A Short Primer on Item Response Theory

Ricardo J. Romeu

The Basics of Item Response Theory

In this section, I give a brief overview of item response theory; the interested reader can consult a number of textbooks for more details (De Ayala, 2013; Embretson & Reise, 2000; Levy & Mislevy, 2017). The basic premise of item response theory (IRT) is to model how a person interacts with items on a survey or inventory; the model generates probabilities of responses by linking *person* and *item* parameters together.

Consider an item with n levels (typically called a polytomous item); for instance, the item might be ”Please indicate your agreement or disagreement with the following statement: I believe that certain people exist solely to harm me” and the $n = 5$ levels might go from ”I don’t believe it at all” (1) to ”I am absolutely certain of this” (5). Item response theory starts from the assumption that there is at least one *latent trait* that dictates the responses given to a set of items. For instance, this item could be part of a scale meant to measure delusion-proneness, and so our model would relate this trait of delusion-proneness to the possible responses on the item. (Note that the content of the scale defines the meaning of the latent trait.) However, we cannot determine with certainty which level of this item will be endorsed, and so our model gives a *probability distribution* over the set of levels on the item, and the latent trait determines what this probability distribution will be.

We can mathematically represent this as follows: let $[p_1(\theta)...p_n(\theta)]^T$ denote a vector where entry j is the probability of endorsing level j. Then an item response model is a function mapping a real value (the latent trait θ) to a vector of probabilities, as so:

$$\theta \mapsto \begin{bmatrix} p_1(\theta; a_n, b_n) \\ ... \\ p_n(\theta; a_n, b_n) \end{bmatrix}. \tag{1}$$

Equation (1) also introduces the *item parameters* $\{a_n\}$ and $\{b_n\}$ associated with this particular question. These item parameters give us information about how this particular item relates the latent trait to the probability of endorsement for each level. These item parameters are crucial for understanding how the scale is ”behaving” with respect to our sample (De Ayala, 2013; Embretson & Reise, 2000; Levy & Mislevy, 2017).

As a concrete example, consider a question with only 2 levels, i.e. a binary question. A popular model for this kind of item is the ”2PL” model, or the two-parameter logistic model (Embretson & Reise, 2000). Such a model can be represented as:

