Item response theory

Item Response Theory (IRT)

Item response theory or IRT is one of the psychometric theories. It can be used

- *to evaluate the psychometric properties of an existing scale and its items, to optimally shorten the scale when necessary, and to evaluate the performance of the reduced scale. When used appropriately, IRT modeling can produce precise, valid, and relatively brief instruments resulting in minimal response burden.*[1]

IRT represents a family of mathematical models describing the probability with which examinees with different knowledge levels will provide an correct (or incorrect) answer to an assessment item, and is based on two postulates[2]:

- *The performance of an examine on an item can be predicted (or explained) by a set of factors called traits or latent traits, or abilities.*
- *The relationship between examinees performance and the set of traits can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC).*

Although principles of adapting the test to the examinee was recognized as a good practice in psychological measurement even from the middle of the 20th century (for example Carl Binet's, later Stanford-Binet IQ Test[3]), it was an unscalable time-consuming task. But computers enabled implementation and automation of IRT procedures.

Item Characteristic Curve (ICC)

Every possible test item is described with its *item characteristic curve* (ICC) defined by the function P(θ). This function describes the probability that a subject with the knowledge level of θ will answer the given test item correctly. It is defined on the set of real numbers, but often plotted only on the interval of [-3,3]. There are various kinds of item characteristic curves, depending on how many parameters are included. Most commonly, one, two or three.

The One-Parameter Model

The one-parameter model, also called the Rasch model was introduced in 1960s by Georg Rasch.[4] The equation describing the Rasch model is:

- $P(\theta) = P(u=1|\theta) = 1/(1+e^{\wedge}[-(\theta-b)])$, where

- **P(θ)** is the **probability of a correct response** to an assessment item for the skill/knowledge level of θ
- **θ** is the **ability level**
- **u=1** indicates that the student has given the **correct answer** to the item, and u=0 indicates an incorrect answer, and
- **b** is the difficulty factor.

The Two-Parameter Mode

The two-parameter model or the logistic function, which was first used as a model to describe the item characteristic curve in the 1950s is defined as:
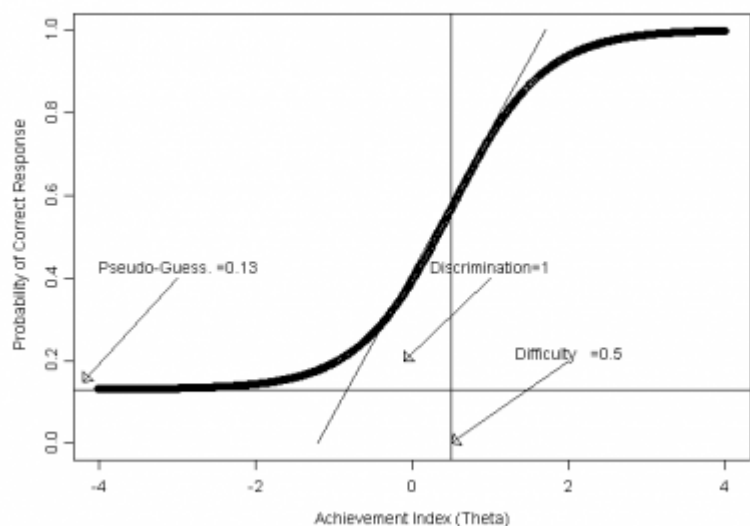
- $P(\theta) = P(u=1|\theta) = 1/(1+e^{-a(\theta-b)})$, where

- **P(θ)** is the **probability of a correct response** to an assessment item for the skill/knowledge level of θ
- **θ** is the **ability level**
- **u=1** indicates that the student has given the **correct answer** to the item, and u=0 indicates an incorrect answer, and
- **a** is the **discrimination factor**, describing how good the item can distinguish between subjects with different level of knowledge,
- **b** is the difficulty factor.

