulties are the same as in Figure 2. The discrimination parameters are as follows: $\alpha = 2$ (blue), $\alpha = .5$ (red), and $\alpha = 1$ (green).

Properties of IRFs from the 2PLM:

- The 2PLM is also a *cumulative* model: The probability of a correct response is expected to increase with θ.

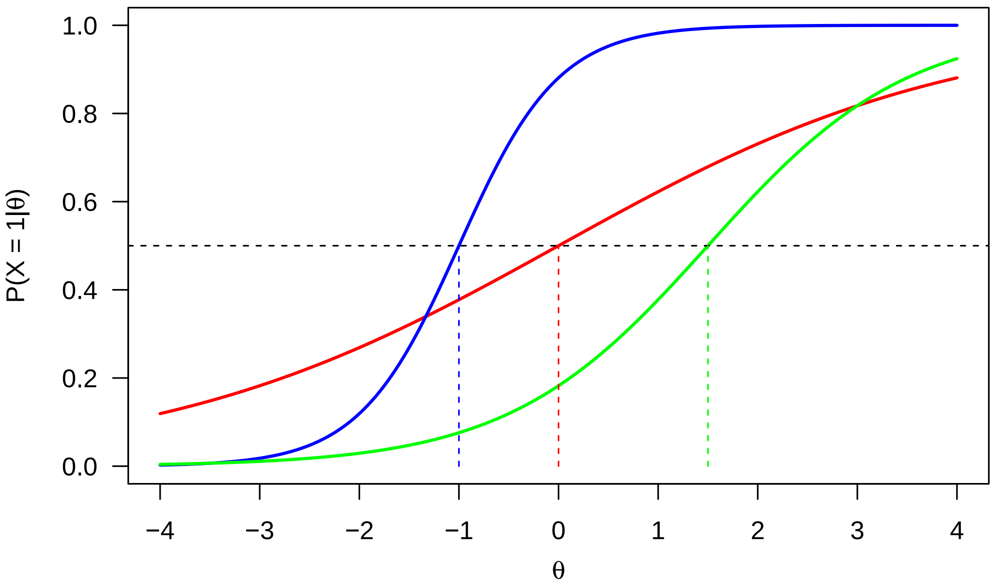

Figure 3: 2PLM

- The discrimination parameter determines the steepness of the curve at the difficulty point: The larger α, the steeper the curve. Steeper curves are better to distinguish (*discriminate*) among persons standing close to each other in this reagion of the latent trait.
- The difficulty parameters δ are equivalently interpreted as in the 1PLM.
- The IRFs may now intersect, as Figure 3 shows. So, items may be differently ranked in terms of difficulty for varying levels of θ.
- The 2PLM is more flexible than the 1PLM because it allows items to differ in terms of discrimination. When all items have similar discrimination then the 1PLM is preferred (for parsimony).

4 The 3PL model (3PLM)

The 3PLM generalizes the 2PLM by adding a new item parameter known as the *(pseudo)guessing* parameter. The IRF for item i is given by

$$P(X_i = 1|\theta) = \gamma_i + (1-\gamma_i)\frac{\exp[\alpha_i(\theta-\delta_i)]}{1+\exp[\alpha_i(\theta-\delta_i)]}, \tag{3}$$

where γ_i is the pseudoguessing parameter for item i. This parameter takes on values between 0 and 0.5. It expresses the property that even very low ability persons have a positive probability of answering to an item correctly simply by randomly guessing the correct answer (think of multiple choice items). For example, γ_i equals .25 for a four-options multiple choice item, which means that a respondent purely guessing the answer to an item would choose the correct answer with probability 1/4.

Figure 4 shows three examples of IRFs under the 3PLM. The items are the same as in Figure 3. The pseudoguessing parameters are as follows: $\gamma = 0$ (blue), $\gamma = .25$ (red), and $\gamma = .20$ (green).

Properties of IRFs from the 3PLM:

- The 3PLM is a *cumulative* model: The probability of a correct response is expected to increase with θ.