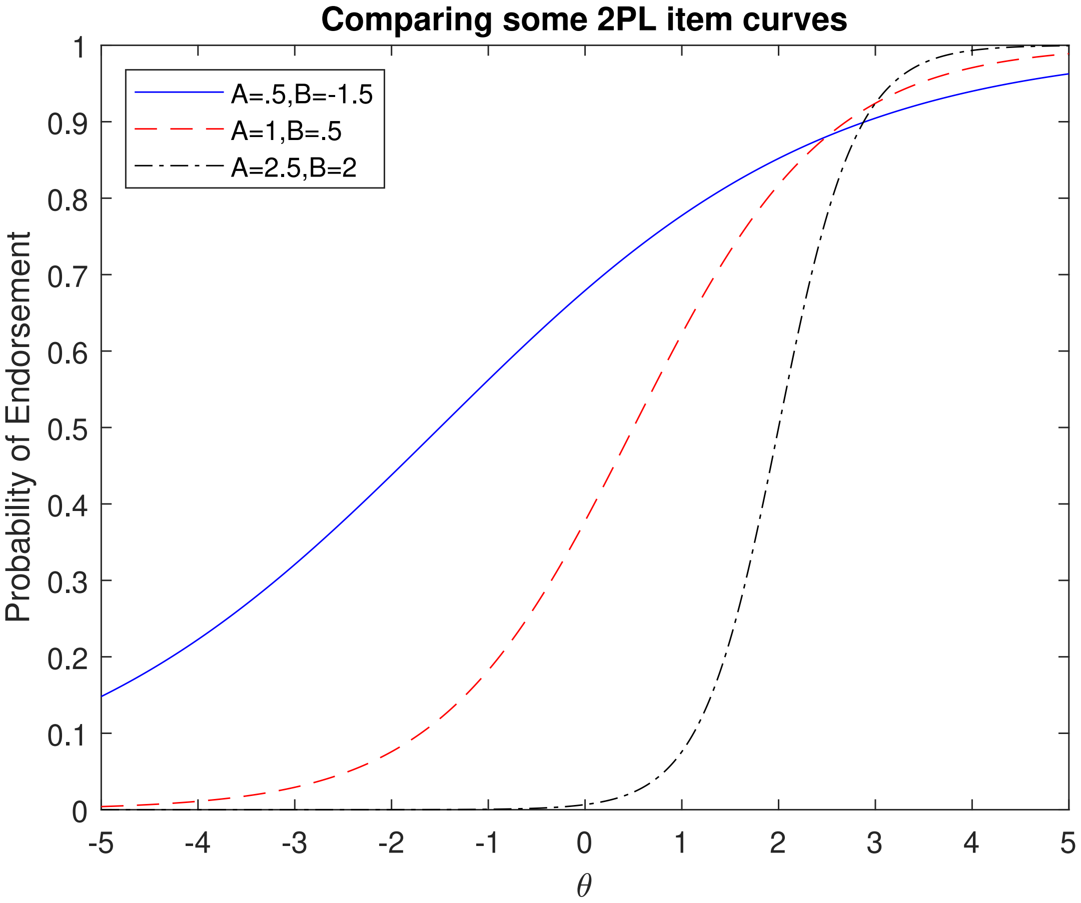

Figure 1: Some examples of two-parameter logistic item response curves. These functions give, at every latent trait θ, the probability of endorsing the item. The left-most function is slowest to ascend and is ”centered” at a negative θ value; this indicates that the item is not very discriminating and requires little of the latent trait to have a high probability of endorsement. In contrast, the right-most function is highly discriminating and requires a great deal of the latent trait to have a high probability of endorsement.

$$\theta \mapsto \begin{bmatrix} 1 - \frac{1}{1+\exp(-a(\theta - b))} \\ \frac{1}{1+\exp(-a(\theta - b))} \end{bmatrix}. \tag{2}$$

A few graphical examples of Equation (2) can be found in Figure 1.

Some examples of this kind of item response model can be found in Figure 1. Here, the probability of endorsing the item (or, marking it as ”True”) is given by the second entry in the probability vector. The item parameter $b \in \mathbb{R}$ is the difficulty of the scale, and corresponds to the latent trait value at which endorsement and non-endorsement are equally likely (Embretson & Reise, 2000). It can intuitively be thought of as where the logistic function is ”centered” on the scale. The second item parameter $a > 0$ is the discrimination of the item, and it is defined as the slope of the logistic function at the difficulty b: $f'(b) = a$, where f' denotes the derivative of the logistic function (Embretson & Reise, 2000). The larger the value of a, the closer the function f is to a step function; a step function perfectly separates persons into those with a ”lower” trait (traits less than b) and those with a ”higher” trait (traits greater than b), and so a larger a means the item is better able to separate low from high levels of the trait.

The item response model can be generalized in a number of ways to polytomous items; the version we focus on here is Samejima’s Graded Response Model (De Ayala, 2013; Embretson & Reise, 2000; Levy & Mislevy, 2017; Samejima, 2016), which is the model we employ in our Bayesian approach below. The basic idea of Samejima’s model is to use the models we have for the binary case and employ that in the polytomous case. This entails finding a hidden set of binary choices among the Likert scale levels. Samejima did this not by modeling the probability of endorsing a particular level directly, but by modeling the probability of crossing thresholds that were postulated to exist between the different levels. (Figure 2 gives an example of this model.) To better understand this, imagine partitioning the latent scale into n bins (which involves $n-1$ thresholds). Then a person will endorse level k if that person’s latent trait surpasses the threshold separating levels $k-1$ and k, but *does not* surpass the threshold separating levels k and $k+1$. This places that person’s latent trait firmly within bin k[1]. Letting X_i denote the response on item i, we can summarize this as:

$$P(X_i = k) = P(X_i \geq k) - P(X_i \geq k+1). \tag{3}$$

[1] Samejima (1972) poetically described this process as the level ”attracting” the person, and the person deciding to ”reject” or ”accept” this atrtaction.

