Item Response Theory Analysis Report

TestAnaAPP

06/09/2024

**Table of Contents**

[1. Introduction to IRT Models 1](#_Toc179897872)

[1.1. IRT Models for Dichotomous Scoring Data 1](#_Toc179897873)

[1.1.1. Rasch model (1PL) 1](#_Toc179897874)

[1.1.2. Two-parameter logistic model (2PL) 1](#_Toc179897875)

[1.1.3. Three-parameter logistic model (3PL) 1](#_Toc179897876)

[1.1.4. Four-parameter logistic model (4PL) 1](#_Toc179897877)

[1.2. IRT Models for Polytomous Data 1](#_Toc179897878)

[1.2.1. Partial credit model (PCM) 1](#_Toc179897879)

[1.2.2. Generalized partial credit model (GPCM) 1](#_Toc179897880)

[1.2.3. Graded response model (GRM) 1](#_Toc179897881)

[2. Selected Model and Settings 1](#_Toc179897882)

[3. Model Fit 1](#_Toc179897883)

[3.1.Relative Fit Indices 1](#_Toc179897884)

[3.2.Absolute Fit Indices 1](#_Toc179897885)

[4. Hypothesis Testing 1](#_Toc179897886)

[4.1.Unidimensionality Test 1](#_Toc179897887)

[4.2.Independence Test 1](#_Toc179897888)

[5. Item Fit 1](#_Toc179897889)

[6. Item Parameters 1](#_Toc179897890)

[7. Wright Map 1](#_Toc179897891)

[8. Item Characteristic Curve (ICC) 1](#_Toc179897892)

[9. Item Information Curve (IIC) 1](#_Toc179897893)

[10. Test Information Curve (TIC) 1](#_Toc179897894)

[11. Reference 1](#_Toc179897895)

**List of Figures**

[**Figure** The item characteristic curve of this test 1](#_Toc179897896)

[**Figure** The item information curve of this test 1](#_Toc179897897)

[**Figure** The test information curve and measurement error 1](#_Toc179897898)

**List of Tables**

[**Table** The relative fit indices of the Graded response model (GRM) 1](#_Toc179897899)

[**Table** The absolute fit indices of the Graded response model (GRM) 1](#_Toc179897900)

[**Table** Results of unidimensionality test (EFA eigenvalue) 1](#_Toc179897901)

[**Table** Independence test (Q3) 1](#_Toc179897902)

[**Table** The results of item fit (Graded response model (GRM)-chi-squared test (Kang & Chen, 2007)) 1](#_Toc179897903)

[**Table** The item parameters of the Graded response model (GRM) 1](#_Toc179897904)

In the field of psychological and educational measurement, two commonly used measurement theories are classical test theory (CTT) and item response theory (IRT). CTT typically relies on the total score as a reflection of the test taker’s underlying trait level. It possesses characteristics that are easily understandable and practical. However, CTT has inherent limitations that cannot be overcome due to its simplicity. These limitations include the high dependency of test scores on test items and the dependency of item parameters on the test sample, and the lack of comparability between item difficulty and test taker scores. On the other hand, IRT overcomes some of these limitations through extensive mathematical calculations, large samples, and strong assumptions. IRT provides estimates of item and test taker latent traits. It offers several specific advantages:

(1) Invariance of test taker ability and item parameters.  
(2) Test taker ability and item difficulty parameters on the same scale, allowing for comparability.  
(3) IRT provides different measurement errors for test takers at different ability levels, which is more reasonable.  
(4) Others.

The ‘TestAnaAPP’ is an interactive program developed in the R language for test analysis. It addresses the high computational cost and complexity of IRT models and provides practitioners of test quality analysis with a comprehensive set of evaluation indicators for items. Additionally, it offers professional visualization for the results.

# Introduction to IRT Models

‘TestAnaAPP’ offers a range of commonly used scoring models in the field of selection, catering to both dichotomous and polytomous data. The dichotomous scoring models consist of the Rasch model, the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), the three-parameter logistic model (3PL), and the four-parameter logistic model (4PL). On the other hand, the polytomous scoring models encompass the graded response model (GRM), the partial credit model (PCM), and the generalized partial credit model (GPCM). These models are widely recognized and utilized in the field.

