sciendo

Acta Educationis Generalis
Volume 13, 2023, Issue 3

DOI: 10.2478/atd-2023-0022

Parameters and Models of Item Response Theory (IRT): A Review of Literature

Abraham Gyamfi - Rosemary Acquaye*

Received: April 20, 2023; received in revised form: June 4, 2023;
accepted: June 6, 2023

Abstract:

Introduction: Item response theory (IRT) has received much attention in validation of assessment instrument because it allows the estimation of students' ability from any set of the items. Item response theory allows the difficulty and discrimination levels of each item on the test to be estimated. In the framework of IRT, item characteristics are independent of the sample and latent traits of the person are independent of the test on the account that the selected models perfectly fit the data. Therefore, scores that describe examinee performance are independent on test difficulty. The scores of the examinee may be lower on a difficult test and higher on easier tests, but the ability level of the examinee remains the same over any test at the time of testing. The IRT model allows the estimation of item parameters. The line of difference between the models and parameters of IRT is not clear to many students in assessment.

Purpose: This paper reviews the parameters that are estimated using IRT and the models available in IRT. Also, the paper highlights the difference between the parameters and models and the various models under each set of data.

Methods: Various literatures on IRT relating to the parameters and models of IRT are reviewed.

Conclusions: There are four parameters estimated with IRT but the models are not four. Again, the models of IRT depends on the type of data. Dichotomous data has four models for the four parameters. However, polytomous data has two parameters: item difficulty and item discrimination for the models.

Key words: item parameter, models of item response theory, item difficulty, items discrimination.

* Abraham Gyamfi, Wesley College of Education, Kumasi, Ghana; abrahamgyamfi84@gmail.com
Rosemary Acquaye, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; rosemaryacquaye06@gmail.com

Introduction

Item response theory (IRT) is also known as latent trait theory or item characteristic curve theory, according to de Ayalaya (2009). Item response theory is a model system for determining the relationship between latent variables and their manifestations (Gyamfi & Wren, 2022; Butakor, 2022). Item response theory does not explain why a person responds a certain way to a question or how they determine what to answer (Bichi, Hafiz, & Bello, 2016; Bichi, Embong, Mamat, & Maiwada, 2015). This distinguishes it in the traditional sense. Instead, IRT is like statistical estimation theory, in which latent characterizations of the examinee and items are used as predictors of observed scores (Gyamfi, 2023). Comprehensive and accurate knowledge on the assumptions, models, and parameters of IRT is required for better utilization of the theory. However, the distinction between the models and parameters of IRT in respect to a particular type of data continues to be a challenge to many students of measurement.

1 Assumptions of IRT for polytomous items

There are two basic assumptions that underpin IRT for polytomous items; unidimensionality and local dependence (Bulut, 2015) and in addition, a third one-measurement invariance, for monotonous items.

- Unidimensionality: Annan-Brew (2020) stated that most popular assumption of IRT models is unidimensionality. “It is a specified form for the Item Response Function (IRF) that can be checked empirically. Unidimensionality means that all items of the test measure the same latent trait and that with the result examinees can be ordered on a linear scale” (p. 29). The unidimensionality makes it possible to estimate the ability of an examinee on the same ability scale from any pool of items in the universe of items. There should be a dominant component that is being measured by the test. Carvalho, Primi and Baptista (2015) reported that polytomous items in mathematics meet theunidimensionality assumption. There has been suggested statistical approach of checking unidimensionality assumption for both polytomous and monotonous items such as confirmatory factor analysis, IRT and principal component analysis (DeMars, 2018).
- Local independence: The local independence assumption is related to unidimensionality. The assumption holds that the responses to items of a test are statistically independent conditional on the ability level of the examinee θ (Annan-Brew, 2020). It indicates that an examinee's performance on one item must not have a positive or negative impact on their responses to subsequent test items. In other words, an item's content cannot give away the answers to other things on the test. Furthermore, it is believed that the likelihood of

correctly answering to an item does not decrease monotonically with skill level. Item response theory (IRT) is "a set of latent variable techniques that are specifically designed to explain the interaction between a subject's ability and item level stimuli such as difficulty, discriminating, guessing, and others." (Janssen, Meier, & Trace, 2014, p. 9) The IRT is concerned with the pattern of responses rather than the total score variables and linear regression theory. Le (2013) pointed out that performance-based assessment produces more local independence than traditional assessment. A residual correction coefficient of 0.7 is used to check the local independence assumption for both polytomous and monotonous (Lord, 2008; Min & He, 2014).