This means that Samejima finds the hidden binary response within the crossing of the thresholds (given as survivor probabilities, as Equation (3) suggests), and calculates the probabilities of each level as crossing one threshold but not the next one. Each threshold curve can then have its own set of item parameters; however, in order to keep the number of parameters to a minimum, it is usually assumed that each threshold curve has the same discrimination (a; known as the *homogenous* case; Samejima 1969, 1972, 2016). A graphical example of Samejima's model, with 5 levels, is given in Figure 2.

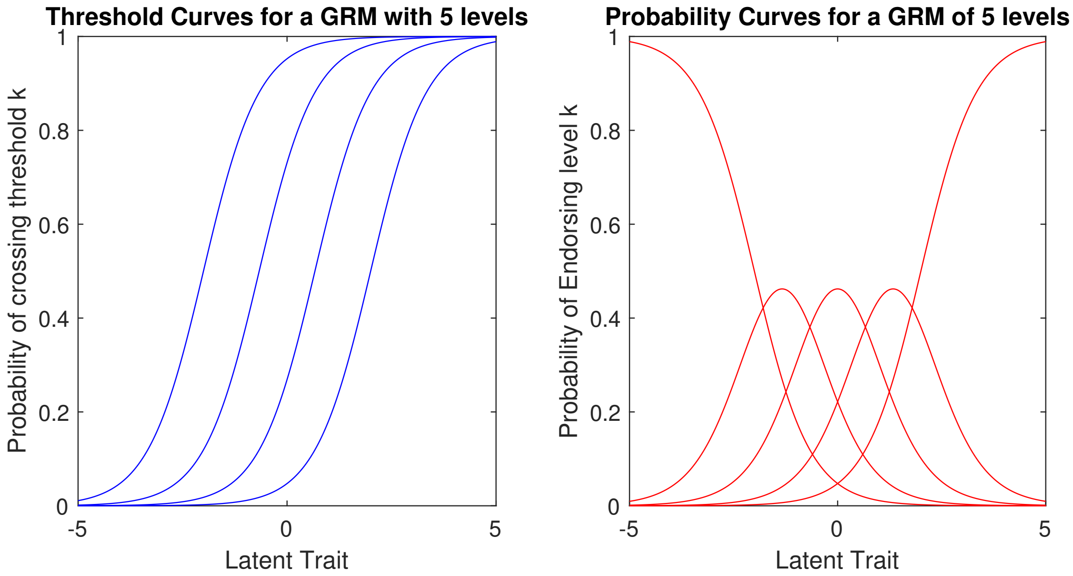

Figure 2: The Samejima (Graded Response) Model, separated into its threshold curves (left) and its probability curves (right). The threshold curves give the probability that a person will cross from one level of the scale to the next level, given the latent trait. The probability curves are (mostly) obtained by taking the difference of successive threshold curves, as they are survivor functions. The exceptions are the probability curves for the lowest (1 minus the first threshold curve) and the highest (the last threshold curve minus 0) levels. Note that the threshold curves are equally spaced in the left panel; this is the simplifying assumption that is not always accepted.

If the number of items on the scale is relatively large, one simple way to summarize the scale is to look at the *total information curve* of the scale, found by summing the (Fisher) information for each item (De Ayala, 2013; Embretson & Reise, 2000; Samejima, 1969, 1994). This summing is justified under the commonly held assumption of *local independence,* which assumes that the responses to each item are independent given the latent trait (Embretson & Reise, 2000). The Fisher information is related to the variance in our estimate of the parameter in question; in particular, the *standard error of measurement* is (asymptotically) inversely related to the (square of the) total information: $1/\sqrt{I(\theta)} \rightarrow SEM(\theta)$. Note that the standard error of measurement is a function of

θ (Embretson & Reise, 2000; Samejima, 1994). The general Fisher information function for a polytomous item with n levels is given by the formula:

$$I(p(\theta)) = \sum_{j=1}^{n} \frac{(p_j'(\theta))^2}{p_j(\theta)}, \tag{4}$$

where $p(\theta) = \Big[p_1(\theta), ..., p_n(\theta)\Big]$ is our probability distribution over the n levels (De Ayala, 2013; Samejima, 1969). At a glance, the total information curve tells us where on the latent trait our estimates are strongest, which is an indication of where the scale is most useful.