It is important to note that in IRT, the Rasch model is equivalent to the One-Parameter Logistic (1PL) model. The terms “one-parameter” and “two-parameter” refer to the number of item parameters in the IRT models. The specific models are described below:

## IRT Models for Dichotomous Scoring Data

Let denote the response of test taker (where ) to item (where ) in dichotomous scoring data. In this type of data, the value of is assigned as 0 (for incorrect responses) or 1 (for correct responses). Moreover, let represent the item difficulty parameter, and represent the latent trait parameter of the test taker. In the context of IRT analysis, the difficulty parameters and the latent trait parameters of test takers typically follow a standard normal distribution, allowing for comparability. This comparability enables the determination of the difficulty level of an item for test takers at specific levels, based on the relative magnitudes of these two parameters.

The ‘TestAnaAPP’ program supports the following specific IRT models for dichotomous scoring data:

### Rasch model (1PL)

The Rasch model, also known as the simple Rasch model, is employed for items that are scored dichotomously. The item response function (IRF) of the Rasch model was originally defined by Georg Rasch in 1960. The formula for the IRF is as follows:

The Rasch model (1PL) assumes that the probability of test taker answering item correctly (represented by ) is determined by the relative difficulty of item for test taker , denoted as .

### Two-parameter logistic model (2PL)

The 2PL model, originally proposed by Birnbaum in 1957, characterizes the psychometric features of items through the parameters of item difficulty () and item discrimination (). Item discrimination reflects the item’s capacity to distinguish between subjects with varying levels of latent traits, and is a crucial measure of item quality. The item response function (IRF) of the 2PL model is illustrated below.

### Three-parameter logistic model (3PL)

Birnbaum (1968) expanded upon the 2PL model by introducing the item guessing parameter, denoted as , and proposing the 3PL model. The item guessing parameter, , signifies the likelihood of a subject correctly guessing item . In the 3PL model, the probability that subject answers item correctly, defined as , is influenced by the relative difficulty of item for subject (), the item discrimination level (), and the guessing parameter (). The item response function (IRF) of the 3PL model is displayed below.

### Four-parameter logistic model (4PL)

The 3PL model addresses the guessing behavior of subjects with very low ability on items, while the 4PL model expands on this by also accounting for slip made by highly capable subjects. In addition to the components of the 3PL model, the 4PL model introduces an extra item slip parameter, denoted as , to represent the probability of high-ability subjects making mistakes on item . The formula for the item response function (IRF) of the 4PL model is displayed below.

For more details on the 4PL model, refer to Barton and Lord (1981).

## IRT Models for Polytomous Data

Let denote the response of subject (where ) to item (where ) in a polytomous data context, where there exist response categories. Specifically, , where . Moreover, in the polytomous item response theory (IRT) model, the item threshold (or difficulty) parameters are represented as , with , while the latent trait parameter for the subject is denoted as .

### Partial credit model (PCM)

The partial credit model (PCM) is a variant of the Rasch model (Masters, 1982) specifically designed to analyze response data with more than two categories. In the PCM model, the item parameters solely consist of difficulty parameters . Here, represents the threshold parameter (difficulty) for category on item .

The item response function (IRF) for the PCM is represented by the following equation:

where represents the probability that subject responds to item with category .

### Generalized partial credit model (GPCM)

The generalized partial credit model (GPCM) is a modification of the PCM proposed by Muraki in 1992. The GPCM enables the estimation of item discrimination parameters, denoted as . The Item Response Function (IRF) for the GPCM model can be described as:

### Graded response model (GRM)

Both the PCM and GPCM models are founded on the modeling of the probability of response categories for item . In contrast, the GRM model estimates the cumulative probability of response categories or higher. The cumulative response probability function for the GRM model can be expressed as:

where represents the probability that subject responds to item with category or a higher category. The parameter corresponds to the threshold (difficulty) of response category on item . To calculate the probability of a specific response category, it is necessary to subtract the adjoining cumulative probabilities:

It is important to note that the GRM model, in contrast to the GPCM and PCM models, assumes the inequality . This inequality indicates that the difficulty of higher response categories is consistently greater than that of lower categories.