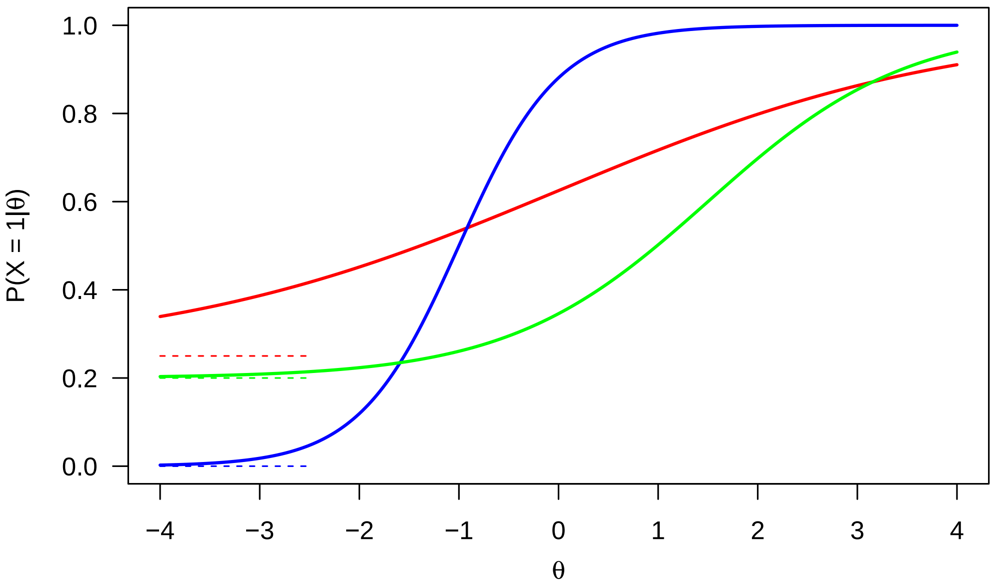

Figure 4: 3PLM

- The pseudoguessing parameter introduces a lower asymptote to the IRF (symbolized by the dashed horizontal segments on the low-left section of Figure 4.
- The 3PLM is more flexible than the 2PLM because it allows for different lower asymptotes. However, pseudoguessing parameters are often diffcult to estimate accurately (they are often associated to relatively large SEs). Also, it has been argued that seldom respondents answer to multiple choice items by purely guessing, that is, it is usually the case that one of more distractors are identified. In this case, γ underestimates the real guessing probability involved in the answering process. For these reasons, one should carefully check whether the 3PLM is suitable over and above the 2PLM.
- The difficulty parameters δ are now slightly differently interpreted. For item i, δ_i is the point of the latent trait at which the probability of a correct answer equals $\frac{\gamma_i+1}{2}$ (i.e., the midpoint between γ_i and 1). For example, for the green item ($\delta = 1.5$, $\alpha = 1$, $\gamma = .20$), we have that (check this using the IRF!)

$$P(X_i = 1|\theta = 1.5) = \frac{.20 + 1}{2} = .60.$$

- The discrimination parameters are similarly interpreted as in the 2PLM.
- The 3PLM has many similarities with the MMH model, however for the 3PLM the IRFs are parametric logistic functions.

5 Model estimation

Two sets of parameters need to be estimated when fitting an IRT model: The item parameters (difficulty, discrimination, pseudoguessing) and the person parameters (θ). There are various estimation algorithms available for items (joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, and Bayesian MCMC). Given the estimated item parameters, there are then various estimation algorithms for the person parameters (maximum likelihood, MAP, EAP). Here we will only lay out the general principle concerning *maximum likelihood estimation* for person scoring. So, we will assume that the item parameters are known and explain how the θ parameters are then estimated.

5.1 Person scoring

Suppose we have scores of N persons on I items collected in an $N \times I$ data matrix. For person s $(s = 1, \ldots, N)$ the response pattern is denoted by $X_s = (u_{s1}, u_{s2}, \ldots, u_{sI})$, where each item score u_{si} $(i = 1, \ldots, I)$ is either 0 (incorrect answer) or 1 (correct answer). The probability of observing this response pattern under the IRT model of interest (1PLM, 2PLM, or 3PLM) is given by

$$P(X_s|\theta_s) = \prod_{i=1}^{I} P_{si}^{u_{si}}(1 - P_{si})^{1-u_{si}}, \tag{4}$$

where $P_{si} = P(X_i = 1|\theta_s)$, that is, the probability of person s providing a correct answer to item i. The explicit formula for P_{si} depends on the IRT model one wants to fit to the data (Equation (1) for the 1PLM, Equation (2) for the 2PLM, or Equation (3) for the 3PLM). For example, if $I = 4$ and the response pattern is $X_s = (1, 1, 1, 0)$ then Equation (4) is equal to

$$P(X_s|\theta_s) = P_{s1}P_{s2}P_{s3}(1 - P_{s4}).$$

We are allowed to multiply the item probabilities across person's s response pattern due to the local independence assumption.