In sum, item response theory aims to relate the hypothesized latent trait to probabilities of responses on the scale. It does this by assuming particular functional forms for these probabilities, and these functional forms come with item parameters that can help us better understand how the scale is functioning with our particular sample of respondents. We can summarize the results of our model fitting by observing the total information curve of the scale; recalling that the total information is related to our standard error of measurement for θ, our latent trait, we can see that the total information function can clue us to where on the latent trait the scale is most useful.

Appendix

In this section, I present my proof of Equation (4), the item information curve for a general polytomous item. I cannot claim to have originally found this equation, but I present a short proof that works for a larger class of item response curves, since I do not make assumptions on the functional form of the curves and only minimal assumptions on differentiability (Samejima, 1969, 1972, 1994). Samejima (1969) began her proof with a simplification of the Fisher Information function based on the notion of local independence (p. 37, Chap 6); however, this proof here begins from the definition of Fisher Information, and as a result, the probability functions need only be continuously differentiable, instead of twice differentiable. The trick to the proof is to think of the n level item as being a vector in an appropriate Hilbert space. We begin by precisely stating the theorem.

Let $p = \Big[p_1(\theta), ..., p_n(\theta)\Big] \in \mathbb{R}^n$ be a vector of continuously differentiable functions with $p_j : S \subset \mathbb{R} \to \mathbb{R}$, $p_j(\theta) > 0$ for all $\theta \in \mathbb{R}$, and $\sum_{j=1}^{n} p_j(\theta) = 1$ for all $\theta \in S \subset \mathbb{R}$. Then the Fisher Information function for p is given by $I(\theta) = \sum_{j=1}^{n} \frac{(p_j'(\theta))^2}{p_j(\theta)}$. Moreover, $I(\theta)$ is continuous on the domain of p.

Proof. Fisher information is defined to be the following:

$$I(\theta) = E[\{\frac{\partial}{\partial\theta}\log(f(X;\theta))\}^2|\theta], \tag{5}$$

where $f(X;\theta)$ is the likelihood of the data X, with parameter θ (De Ayala, 2013). This means our first step is to obtain a general likelihood function for the n level Likert scale.

To obtain this function, suppose we have an item with n levels, and these levels take on values $\{a_1,...,a_n\}$. That is, if we let the random variable X denote the value chosen on this item, then we have $X = a_j$ with probability $p_j(\theta)$, for $j = 1,...,n$. We can then represent the likelihood function as follows:

$$f(X;\theta) = p_1(\theta)^{\chi_{\{X=a_1\}}}...p_n(\theta)^{\chi_{\{X=a_n\}}} = \Pi_{j=1}^{n} p_j(\theta)^{\chi_{\{X=a_j\}}}, \tag{6}$$

where, in general, the characteristic function χ_A equals 1 if the event A occurs and is 0 otherwise. So, for example, if we happen to observe that $X = a_3$, say, then all the χ's with $a_j \neq a_3$ will be zero, leaving only $p_3(\theta)$ leftover. The log-likelihood will then be given by:

$$\log(f(X;\theta)) = \sum_{j=1}^{n} \chi_{\{X=a_j\}} \log(p_j(\theta)). \tag{7}$$

Now, if we take the partial derivative with respect to θ, we obtain (since χ_A is a constant relative to θ for all events A):

$$\frac{\partial}{\partial\theta}\log(f(X;\theta)) = \sum_{j=1}^{n} \chi_{\{X=a_j\}} \frac{p'_j(\theta)}{p_j(\theta)}. \tag{8}$$