# Selected Model and Settings

**The results presented in this document were estimated based on the settings you selected in the interactive ‘TestAnaAPP’ program, along with default options within the program. The following settings were employed:**

* **IRT model:** Graded response model (GRM)；
* **Model parameter estimation method:** standard EM algorithm；
* **Person parameter estimation method:** expected a-posteriori (EAP)；
* **EFA estimation method:** Principal Component Analysis；
* **EFA factor rotation method:** Varimax；
* **Selected indicator for independence tests:** Q3；
* **Selected indicator for item fit tests:** chi-squared test (Kang & Chen, 2007)；

# Model Fit

The model’s fitting results included both relative fit indices and absolute fit indices. Relative fit indices serve as a means to compare the level of fit between models. When two IRT models are used to fit the same dataset and produce different sets of relative fit index values, the model with the lower relative fit index value is generally considered to fit the data comparatively better. Absolute fit indices reflect the degree of absolute fit between the model and the data. If the absolute fit indices fall within an acceptable range, it suggests that the model fits the data well.

## Relative Fit Indices

The ‘TestAnaAPP’ program utilizes the ‘mirt’ package for parameter estimation and calculates relative fit indices, including AIC, SABIC, HQ, BIC, and Log-likelihood. Smaller values of AIC, SABIC, HQ, and BIC indicate a higher degree of model fit. A greater value of Log-likelihood signifies a higher degree of model fit. To employ relative fit indices for IRT model selection, you can opt for various models during parameter estimation in ‘TestAnaAPP’ and record the corresponding relative fit index magnitudes.

Table *1* The relative fit indices of the Graded response model (GRM)

| **AIC** | **SABIC** | **HQ** | **BIC** | **logLik** |
| --- | --- | --- | --- | --- |
| **22,443.99** | 22,527.5 | 22,535 | 22,683.12 | -11,172.99 |

## Absolute Fit Indices

‘TestAnaAPP’ generates absolute fit indices, including M2, RMSEA, SRMSR, TLI, and CFI.

* M2: If the corresponding p-value is less than 0.05 (α<0.05), the null hypothesis is rejected, suggesting inadequate fit of the model to the data.
* RMSEA: Browne and Cudeck (1993) provided a well-regarded guide to interpreting RMSEA. A value below 0.05 indicates a very good fit (close fit), 0.05 to 0.08 suggests a relatively good fit (fair fit), 0.08 to 0.10 implies a mediocre fit, and a value above 0.10 indicates a poor fit.
* SRMSR: A value of SRMSR≤0.05 indicates acceptable fit between the model and the data. TLI and CFI values greater than 0.9 suggest satisfactory fit of the model to the data.

Table *2* The absolute fit indices of the Graded response model (GRM)

| **M2** | **df** | **p** | **RMSEA** | **RMSEA\_5** | **RMSEA\_95** | **SRMSR** | **TLI** | **CFI** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **158.846** | 14 | 0 | 0.103 | 0.089 | 0.118 | 0.067 | 0.922 | 0.948 |

# Hypothesis Testing

## Unidimensionality Test

Common methods used for testing unidimensionality include exploratory and confirmatory factor analysis. The software ‘TestAnaAPP’ generates exploratory factor analysis results for the unidimensionality testing. To assess conformity to the unidimensionality hypothesis, the criterion is that the eigenvalue of the first factor is at least 3-5 times greater than the eigenvalue of the second factor. If Principal Component Analysis is used for exploratory factor analysis, the eigenvalue of the first factor is 3.48 times higher than the eigenvalue of the second factor. Therefore, **it conforms to the unidimensionality hypothesis.**

Table *3* Results of unidimensionality test (EFA eigenvalue)

| **rownames.fit.eigenvalues.** | **Eigenvalue** | **Variance..** | **Cumulative..** | **SS.Loading** | **Variance...1** | **Cumulative...1** |
| --- | --- | --- | --- | --- | --- | --- |
| **Component 1** | 3.192 | 45.606 | 45.606 | 3.192 | 45.606 | 45.606 |
| **Component 2** | 0.917 | 13.102 | 58.708 |  |  |  |
| **Component 3** | 0.789 | 11.267 | 69.975 |  |  |  |
| **Component 4** | 0.681 | 9.733 | 79.708 |  |  |  |
| **Component 5** | 0.505 | 7.219 | 86.927 |  |  |  |
| **Component 6** | 0.466 | 6.662 | 93.589 |  |  |  |
| **Component 7** | 0.449 | 6.411 | 100.000 |  |  |  |

## Independence Test

‘TestAnaAPP’ provides the Q3 and LD-X2 statistics for independent testing in IRT. Larger absolute values indicate stronger dependencies between the two items. Due to the lack of uniform standards across different studies, we do not provide the cutoff values for Q3 and LD-X2. Readers are encouraged to refer to professional books or papers for data interpretation.