Equation (4) defines the *likelihood* function for θ_s:

$$L(\theta_s|X_s) = \prod_{i=1}^{I} P_{si}^{u_{si}}(1 - P_{si})^{1-u_{si}}. \tag{5}$$

With known item parameters, Equation (5) is a (rather complex) function of θ_s only. MLE tries to finds the value for θ_s, say $\widehat{\theta}_s$, that maximizes the likelihood function. That is, $\widehat{\theta}_s$ is the value that makes it more likely that a person with this latent trait produces a response pattern such as $X_s = (u_{s1}, u_{s2}, \ldots, u_{sI})$.

Typically it is the natural logarithm of the likelihood function that one maximizes, for numerical reasons (observe that this is allowed because the logarithm function is monotonic). So, the function being maximized is given by

$$\ln L(\theta_s|X_s) = \sum_{i=1}^{I} u_{si}\ln(P_{si}) + (1 - u_{si})\ln(1 - P_{si}). \tag{6}$$

The algorithm used to estimate θ_s is iterative (commonly based on the Newton-Raphson method). The algorithm proceeds as follows:

1. Specify a starting value for θ_s, say, $\theta_s^{(0)}$ (e.g., 0).

2. Compute the first- and second-order derivatives of the log-likelihood function (Equation (6)) at $\theta_s^{(0)}$, say, $D1(\theta_s^{(0)})$ and $D2(\theta_s^{(0)})$. The explicit formulas for these derivatives under the 1PLM, 2PLM, or 3PLM can be found, for example, in Baker and Kim (2000) and Embretson and Reise (2000, for the 1PLM and 2PLM only).

3. Compute $\varepsilon_1 = D1(\theta_s^{(0)})/D2(\theta_s^{(0)})$.

4. Update θ_s:

$$\theta_s^{(1)} = \theta_s^{(0)} - \varepsilon_1$$

.

Table 1: Toy example

Item	Item score	alpha	delta
1	1	1.0	-1.5
2	1	1.5	0.0
3	0	2.0	1.5

5. Repeat steps 2-4 using the updated θ from the last step. Proceed until $|\varepsilon|$ is smaller than a predefined threshold (e.g., .001) or until a maximum number of iteration steps has been reached (e.g., 100).

This method is expected to converge to the maximum of the (log) likelihood function.

5.2 Information, SE

The standard error (SE) of the estimated person parameter $\widehat{\theta}_s$ provides a measure of uncertainty of the estimate. The SE is directly related to the so-called *item information function*. This function indicates what regions of the theta scale are measured more precisely by the item (higher information) and what regions are measured less precisely (lower information).

For a dichotomous item, the item information function is given by (see Baker & Kim, 2004)

$$Inf_i(\theta) = \left[\alpha_i^2 \frac{1 - P_i(\theta)}{P_i(\theta)}\right] \left[\frac{P_i(\theta) - \gamma_i}{1 - \gamma_i}\right]^2. \tag{7}$$

This function applies for the 3PLM, the 2PLM ($\gamma_i = 0$), and the 1PLM ($\gamma_i = 0$, $\alpha_i = 1$).

The *test information function* is simply the sum of all item information functions:

$$TInf(\theta) = \sum_{i=1}^{I} Inf_i(\theta). \tag{8}$$

Finally, it can be shown that

$$SE(\theta) = \frac{1}{\sqrt{TInf(\theta)}}. \tag{9}$$

5.3 Small example

Let's focus on a small toy example (Embretson & Reise, 2000). We want to estimate θ based on the 2PLM and three items, as shown below:

The item response functions look like this (see Equation (2)):

The likelihood function is given by

$$L(\theta_s | X_s = (1,1,0)) = P_{s1} P_{s2} (1 - P_{s3}), \tag{10}$$

where:

- $P_{s1} = \frac{\exp[1.0(\theta - (-1.5))]}{1 + \exp[1.0(\theta - (-1.5))]}$,

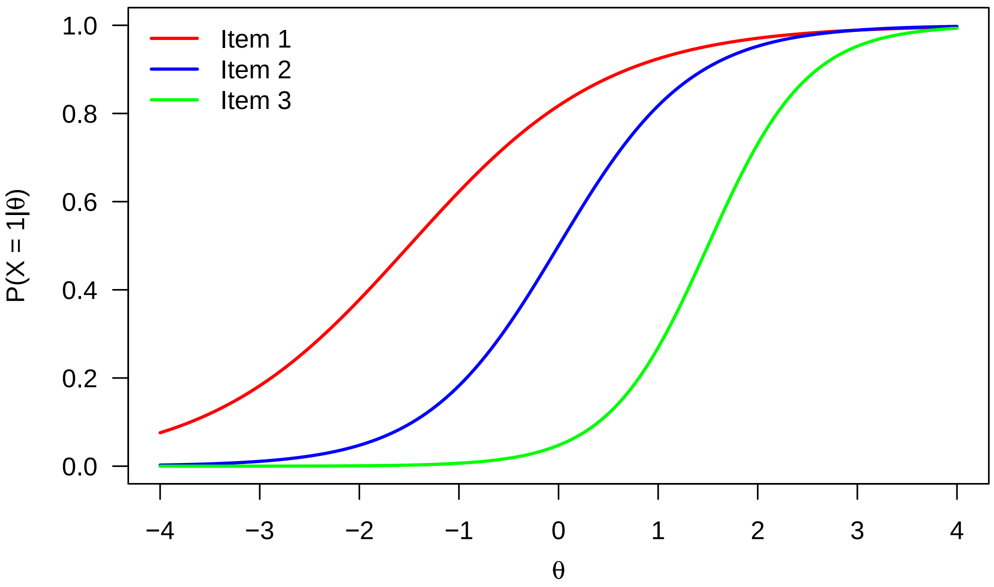

Figure 5: IRFs from the toy example

- $P_{s2} = \frac{\exp[1.5(\theta - 0.0)]}{1 + \exp[1.5(\theta - 0.0)]}$,

and

- $P_{s3} = \frac{\exp[2.0(\theta - 1.5)]}{1 + \exp[2.0(\theta - 1.5)]}$.

The log-likelihood function is given by

$$\ln L(\theta_s | X_s = (1,1,0)) = \ln(P_{s1}) + \ln(P_{s2}) + \ln(1 - P_{s3}). \tag{11}$$

The plot of the likelihood function in Equation (10) is displayed in Figure 6 (top panel), whereas the log-likelihood function in Equation (11) is displayed in Figure 6 (bottom panel).

The MLE for parameter θ_s is the value of θ that maximizes the likelihood/ log-likelihood function. Observe that the same θ value (slightly below 1) maximizes both the likelihood and the likelihood functions simultaneously; this is indicated in both panels by the vertical red dashed line. This is always the case. Again, the log-likelihood is used for numerical convenience.

The seeked value is $\widehat{\theta}_s = .838$. This value is found by the Newton-Raphson method (initial θ value: 0; convergence threshold: .0001; number of steps required: 5).

Next we look at the item and test information functions and the SE of the person parameter estimate. The three item information functions look as follows (see Equation (7)):

Two important properties of item information functions are immediately visible:

- The item information is maximum at its difficulty level (marked by the vertical dashed lines). This is true for the 1PLM and the 2PLM (not for the 3PLM due to the effect of the pseudoguessing parameter).
- Item information increases with the discrimination parameter.

The test information function (Equation (8)) is the sum of the three item information functions (black curve in Figure 7). It can be seen that this "test" is mostly useful to assess persons with latent ability, say, between 0 and 2 (where the test information function attains high values). This test does not measure persons reliably in the rest of the latent trait because no items were located in these regions.