The Three-Parameter Model

A component that accounts for the probability of guessing the correct answer was introduced into the two-parameter model in the 1968 by Birnbaum, thereby creating the three-parameter model described with the following formula:

- $P(\theta) = P(u=1|\theta) = c+(1-c)/(1+e^{-a\cdot(\theta-b)})$

Often, another extension in the formula is used[5)6)]:

- $P(\theta) = P(u=1|\theta) = c+(1-c)/(1+e^{-D\cdot a\cdot(\theta-b)})$, where

- **P(θ)** is the **probability of a correct response** to an assessment item for the skill/knowledge level of θ
- **u=1** indicates that the student has given the **correct answer** to the item, and u=0 indicates an incorrect answer

- **D** is a **constant** usually with value of 1.7,
- **a** is the **discrimination factor**, describing how good the item can distinguish between subjects with different level of knowledge and usually having the value of the maximal slope of the ICC, theoretically any real number but usually limited to [-2.80,2.80],
- **b** represents **difficulty**, theoretically also any real number value, but usually restricted to [-3,3] and
- **c** is the **guessing factor** or *pseudo-chance level*, suggesting how likely is it to occur that a student without any knowledge will provide the correct answer to the item, theoretically ranging from [0,1], but in practice, values above 0.35 are not considered acceptable.

The Item Information Function (IIF)
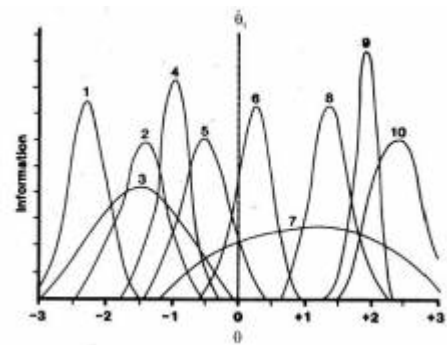
The item information function or just information function provides a measurement of the amount of information provided by an assessment item at a given ability level. The item information function is usually calculated as:

- $I = 1/\sigma^2 = 1/[P(\theta)*(1-P(\theta))]$

On the IIF graph, the location of the center of the IIF reflects the difficulty of the item, the height of the IIF reflects the item discrimination, and its asymmetry reflects the magnitude of the pseudo guessing parameter.[7]

In the usual CAT approach, at any point of the assessment, the next item that will be presented is the one that will provide most information for the assessed ability level.[8]

In CAT tests

- the test takers do not have to take questions that are to complex or to simple for them[9]

Computer adaptive testing (CAT)

Computer adaptive testing or CAT refers to

- *… the redesign of psychological and educational tests for effective and efficient administration by interactive computers. Its objective is to select, for each examinee, the set of test questions that simultaneously most effectively and efficiently measure that person on the trait… The computer… determines how to begin the test for a given examinee, selects items based on the examinee's scored responses to previous items, and applies one or more rules to terminate an examinee's test.* [10]

IRT-based CAT algorithm

The development of IRT models provided a theoretical basis for principles and procedures of CAT and further their further improvements. A simple CAT algorithm runs as follows:[11])

1. Unless known otherwise, the initial ability level θ is assumed to be in the middle of the scale; θ=0.
2. Define step size S, for example 3 (the larger the value, the sooner the loop coming next will finish).
3. Repeat
    1. Based on the IIF select the item providing most information for the current level θ and present it to the examinee.
    2. If the answer is correct θ=θ+S. Else θ=θ-S.
4. Until a correct and a false answer have been obtained at least once.
5. Repeat
    1. Estimate the a-posteriori probability distribution of the student's knowledge using Bayes' rule

$$\mathrm{P}(\theta|\mathbf{u}) = \prod_{i=1}^{n} P_i(\theta)^{u_i} \left(1 - P_i(\theta)\right)^{(1-u_i)} P(\theta|u_1,\ldots u_{i-1})$$

    2. Estimate the knowledge level as the median of the calculated distribution
    3. Item providing most information for estimated knowledge level (based in the IIF) is presented to the examinee and obtain an answer
6. Until a stopping rule is satisfied (for example: maximum number of presented items is reached, estimated knowledge distribution has small enough variance)

[1]

Edelen, Maria Orlando, and Bryce B. Reeve. Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement. Quality of Life Research 16: 5-18, 2007.
[2]

Hambleton, Ronald K., Hariharan Swaminathan, and H. Jane Rogers. Fundamentals of item response theory. SAGE, 1991.
[3], [8], [10], [11]

Weiss, David J. Computerized adaptive testing for effective and efficient measurement in counseling and education. Measurement and Evaluation in Counseling and Development 37, no. 2: 70-84, July 2004.
[4]

Baker, Frank B. The basics of item response theory. Heinemann, 1985.
[5]

Guzman, E., and R. Conejo. Self-assessment in a feasible, adaptive web-based testing system. IEEE Transactions on Education 48, no. 4: 688- 695, November 2005.
[6], [9]

Matsuoka, Rieko. National College of Nursing, Japan. Computer Assessment System for English Communication. PAAL Japan Proceedings, pp131-140, 2006.
[7]