Here is where we make a critical observation. Suppose we have two vectors $x, y \in \mathbb{R}^n$, i.e., two vectors in the n-dimension Euclidean space $\mathbb{R}^n$. Then we have their inner product defined as: $\langle x, y\rangle = x_1y_1 + ...x_ny_n$, i.e., multiply component-wise and then sum up the products (Conway, 2013). The reader should notice that the above equation is just the inner product between two vectors in $\mathbb{R}^n$, for every θ : the first vector is $\chi = \left[\chi_{\{X=a_j\}},...,\chi_{\{X=a_n\}}\right]$ and the second vector is $q = \left[q_1(\theta),...,q_n(\theta)\right]$, where both vectors are in $\mathbb{R}^n$ and $q_j(\theta) = p'_j(\theta)/p_j(\theta)$. Using the bilinearity of the inner product, we can then simplify the squaring of this sum using the following calculation:

$$L(\theta) = (\frac{\partial\log(f(X;\theta))}{\partial\theta})^2 = \langle\chi, q\rangle^2 = \langle\chi, q\rangle\langle\chi, q\rangle = \langle\langle\chi, q\rangle\chi, q\rangle, \tag{9}$$

which is justified since $\langle x, y\rangle \in \mathbb{R}$ for $x, y \in \mathbb{R}^n$.

Now let us focus on simplifying $\langle\chi, q\rangle\chi$. Since $\langle\chi, q\rangle$ is a real number, we can distribute this to each element in the vector χ, such that the kth entry of $\langle\chi, q\rangle\chi$ is given by:

$$\chi_k = (\sum_{j=1}^{n} \chi_{\{X=a_j\}} q_j(\theta))\chi_{\{X=a_k\}} = \sum_{j=1}^{n} \chi_{\{X=a_k\}}\chi_{\{X=a_j\}} q_j(\theta)), \tag{10}$$

where we have introduced the shorthand χ_k to denote the kth element of χ. Note that X can only take one of the n values $\{a_j\}_{j=1}^{n}$, and so the product $\chi_k\chi_j = 1$ if and only if $j = k$, and so the above equation reduces to $\chi_k = \chi^2_{\{X=a_k\}} q_k(\theta)$, since all other combinations will always turn up zero no matter what value X takes. Since for every characteristic function, $\chi_A^2 = \chi_A$ (Conway, 2013), we can then see that:

$$\langle\chi, q\rangle\chi = \left[\chi_1 q_1(\theta), ..., \chi_n q_n(\theta)\right]. \tag{11}$$

Finally, we take the inner product one more time to arrive at our squared sum:

$$L(\theta) = \langle\langle\chi, q\rangle\chi, q\rangle = \sum_{j=1}^{n} \chi_j^2 q_j^2 = \sum_{j=1}^{n} \chi_j (\frac{p'_j(\theta)}{p_j(\theta)})^2. \tag{12}$$

We arrive at our final equation by taking the (conditional) expected value of $L(\theta)$:

$$E[L(\theta)|\theta] = E[\sum_{j=1}^{n} \chi_j (\frac{p'_j(\theta)}{p_j(\theta)})^2|\theta] = \sum_{j=1}^{n} E[\chi_j (\frac{p'_j(\theta)}{p_j(\theta)})^2|\theta] = \sum_{j=1}^{n} \frac{(p'_j(\theta))^2}{p_j(\theta)} = I(\theta), \tag{13}$$

since $E[\chi_j|\theta] = p_j(\theta)$ and $q^2(\theta)$ is a constant when θ is given. Of course, as p was assumed continuously differentiable, this means p' is continuous; as $p > 0$ for all θ, this makes the composition $(p')^2/p$ continuous; finally, as the finite sum of continuous functions is again continuous, $I(\theta)$ is then continuous wherever p is. QED. □

References

Conway, J. B. (2013). *A course in functional analysis* (Vol. 96). Springer Science & Business Media.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.

Levy, R., & Mislevy, R. J. (2017). *Bayesian psychometric modeling*. Chapman and Hall/CRC.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*.

Samejima, F. (1994). Some critical observations of the test information function as a measure of local accuracy in ability estimation. *Psychometrika*, *59*(3), 307–329.

Samejima, F. (2016). Graded response models. In *Handbook of item response theory, volume one* (pp. 123–136). Chapman and Hall/CRC.