A brief introduction to Item Response Theory

Amanda Luby

Background

Item Response Theory (IRT) is used extensively in psychometrics and educational testing theory to study the relationship between a respondent's (unobserved) ability and their performance on a certain item. For analyzing test results using an IRT approach, the probability of a correct response to a question depends on both a) the difficulty of the question and b) the ability of the respondent. Both 'difficulty' and 'ability', however, are not observable and so we must estimate them using the available data. IRT allows for participants to be compared on the same scale, even if they were shown different sets of questions on an exam. Item Response Models can account for varying difficulty of questions, and adjusts each individual's ability measure accordingly.

These models can be used for descriptive measurement purposes, such as to determine which questions are more difficult than others, or to compare the ability of respondents. IRT can also be used for explanatory purposes, where person characteristics or question properties are incorporated. With an explanatory approach, we can determine, for instance, if there are certain properties that make some questions harder than others.

It is helpful to represent the data as a matrix of binary responses, where 1 indicates a correct answer and 0 represents an incorrect answer. In the context of proficiency exams, if we have n participants and m different questions, we can then express the data as a $n \times m$ matrix of participant responses to the exam:

$$Y = \begin{bmatrix} 1 & - & - & \dots & 0 \\ - & 0 & - & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ - & 1 & 1 & \dots & - \end{bmatrix},$$

where $Y_{pi} = 1$ (row p, column i) if participant p correctly answered question i, $Y_{pi} = 0$ if participant p incorrectly answered question i, and $Y_{pi} = -$ indicates that participant p was not shown question i. For example, the first row of Y corresponds to the responses of participant 1. The first column of the first row, $Y_{1,1}$ is equal to one, so they correctly answered the first question. Both $Y_{1,2} = -$ and $Y_{1,3} = -$, meaning that person 1 was not shown questions 2 or 3. Finally, $Y_{1,m} = 0$, designating that person 1 incorrectly answered question n (the last question). This matrix representation allows for all responses, organized by both respondent and question index, to be included in a single data structure.

Rasch Model

We will denote participant proficiency (ability) as θ_p and question difficulty as b_i. Questions with higher values of b_i are harder items, where participants with low proficiencies have only a small probability of getting the answer right. Easier questions have lower values of b_i, and all participants (even those with low proficiencies) are likely to get the question right.

The *Rasch model* is the most basic item response model, and relates the participant proficiencies and

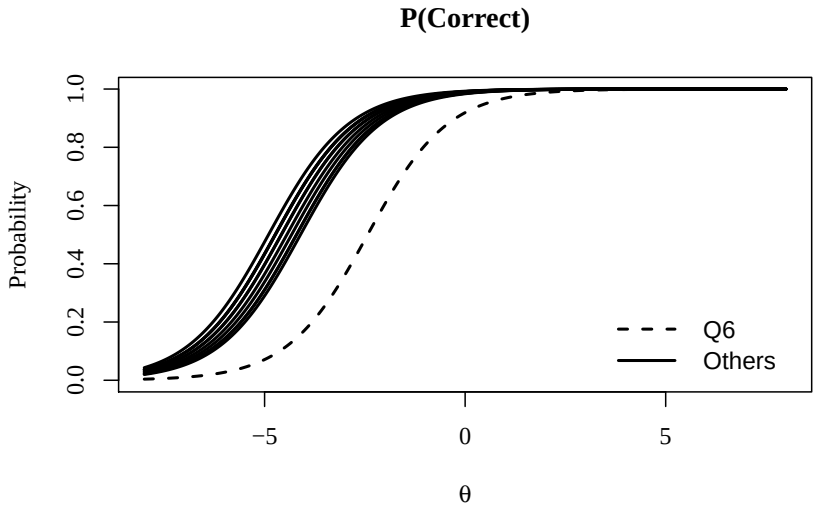

Figure 1: Item Characteristic Curves (ICC) for each question from Latent Print Examination Test No. 16-515/516 under the Rasch model.

question difficulties with a logistic function:

$$P(Y_{pi} = 1) = \frac{1}{1 + \exp(\theta_p - b_i)}. \tag{1}$$

A feature of the Rasch model is that when $\theta_p = b_i$, the probability of a correct response is 0.5. Intuitively, this is what we would expect when both persons and items are located on the same scale. If a person is shown an item with a difficulty exactly equal to their proficiency, they are equally likely to answer the question correctly or incorrectly.

Results from IRT models can be summarized using an *Item Characteristic Curve* (ICC), which represents the probability of a correct response to a question over all possible values of θ. ICCs allow visualization of the differences between questions. Figure 1 illustrates the ICCs for a latent print proficiency exam. Most of the questions have curves that are very close to one another, while question six is further away (designated with the dashed line). Since the ICC for question six is located to the right of the other questions, respondents must have a higher proficiency (θ) in order to increase their probability of correctly answering question six compared to the other questions.

There are two approaches we can take to increase the complexity of the Rasch model. The first is to add additional latent (unobservable) variables for the questions. This increases the amount of computational complexity, and large amounts of data are needed to estimate all of the parameters. The second approach, which was alluded to at the beginning of this document, is to incorporate observable features of respondents or questions into the model. We will first introduce two extensions to the Rasch model using the first approach, and then discuss explanatory modeling using the second approach.