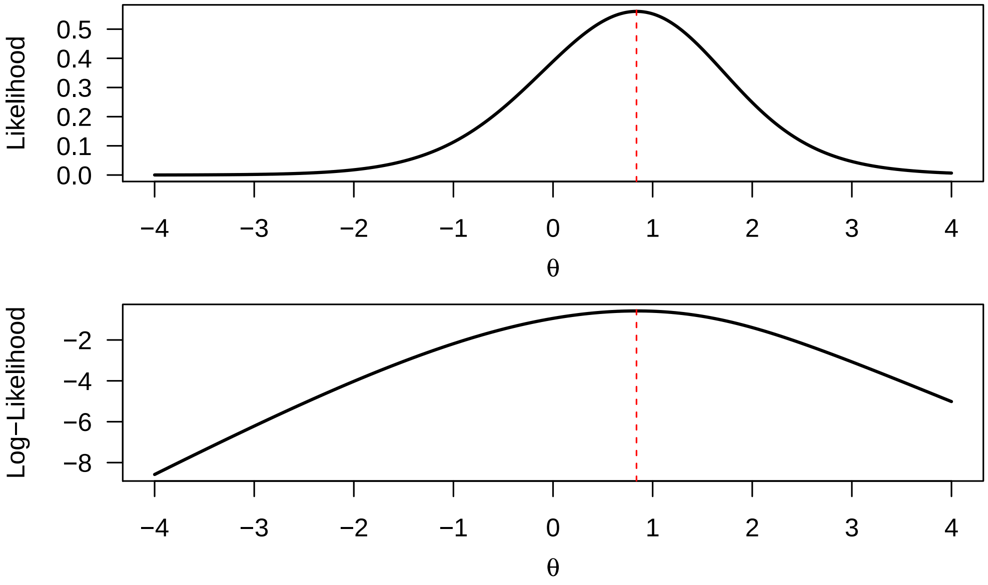

Figure 6: Likelihood and log-likelihood functions from the toy example

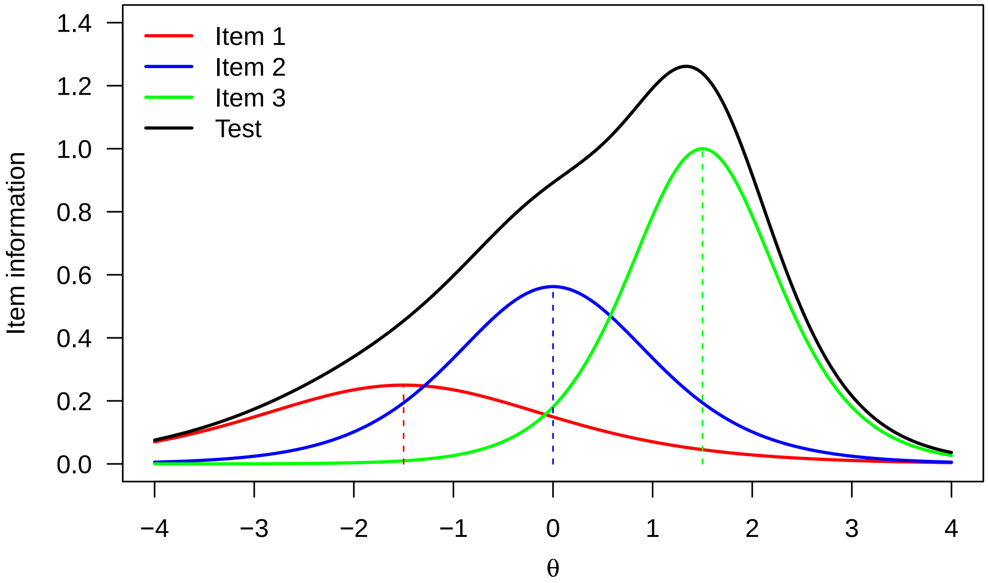

Figure 7: Item and test information functions

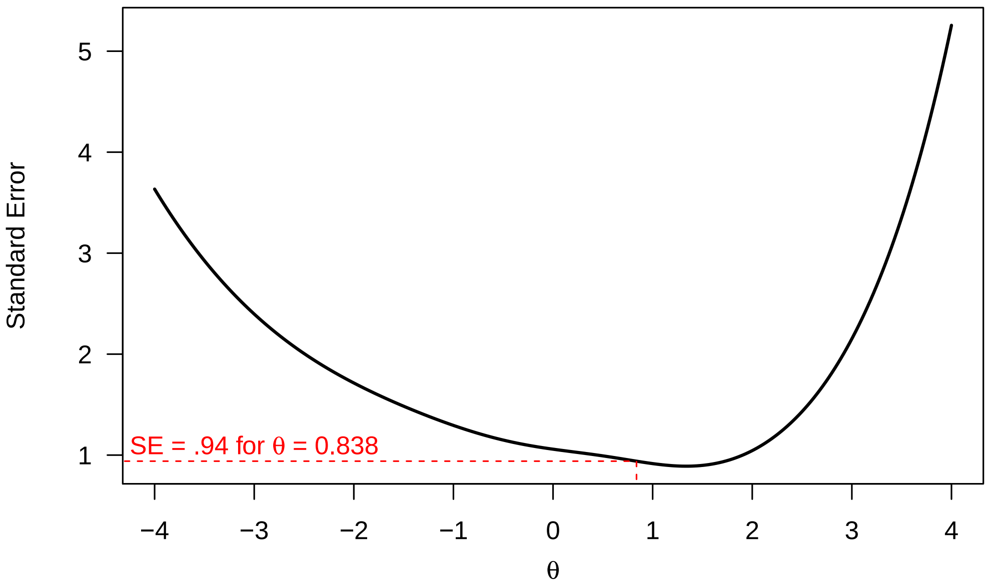

Figure 8: SE of measurement

Finally, the SE conditional on θ (Equation (9)) is shown in Figure 8. Of course, these SE values are unnaceptably large because this toy example is based on three items only. As the number of items increases the measurement precision (and therefore SE) decreases, as expected. That is why a test with more questions is typically more informative than a test with fewer items, all other things being equal. For the person who answered $(1, 1, 0)$ to the three items, we had already seen that $\widehat{\theta}_s = 0.838$. The associated SE is 0.94 (red dashed line).

Some properties of the SE:

- The test information function (Equation (8)), and therefore the SE, does not depend on the particular persons who took the test. This is unlike CTT and its reliability coefficient.
- The SE changes with θ: The test measures persons with more precision (i.e., lower SE) in the latent scale regions with the highest information. This is also unlike CTT, for which the SE is constant for all person scores.

5.4 Model indeterminacy

It should be said that IRT models are typically not identified. That means that there are infinite sets of item and person parameters that lead to the exact same model fit. This is easy to understand by considering the 2PLM. Observe that

$$P(X = 1|\theta) = \frac{\exp[\alpha(\theta - \delta)]}{1 + \exp[\alpha(\theta - \delta)]} = \frac{\exp\{\alpha[(\theta + \lambda) - (\delta + \lambda)]\}}{1 + \exp\{\alpha[(\theta + \lambda) - (\delta + \lambda)]\}}.$$