In this analysis, you selected **Q3** as the statistic for the independence test.

Table *4* Independence test (Q3)

| **Item** | **PIL2** | **PIL5** | **PIL7** | **PIL10** | **PIL17** | **PIL18** | **PIL19** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PIL2** | 1.000 | 0.202 | -0.076 | -0.033 | -0.213 | -0.211 | -0.194 |
| **PIL5** | 0.202 | 1.000 | -0.045 | -0.005 | -0.241 | -0.256 | -0.263 |
| **PIL7** | -0.076 | -0.045 | 1.000 | -0.087 | -0.059 | -0.124 | -0.174 |
| **PIL10** | -0.033 | -0.005 | -0.087 | 1.000 | -0.200 | -0.145 | -0.067 |
| **PIL17** | -0.213 | -0.241 | -0.059 | -0.200 | 1.000 | -0.058 | -0.293 |
| **PIL18** | -0.211 | -0.256 | -0.124 | -0.145 | -0.058 | 1.000 | -0.147 |
| **PIL19** | -0.194 | -0.263 | -0.174 | -0.067 | -0.293 | -0.147 | 1.000 |

*Note*: The red color indicates that the absolute value of the Q3 statistic exceeds the threshold value you set on TestAnaAPP.

# Item Fit

The item fit test is utilized to assess the existence of a significant disparity between the predicted model responses and the actual responses. ‘TestAnaAPP’ provides the and statistics for performing the item fit test. A value less than 0.05 denotes a significant disparity between the predicted model values and the actual data. Furthermore, an RMSEA value greater than 0.1 suggests a substantial discrepancy between the predicted values and the actual data. A comprehensive assessment can be made by considering the test situation and statistical indicators. Selecting an alternative model for the fit could potentially yield divergent outcomes.

Table *5* The results of item fit (Graded response model (GRM)-chi-squared test (Kang & Chen, 2007))

| **item** | **S\_X2** | **df.S\_X2** | **RMSEA.S\_X2** | **p.S\_X2** |
| --- | --- | --- | --- | --- |
| **PIL2** | 219.411 | 120 | 0.029 | 0.000 |
| **PIL5** | 173.070 | 114 | 0.023 | 0.000 |
| **PIL7** | 157.727 | 122 | 0.017 | 0.016 |
| **PIL10** | 183.972 | 128 | 0.021 | 0.001 |
| **PIL17** | 109.340 | 97 | 0.011 | 0.185 |
| **PIL18** | 131.495 | 108 | 0.015 | 0.062 |
| **PIL19** | 135.205 | 86 | 0.024 | 0.001 |

# Item Parameters

Item parameters serve as crucial indicators for assessing item quality. Provided below are the criteria, derived from “**The Basics of Item Response Theory Using R**”, that can be utilized to evaluate item discrimination values. Readers can employ these criteria for evaluating the extent of item discrimination in the test under analysis.

**Table: Labels for Item Discrimination Values**

| Verbal label | Range of values | Typical Value |
| --- | --- | --- |
| None | 0 | 0.00 |
| Very low | 0.01-0.34 | 0.18 |
| Low | 0.35-0.64 | 0.50 |
| Moderate | 0.65-1.34 | 1.00 |
| High | 1.35-1.69 | 1.50 |
| Very high | >1.70 | 2.00 |
| Perfect |  |  |

The Graded response model (GRM) model was chosen for this analysis. Parameter estimation was performed using the standard EM algorithm. The item parameters are presented below:

Table *6* The item parameters of the Graded response model (GRM)

| **Item** | **Discrimination** | **Difficulty1** | **Difficulty2** | **Difficulty3** | **Difficulty4** | **Difficulty5** | **Difficulty6** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PIL2** | 1.252 | -3.032 | -1.880 | -0.868 | 0.090 | 1.053 | 2.309 |
| **PIL5** | 1.518 | -2.879 | -1.669 | -0.921 | -0.295 | 0.274 | 1.154 |
| **PIL7** | 1.303 | -3.095 | -2.400 | -1.847 | -1.306 | -0.741 | 0.080 |
| **PIL10** | 1.042 | -2.333 | -1.471 | -0.678 | 0.467 | 1.163 | 2.324 |
| **PIL17** | 2.035 | -2.535 | -1.779 | -1.045 | -0.263 | 0.396 | 1.480 |
| **PIL18** | 1.739 | -2.713 | -1.877 | -1.208 | -0.425 | 0.348 | 1.243 |
| **PIL19** | 2.410 | -2.505 | -1.820 | -1.117 | -0.247 | 0.552 | 1.494 |

*Note*: The red color indicates that the item discrimination value is below the threshold value you set on TestAnaAPP.

# Wright Map

IRT assess item difficulty and participant latent traits on the same scale, enabling comparability. This provides a significant advantage compared to CTT. The Wright map serves as a visual tool for comparing the distribution of item difficulty and latent traits. Through comparing the scatterplot distribution of item difficulty with latent trait values, the relative difficulty or ease of the test for the subjects, as well as the difficulty level of test, can be determined. Additionally, it enables the identification of harder and easier items.

It is important to note that the test can effectively differentiate the subjects’ latent trait levels only when it possesses appropriate difficulty. Otherwise, ceiling or floor effects may arise, whereby the measurement tool inaccurately assesses the subjects’ true abilities, resulting in all subjects receiving either high or low scores.

## NULL

# Item Characteristic Curve (ICC)

The Item Characteristic Curve (ICC) visualizes the probability of subjects obtaining a specific score based on the predictions of the statistical model.

The ICC allows us to observe the probability of subjects with different levels of latent traits obtaining a specific score on each item.

The ICC provides insights into the characteristics of items for subjects with varying levels of latent traits.

Specialized books offer guidance on interpreting the ICC curve, enabling readers to better understand its implications.



Figure *1* The item characteristic curve of this test

# Item Information Curve (IIC)

Item information represents the level of measurement error associated with each item. Higher values indicate a greater contribution of the item to the accurate measurement of the latent trait of the subjects. The value of item information is often related to the discriminatory power and difficulty level of the item. The Item Information Curve (IIC) visually represents the item information and allows readers to assess the contribution of the items to the accurate measurement of a specific level of latent trait.

Professional books can provide guidance on interpreting the IIC curve for those seeking to gain a comprehensive understanding.



Figure *2* The item information curve of this test

# Test Information Curve (TIC)

Test information is a quantitative measure of test accuracy in IRT, calculated by summing the information obtained from each item. A higher test information value indicates a more precise measurement of the latent trait for the subjects being assessed. ‘TestAnaAPP’ offers test information curves (TIC) and measurement error curves as graphical representations to demonstrate the quality of the test. A consistent conversion relationship exists between test information () and measurement error (), which can be expressed by the following formula:



Figure *3* The test information curve and measurement error

It is important to note that users can download the item information and test information from the user interface. This allows them to create custom figures using the downloaded files.

**Statement:** ‘TestAnaAPP’ is mainly designed to provide convenient item and test analysis programs for practitioners in the field of psychological and educational measurement. The purpose of providing this report is to present the analysis results in a complete and clear manner, and to provide universal and optimal result interpretation is not the primary objective of ‘TestAnaAPP’. **Therefore, while we hope users can benefit from the user-friendly nature of this program, we also encourage maintaining a cautious attitude towards the interpretation of results in order to avoid potential biases. Additionally, if users discover any errors in this program, we would greatly appreciate your contact with us (**[**jiangyouxiang34@163.com**](mailto:jiangyouxiang34@163.com)**).**

# Reference

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item response model (No. 150–453). Princeton, NJ: Educational Testing Service.  
Birnbaum, A.(1957). Efficient design and use of tests of a mental ability for various decision making problems (Series Report No. 58-16). Randolph Air Force Base, TX: USAF School of Aviation Medicine.  
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397–479). Reading, MA: Addison-Wesley.  
Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.  
Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.  
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159–176.  
Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Denmarks Paedagogiske Institut, Copenhagen. Republished in 1980 by the University of Chicago Press.