Extension 1: Additional latent variables

The *Two-parameter logistic model* (2PL) includes a discrimination parameter (a_i) for each of the questions:

$$P(Y_{pi} = 1) = \frac{1}{1 + \exp(-a_i(\theta_p - b_i)}. \tag{2}$$

Higher values of a_i denote a higher discrimination, meaning that respondents with a lower proficiency have a much lower chance of correctly answering the question than respondents with a higher proficiency. A higher discrimination also corresponds to an ICC with a steeper slope, while lower discriminations lead to flat ICCs. This is illustrated in Figure 2 with the Rasch model in black and the 2PL models in blue, all with difficulty $b = 0$ for simplicity. The light blue line corresponds to a lower discrimination, $b = 0.5$, while the dark blue line corresponds to a higher discrimination, $b = 2$. The dark blue line discriminates between participants with proficiency less than 0 and participants with proficiency greater than 0 better than both the Rasch model and the 2PL model with $b = 0.5$, evidenced by the curve increasing from 0 to 1 more rapidly.

It is up to the test constructor whether it is more desirable to have questions that sharply discriminate between two levels of proficiency, or questions that are less discriminating but cover a greater range of proficiencies. If all of the questions are highly discriminating but are near the same difficulty, for instance, we will only know whether respondents have a proficiency above or below that difficulty rather than a broader range of estimates.

The *Three-parameter logistic model* (3PL) includes a pseudo-guessing parameter (c_i) for each of the questions:

$$P(Y_{pi} = 1) = c_i + \frac{1 - c_i}{1 + \exp(-a_i(\theta_p - b_i)}. \tag{3}$$

Including c_i increases the "base probability" of correctly answering a question. For instance, in a classic multiple choice question, even respondents with extremely low proficiencies will sometimes get a question right by chance. There may be options in a multiple choice question that are obviously incorrect, to the point where even low-proficiency respondents will not choose that option, which may not be known before the exam is administered. Using an IRT approach accounts for these effects by using the observed data to estimate c_i, rather than determining the value of c_i from the question itself.

We have illustrated 3PL models in Figure 2 with green lines. These models have the same difficulty ($b = 0$) and discriminations ($a = 0.5$ and $a = 2$) as the other models shown in the graph, but include a guessing parameter of $c = 0.3$. The light green curve is the same difficulty and discrimination as the light blue line, with the guessing parameter included, and the dark green curve is the same difficulty and discrimination as the dark blue curve, with the guessing parameter included. The 3PL models have the same shape as their corresponding 2PL models, but are shifted so that even participants with low proficiencies have at least a 30% chance of getting the answer correct.

Extension 2: Additional observed variables

Rather than increasing the complexity of the model using latent variables such as question discrimination, we can incorporate observed variables that describe the respondents or the questions. We return to the Rasch model (Equation 2), but instead of adding additional latent variables to the model itself, we take a multilevel

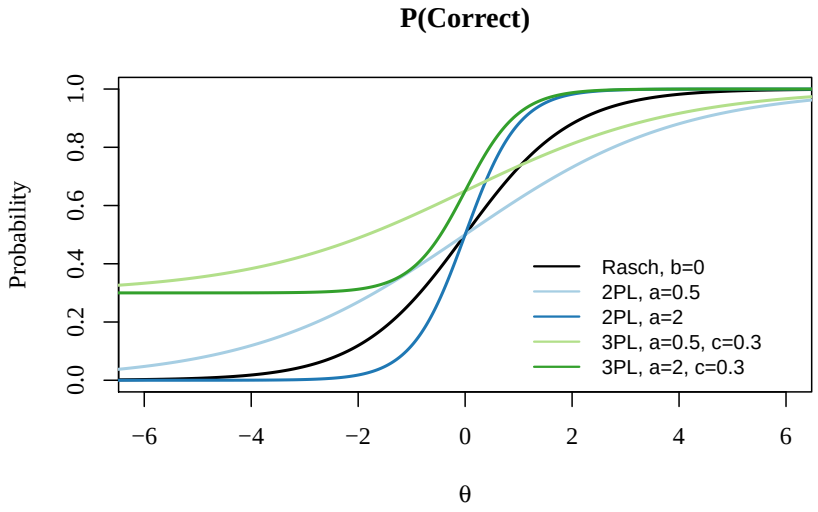

Figure 2: ICCs for five models, all with difficulty $b = 0$. The 2PL models are shown in blue and the 3PL models with guessing parameter $c = 0.3$ are in green. Darker shades correspond to discrimination $a = 2$ while lighter colors correspond to discrimination $a = 0.5$

model approach and give the proficiency (θ) and difficulty (b_i) parameters their own equations.

For instance, we could model the relationship between the difficulties of the questions, b_i, and variables such as the number of minutiae, whether the core and delta are visible, or a quality metric score with a linear regression model. That is,

$$b_i = \beta_0 + \beta_1 \cdot \texttt{Minutiae} + \beta_2 \cdot \texttt{CoreVisible} + \beta_3 \cdot \texttt{DeltaVisible} + \beta_4 \cdot \texttt{QualityMetric} + \epsilon_i$$

Each of `Minutiae`, `CoreVisible`, `DeltaVisible`, and `QualityMetric` are observable, and $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 can be estimated using standard statistical methods such as linear regression. Whatever variation in b_i is not explained by the observed variables is expressed in the residual term, ϵ_i. If ϵ_i is large compared to the estimated βs, then the observed variables do not explain the difficulty of the question well.

There are, of course, many alternatives to a linear regression model to explain the difficulty and/or proficiency variables. Transformations and combinations of variables should also be considered. Adding an additional level to the model in which we try to explain θ_p or b_i based on additional observed variables, however, remains the same regardless of which model we choose.

Summary

We have introduced the IRT modeling framework for forensic proficiency exams. The Rasch model is used as the base model for our framework, and we discussed two possible methods of extension. The first is to introduce additional latent variables describing the questions using the 2PL and 3PL models. The second is a multilevel modeling approach to explain the difficulty and proficiency parameters using additional observed variables describing either the questions or participants.