

## Description

Item response theory (IRT) is used in the design, analysis, scoring, and comparison of tests and similar instruments whose purpose is to measure unobservable characteristics of the respondents. This entry discusses some fundamental and theoretical aspects of IRT and illustrates these with worked examples.

The entries that follow describe how you can use the `irt` suite of commands to fit a variety of IRT models and to evaluate the results. The commands for fitting models can be grouped by the type of responses you are modeling.

### Binary response models

<code>irt 1pl</code>	One-parameter logistic model
<code>irt 2pl</code>	Two-parameter logistic model
<code>irt 3pl</code>	Three-parameter logistic model

### Categorical response models

<code>irt grm</code>	Graded response model
<code>irt nrm</code>	Nominal response model
<code>irt pcm</code>	Partial credit model
<code>irt rsm</code>	Rating scale model

### Multiple IRT models combined

<code>irt hybrid</code>	Hybrid IRT models
-------------------------	-------------------

These models can allow for differences across groups in the population.

### Multiple-group IRT models

<code>irt, group()</code>	IRT models for multiple groups
---------------------------	--------------------------------

Constraints can be applied when fitting any IRT model, and they are particularly useful for constraining parameters across groups in multiple-group models.

### Constraints

<code>irt constraints</code>	Specifying constraints
------------------------------	------------------------

After fitting any IRT model, results can be reported, interpreted, and evaluated using postestimation commands.

### IRT graphs

<code>irtgraph icc</code>	Item characteristic curve plot
<code>irtgraph tcc</code>	Test characteristic curve plot
<code>irtgraph iif</code>	Item information function plot
<code>irtgraph tif</code>	Test information function plot

### IRT reports

<code>estat report</code>	Report estimated IRT parameters
<code>estat greport</code>	Report estimated group IRT parameters

### Model-specific postestimation overview

<code>irt 1pl postestimation</code>	Postestimation tools for irt 1pl
<code>irt 2pl postestimation</code>	Postestimation tools for irt 2pl
<code>irt 3pl postestimation</code>	Postestimation tools for irt 3pl
<code>irt grm postestimation</code>	Postestimation tools for irt grm
<code>irt nrm postestimation</code>	Postestimation tools for irt nrm
<code>irt pcm postestimation</code>	Postestimation tools for irt pcm
<code>irt rsm postestimation</code>	Postestimation tools for irt rsm
<code>irt hybrid postestimation</code>	Postestimation tools for irt hybrid
<code>irt, group() postestimation</code>	Postestimation tools for group IRT

Differential item functioning (DIF) occurs when items that are intended to measure a trait are unfair, favoring one group of individuals over another. DIF can be evaluated by fitting a multiple-group IRT model using `irt`, `group()` or by using a logistic regression or Mantel–Haenszel DIF test.

### Differential item functioning

<code>DIF</code>	Introduction to differential item functioning
<code>diflogistic</code>	Logistic regression DIF
<code>difmh</code>	Mantel–Haenszel DIF

## Remarks and examples

Researchers are often interested in studying abilities, personality traits, and other unobservable characteristics. Throughout this manual, we most often refer to the unobserved characteristic of interest as the latent trait, but we will sometimes also use the term ability.

Latent traits cannot be measured directly, because they are unobservable, but they can be quantified with an instrument. An instrument is simply a collection of items designed to measure a person's level of the latent trait. For example, a researcher interested in measuring mathematical ability (latent trait) may design a test (instrument) consisting of 100 questions (items).

When designing the instrument or analyzing data from the instrument, the researcher is interested in how each individual item relates to the trait and how the group of items as a whole relates to this trait. IRT models allow us to study these relationships.

IRT models are used extensively in the study of cognitive and personality traits, health outcomes, and in the development of item banks and computerized adaptive testing. Some examples of applied work include measuring computer anxiety in grade school children (King and Bond 1996), assessing physical functioning in adults with HIV (Wu et al. 1997), and measuring the degree of public policy involvement of nutritional professionals (Boardley, Fox, and Robinson 1999).

The bulk of the theoretical work in IRT comes from the fields of psychometrics and educational measurement with key early contributions from Rasch (1960), Birnbaum (1968), Wright and Stone (1979), and Lord (1980). Some good introductory IRT reading includes Hambleton, Swaminathan, and Rogers (1991), McDonald (1999), Embretson and Reise (2000), Bond and Fox (2015), and de Ayala (2022). More advanced treatments are presented, for example, in Fischer and Molenaar (1995), van der Linden and Hambleton (1997), Baker and Kim (2004), and De Boeck and Wilson (2004). Raykov and Marcoulides (2018) provide a comprehensive treatment of IRT using Stata.

Benjamin Drake Wright (1926–2015) was born in Wilkes-Barre, Pennsylvania. Wright joined the US Navy in 1944 and went on to study physics at Cornell University. He interned with American physicist Charles H. Townes, and after joining the physics department at the University of Chicago, he became Robert S. Mulliken’s research assistant.

His interests began to shift, and in 1957 he obtained a PhD in the philosophy of human development. When the University of Chicago received an IBM computer, Wright wrote a program for factor analysis and regression. While performing factor analyses for a market research firm, Wright became discomfited by the inconsistency of the results.