- Measurement invariance: This assumption of IRT holds that the items are the same for all groups across the test. This means that all examines respond to the same set of items. It is therefore an assumption based on system rather than the responses. Thus, if a system is comprised of measurements of several items using a single instrument, measurement invariance shall be defined as observing the same relationships between measurements when a second measurement instrument is used in the assessment. It can be defined as the equality of item and examinee parameters from different examinee populations or measurement conditions (Annan-Brew, 2020). This may not be applicable in polytomous items where examinees have options to select from when taking the tests. That is, when all examinees are not mandated to respond to the same set of items, the measurement invariance assumption is not applicable. A typical situation is in the case of performance-based assessment.

2 Parameters in item response theory

Parameters of IRT are the characteristics of the items that are estimated using IRT. There are four basic parameters that IRT estimates. These are item difficulty, item discrimination, guessing and ceiling effect (Annan-Brew, 2020). They are denoted by letters bi, ai, ci and di respectively.

- Item difficulty: According to Liaquat, Asif, Siraji, and Maroof (2012), item difficulty means the percentage of students who answer correctly each test item. Item difficulty indices is an indication of the proportion of the examinees who responded to the item correctly. The lesser the proportion, the difficulty the item is. The bi is the location of the examinee on the item. The b is called ability level needed to respond above a specific threshold with 50% probability.
- Item discrimination: As Nitko (2001) puts it, item discrimination (ai) is the difference between the fraction of the upper group answering the item correctly and the fraction of the lower group answering the item correctly. The ai indicates the extent to which the item is able to differentiate between

higher achieving students and lower achieving students. According to Nitko (2001), item discrimination is important because it is able to indicate both the absolute achievement and relative achievement of the students. By absolute achievement, item discrimination is able to determine the level of subject matter a student has accurately learned.

- Guessing effect: The parameter ci represents the likelihood of properly answering the item by guessing alone. The value of ci does not change with the level of ability. It is the same regardless of skill level. Examinees with high ability and those with low ability both have the same chance of guessing correctly on an item (Annan-Brew, 2020).
- Ceiling effect: The di-parameter is described as the item upper asymptote of carelessness. Response time and the slowness parameter are combined in the suggested 4-parameter logistic model (Zanon, Hutz, Yoo, & Hambleton, 2016). The 4PL has not really been formally added to the traditional IRT models. Also, softwares are not available for analysing it.

3 Models in item response theory

The IRT framework encompasses a group of models. There are two major categories of IRT models depending on the type of data set: models for dichotomous items and models for polytomous items.

3.1 IRT models for dichotomous items

For dichotomously scored test questions, there are four IRT models, three of which are considered standard IRT models, known as 1, 2, and 3 parameter IRT models (Annan-Brew, 2020; Butakor, 2022). The number of parameters estimated gives it names as 1PLM, 2PLM, 3PLM and 4PLM.

1) Four-parameter logistic (4PL) model

The 4th paramater was introduced by Barton and Lord (as cited in Annan-Brew, 2020) as an upper asymptote parameter. It was added to the difficulty, discrimination and guessing make it 4PL. It is also referred to as the ceiling effect parameter, expressed by *d*, into the 3PL model, resulting in the 4PL model. It is represented by the equation:

$$P_4(\theta)=C+(di-ci)\frac{1}{1+\exp[-1.702ai\theta-bi]}$$

"Where d_i = upper asymptote parameter, ci = guessing parameter of the item, a_i,= discrimination parameter of item commonly known as item slope, b_i= difficulty parameter of item known as item location parameter a and θ (Theta) = the ability level of a particular examinee. The P4PL(θ) ranges from the lower asymptote *ci*

to 1, and P4PL(θ) ranges from *ci* to the upper asymptote parameter *di*. The *di*-parameter is described as the item upper asymptote of carelessness."