Visually, this means that we can move all persons and items together left or right on the latent scale by an amount given by λ without affecting the probability of a correct answer. This is a *location* indeterminacy. There is also a *scale* indeterminacy: α can be multiplied by any constant as long as $(\theta + \lambda)$ and $(\delta + \lambda)$ are divided by that same constant.

To solve this problem, model constraints need to be introduced that solve both the *location* and *scale* indeterminacies.

Table 2: Maximum likelihood scoring

Person	Resp. pattern	Theta (Test A)	SE (Test A)	Theta (Test B)	SE (Test B)
1	1111100000	0.00	0.54	-0.23	0.43
2	0111110000	0.00	0.54	-0.13	0.43
3	0011111000	0.00	0.54	-0.04	0.42
4	0001111100	0.00	0.54	-0.04	0.42
5	0000111110	0.00	0.54	0.13	0.43
6	0000011111	0.00	0.54	0.23	0.43
7	1110000000	-0.92	0.57	-0.82	0.51
8	0111000000	-0.92	0.57	-0.74	0.50
9	0011100000	-0.92	0.57	-0.66	0.48
10	0001110000	-0.92	0.57	-0.59	0.47
11	0000111000	-0.92	0.57	-0.53	0.46
12	0000011100	-0.92	0.57	-0.46	0.45
13	0000001110	-0.92	0.57	-0.4	0.45
14	0000000111	-0.92	0.57	-0.34	0.44
15	1000000000	-2.20	0.78	-1.8	0.88
16	1100000000	-1.47	0.63	-1.2	0.62
17	1110000000	-0.92	0.57	-0.82	0.51
18	1111000000	-0.45	0.55	-0.51	0.46
19	1111100000	0.00	0.54	-0.23	0.43
20	1111110000	0.45	0.55	0.04	0.42
21	1111111000	0.92	0.57	0.34	0.44
22	1111111100	1.47	0.63	0.71	0.49
23	1111111110	2.20	0.78	1.28	0.65

5.5 MLE properties

MLE enjoys some good asymptotic (large sample) properties, namely:

- Bias: MLEs converge in probability to the true parameter;
- Efficiency: MLEs have the smallest mean square error among all consistent estimators;
- Residuals are normally distributed.