The parameter c represents the likelihood of properly answering the item by guessing alone. The value of c does not change with the level of ability. It is the same regardless of skill level. Examinees with high ability and those with low ability both have the same chance of guessing correctly on an item. Although the parameter c has a theoretical range of $0 \leq c \leq 1$, the acceptable range in practice is $\theta \leq c \leq 0.35$.

The location parameter is also known as the difficulty parameter. It is represented by the letter b, and it's defined as the point on the ability scale where the probability of properly answering the item is 0.5. Theoretically, range is $-\infty \leq b \leq \infty$, although the typical value of range is $-3 \leq b \leq 3$. "It is the slope of the tangent line of the item characteristics curve at the point of the location parameter", which is the item discrimination parameter (the slope parameter). It has normal values in the $-3 \leq b \leq 3$ range. While most test items discriminate positively (the likelihood of responding correctly to an item rises as ability level rises), some items discriminate negatively (the probability of responding correctly to the item decrease as the ability level increases from low to high) (Mozgalina, 2015; Butakor, 2022).

2) Three-parameter model

The 3PL estimates three parameters that is difficulty, discrimination and guessing size in the responses. It assumes that the d parameter is constant hence taking out of the analysis. It is represented by the equation:

$$P(\theta) = c_i + (1 - c_i) \frac{1}{1+e-ai(\theta - bi)}$$

"Where c_i = guessing parameter of the item, a_i,= discrimination parameter of item commonly known as item slope, b_i= difficulty parameter of item known as item location parameter and θ (Theta) = the ability level of a particular examinee." Figure 1 illustrates a 3-PLM (Park, 2012).

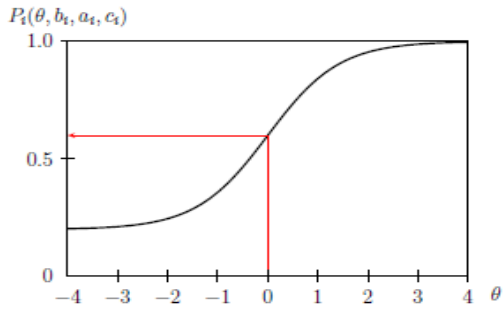

Figure 1. Three-parameter model.

3) Two-parameter model

For the 2PL, the guessing factor c is assumed or constrained to be zero. It is assumed that the guessing factor is constant; therefore, the 3-parameter model is reduced to the 2- parameter model for which only item location (difficulty) and item slope (discrimination) parameters are estimated. Mathematically, the two-parameter is represented as follows (Park, 2012) and Figure 2 illustrates a two-parameter model.

$$P(\theta) = \frac{1}{1+e-ai(\theta - bi)}$$

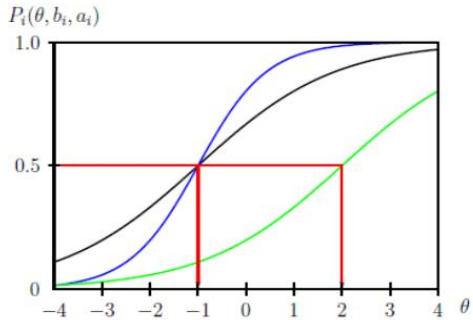

Figure 2. Two-parameter model.

4) One-parameter model

The Rasch model is another name for the one-parameter model. It was named after a scientist who first worked in the area. The 2PL is also subject to a constraint that stipulates that all things have the same and fixed discrimination.