However:

- $\widehat{\theta}_s$ does not exist for all-0s response patterns (e.g., $(0, 0, 0, 0)$) or all-1s response patterns (e.g., $(1, 1, 1, 1)$).
- The good properties above hold for large samples only.
- The model must be well specified.

5.6 Example

This example is to be found in Embretson and Reise (2000, Chapter 7). The table below shows an example of MLE θ estimates based on known item parameters:

- Test A: All $\alpha = 1.5$, and $\delta = [-2.0, -1.5, -1.0, -0.5, 0.0, 0.0, 0.5, 1.0, 1.5, 2.0]$
- Test B: All $\delta = 0.0$ and $\alpha = [1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9]$

Note that:

- For test A, it does not matter which items are answered correctly. Thus ML scoring is insensitive to the consistency of an examinee's response pattern. This is common to the 1PLM (the number-correct score is a sufficient statistic for θ).
- The SEs are rather large (small test, only consisting of 10 items).
- On test B it is possible to obtain a higher raw score and receive a lower latent trait score (e.g., examinee 18 and examinee 14). This is typical to the 2PLM and the 3PLM.
- Examinees who endorse the items with the highest discrimination parameters receive the highest trait scores.

6 Model fit

There are many fit statistics that can be used at various levels (item fit, person fit, model fit). There is no agreed-upon best fit measure, hence software typically offers a plethora of fit outcomes. IRTPRO, for example, reports the following goodness of fit measures:

- The values of -2 loglikelihood, Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) for model comparison.
- (In some cases) the overall likelihood ratio test against the general multinomial alternative.
- For some models, the M2 statistic (Maydeu-Olivares & Joe, 2005, 2006).
- LD indexes (Chen & Thissen, 1997) to detect violations of local dependence.
- $S - \chi^2$ item fit statistic (Orlando & Thissen, 2000, 2003).

Below we only focus on Orlando and Thissen's item fit statistics.

As a general principle, fit statistics look at residuals. For a given IRT model, the item and ability parameters are estimated. This allows computing model-predicted scores (expected scores). The predicted results are then compared to the observed results. The residuals are the differences between the observed and the expected scores. If a model fits well, the residuals are expected to be small.

Yen (1981) suggested an item fit statistic known as Q_1. It is based on ordering the examinees by their θ estimates and then dividing them in 10 subgroups such that the number of examinees per group is very similar. The formula is

$$Q_{1i} = \sum_{k=1}^{10} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}, \tag{12}$$

where

- O_{ik} is the observed proportion of examinees in subgroup k who answered item i correctly;
- E_{ik} is the expected (from the model) proportion of examinees in subgroup k who should answer item i correctly, computed at the subgroup's mean θ value;
- N_k is the number of examinees in subgroup k.

Yen's Q_1 is very similar to Bock's (1972) χ^2 index (Equation 9.3 in Embretson & Reise, 2000), except that Bock's statistic did not rely on a fixed number of subgroups and the E_{ik} values were computed at the median (instead of the mean) θ value for subgroup k.

One problem with the Q_1 index is that it depends on the estimated θ values to obtain the observed proportions in each subgroup. That is, the observed proportions depend on the model fit. This is undesirable. Moreover, the number of subgroups and the cutoff values between subgroups are arbitrary and may affect the results.

Orlando and Thissen (2000) proposed grouping examinees based on the observed data only (thus, not on θ). The formula is given by

$$S - X_i^2 = \sum_{k=1}^{I-1} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}. \tag{13}$$

Now the groups are based on the observed number-correct score, that is, the distribution of the total number of correct answers in the sample of examinees. The observed proportions O_{ik} are straightforwardly computed. The expected proportions E_{ik} are more difficult to compute (see Orlando & Thissen, 2000, for the formulas). This statistic is χ^2 distributed with $(I - 1 - m)$ degrees of freedom (I = number of items, m = number of item i's parameters). As usual in χ^2 tests, cells with low frequencies need to be collapsed (and the degrees of freedom accordingly adjusted).