The difficulty parameter 'b' is the only one that is estimated. As a result, parameter 'a' is treated as a constant rather than a variable, and it is not approximated. The IRT model then becomes one parameter model as illustrated by the equation and Figure 3.

$$P(\theta) = \frac{1}{1+e(\theta - bi)}$$

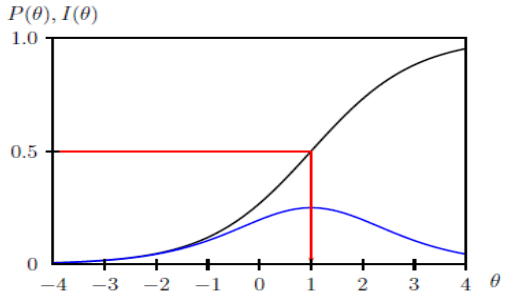

Figure 3. One-parameter model

3.1.1 Comparison of the models

Wyatt (2016, p. 234) claimed that "the 2-parameter model is the best for norm-referenced tests". This is because, aside the difficulty index of each item acquired from the Rasch model, the discrimination powers of the items are evaluated. The 3-parameter model can also be used to estimate discrimination power. The value of c (the guessing parameter) is unaffected by the level of ability. As a result, the examinees with the lowest and best abilities have the same probability of guessing right on the item (Royal & Gonzalez, 2016).

Under the 3-parameter model, the discrimination parameter 'a' could be read as proportional to the slope of the item characteristic curve at θ = b. The slope of the item characteristic curve at θ=b, however, is really $a\frac{1-c}{4}$. While the differences in the values of parameters 'b' and 'a' may be little, they are significant when evaluating the findings of test studies (Wyatt, 2016).

3.2 IRT models for polytomous items

There are two major types of polytomous IRT models; generalized partial credit model and the graded response models. The two focuses on the two types of response probability, which are peculiar to polytomous models and their corresponding response functions. The response functions are modelled differently by the different types of polytomous IRT model.

3.2.1 Generalized partial credit model (GPCM)

Generalized partial credit model (GPCM) provides both conceptual and mathematical description of the major specific polytomous models. "Such models include the Nominal Response Model (NRM), the Partial Credit Model (PCM), the Rating Scale model (RSM) which are all variations of the generalized partial credit model. Partial Credit Model uses the Rasch model to specify the probability of success." That is the partial credit model is the counterpart of the Rasch model which estimates the difficulty 'b' parameter. The PCM is quite popular in assessment contexts because it has limited number of assumptions and steps in the analysis. For PCM, sample size as small as 300 could produce a reliable estimate of the trait (de Ayala, 2009).

Unlike the PCM, Generalized Partial Credit Model includes the item-level discrimination parameter. It is a counterpart of the 2PL of dichotomous items. It is represented by the equation similar to that of the 2PL.

$$Pij(\theta) = \frac{1}{1 + \exp[ai(\theta - bij)]}$$

The GPCM has flexibility characteristics. It offers the possibility to identify item response options that may be duplicated with each other. Example is a situation where the IRFs for some response options may be centred at the same ability level (de Ayala, 2009). One significant distinction between PCM and RSM is that RSM assumes that the gap between item difficulty levels is the same for all things on the test. When a common set of anchor items (Likert-type anchors) elicits item replies, this is usual.

3.2.2 Graded response model (GRM)

The Graded Response Model (GRM) by Samejima (1969) belongs to the cumulative approach where all categories of scores are used to quantify the probability of success or failure (de Ayala, 2009). The GRM estimates probabilities based on the specification of 2PL. Separate b_i parameters are estimated for each step of the item. However, it uses one a_i parameter for all steps for each item. The GRM indicates m-1 "boundary" response functions which are an indication of the cumulative probability for a response category greater than the option of interest. It is represented by the equation:

$$Pij(\theta) = \frac{\exp[ai(\theta - bij)]}{1 + \exp[ai(\theta - bij)]}$$

The reason for using GRM, or any model is based on ordered response categories, with testlet-based scores (group of items based on the same or similar

content developed as a unit with predetermined procedures that the examine may follow) is that, theoretically, testlet-based scores can have an ordered quality if scores "correspond to the degree of completeness of the examinee's reasoning process within a testlet" (Le, 2013, p. 58). Table 3 shows the summary of the IRT models.

Table 1

Summary of IRT models

IRT model	*Item format*	*Model characteristics*
Rasch model	Dichotomous	Equal discrimination across all items is assumed; estimation of difficulty location parameter for each item
Two parameter logistic (2PL)		Estimation of discrimination (slope) and difficulty (location) parameters for each item
Three parameter logistic (3PL)		Estimation of discrimination (slope), difficulty (location), and guessing parameters for each item
Four parameter logistic (4PL)		Estimation of discrimination (slope), difficulty (location), guessing and ceiling parameters for each item
Graded Response Model	Polytomous	Used for ordered responses and discrimination varies across items
Generalized Partial Credit Model		Used for ordered responses and discrimination varies across items. Could be used as an alternative to graded response model
Partial Credit Model		Assumes equal discrimination across all items. Estimation of separate category location parameters for each item
Rating Scale Model		Equal discrimination across all items. Estimate a single set of categorical location parameters for all items
Bock's Nominal Model		Used for unordered responses. Discrimination allowed to vary across items

Conclusion

It should be noted that there are not four models of IRT. But there are four parameters estimated with IRT. Again, the models of IRT depend on the type of data. Dichotomous data have four models for the four parameters. However, polytomous data have two parameters: item difficulty and item discrimination. This is because there is no guessing effect in polytomous items. Therefore, only two parameters are estimated for models of polytomous data.

References

Annan-Brew, R. (2020). *Differential Item Functioning of West African Senior Secondary Certificate Examination in Core Subjects in Southern Ghana* [Doctoral Thesis]. Ghana: UCC.

Bichi, A. A., Hafiz, H., & Bello, S. A. (2016). Evaluation of Northwest University, Kano post-UTME test items using Item response theory. *International Journal of Evaluation and Research in Education (IJERE), 5*(4), 261-270.

Bichi, A. A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. *Austrian Journal of Basic & Applied Science 9*(7), 549-556.

Bulut, O. (2015). Applying item response theory models to entrance examination for graduate studies: Practical issues and insights. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 313-330.

Butakor P. K. (2022). Using classical test and item response theories to evaluate psychometric quality of teacher-made test in Ghana. *European Scientific Journal, ESJ, 18*(1), 139. https://doi.org/10.19044/esj.2022.v18n1p139

Carvalho, L. F., Primi, R., & Baptista, M. N. (2015). IRT Application to verify psychometric properties of the Beck Depression Inventory (BDI) University Psychological. *Bogotá, Colombia, 14*(1), 91-102.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

DeMars, C. E. (2018). Item information function. In *SAGE Encyclopaedia of Educational Research, Measurement, and Evaluation.* Thousand Oaks: SAGE Publications.

Gyamfi, A. & Wren, D. (2022). Determining the difficulty and discrimination parameters of a Mathematics performance-based assessment. *Creative Education, 13*(11), 3483-3489.

Gyamfi, A. (2023). Differential item functioning of performance-based assessment in mathematics for senior high schools. *Jurnal Evaluasi Dan Pembelajaran*, *5*(1), 20-34.

Janssen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal, 16*(2), 37-54.

Le, D. (2013). *Applying Item Response Theory Modeling in Educational Research* [Master's Thesis]. Iowa *State University.*

Liaquat, H., Asif, J. M., Siraji, J., & Maroof, K. (2012*). Development and standardization of intelligence test for children. International Journal of Learning & Development, 2*(5), 190-202.

Lord, F. M. (2008). *Application of item response theory to practical testing problems.* London: Routledge Taylor & Francis Group.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477.

Mozgalina, A. (2015). *Applying an Argument-Based Approach for Validating Language Proficiency Assessments in Second Language Acquisition Research: The Elicited Imitation Test for Russian* [Doctoral Thesis]. Georgetown University.

Nitko, A. J. (2001). *Educational Measurements* (3rd ed.). USA: American council on education.

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics: An Argument-Based Approach to Validation* [Doctoral Thesis]. University of Minnesota.

Royal, K. D., & Gonzalez, L. M. (2016). An evaluation of the psychometric properties of an advising survey for medical and professional program students. *Journal of Educational and Developmental Psychology, 6*(1), 195-203.

Wyatt, C. (2016). *The Development and Validation of an Instrument to Measure Teachers' Perceptions of the Effect of Mobile Technology Initiatives on Classroom Climate*. University of Poland.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica, 29*(18), 345-367.