In 1960, psychometrician Georg Rasch gave a series of lectures on his measurement models at the University of Chicago, and Wright was won over by their stability. Together with Bruce Choppin, he wrote computer programs that would fit the Rasch measurement models. His advocacy in these models is reflected in his cofounding of the Rasch Measurement Social Interest Group, as part of the American Education Research Association (AERA), and the Institute for Objective Measurement, which publishes the *Journal of Applied Measurement* on a quarterly basis. Wright also developed a type of map for presenting the overall performance levels of students; this KIDMAP concept was implemented first by the Los Angeles Independent County School District in the 1980s and later by the Australian Council for Educational Research.

For his many contributions to measurement, which spanned multiple fields, Wright was honored with two conferences celebrating his work.

Frederic M. Lord (1912–2000) was born in Hanover, New Hampshire. He obtained a master's degree in educational psychology from the University of Minnesota and a PhD in psychology from Princeton University. In 1949, he became the director of statistical analysis for the Education Testing Service (ETS), where he would work for 33 years.

Lord devised models to categorize test questions based on difficulty and thus laid the foundation for item response theory. His work with the ETS had impacts on the Law School Admissions Test, the test of English as a Foreign Language, and the Graduate Record exam. Additionally, he coauthored a book with Melvin R. Novick on test theory, which was an expansion of his dissertation. His dissertation alone made a lasting impact on psychometrics, as did his other publications.

In 2000, the ETS created the Frederic M. Lord Chair in Measurement and Statistics in his honor. Because of his pioneering contributions, Lord is regarded as the “Father of Modern Testing”.

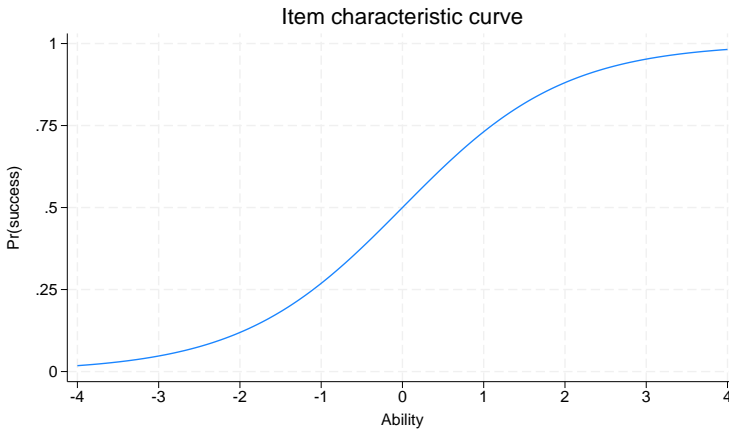
Allan Birnbaum (1923–1976) was born in San Francisco, California. He completed a premedical program prior to obtaining his PhD in mathematical statistics from Columbia University in 1954. There, among other projects, he worked on developing statistical methods applicable to the social sciences. In 1959, he joined the faculty of New York University, where he would teach statistics.

He published a total of 41 papers, but the paper published in 1962 stands out as his most significant contribution to the field of statistical theory. In this publication, he advocated for the likelihood principle, providing proof that the same inference can be made across two experiments that provided proportional likelihood functions. His approach departed from that of Abraham Wald and Erich Leo Lehmann, who influenced his dissertation. Although met simultaneously with appraise and opposition, his work had an impact on meta-analysis and predictions with missing data. Notably, renowned statistician Leonard Jimmie Savage regarded Birnbaum's work on the likelihood principle as highly influential in the field of statistics.

Birnbaum also published in the areas of classification and discrimination, and he applied his medical background to research on experimental genetics. He held faculty positions at Stanford University, New York University, and Cambridge University. The last position he held was chair of statistics at City, University of London. He was honored with election to fellowship by the American Association for the Advancement of Sciences, the American Statistical Association, and the Institute of Mathematical Statistics.

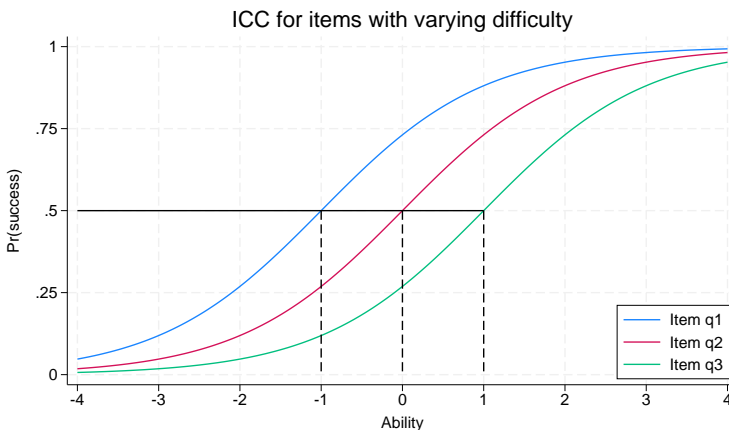
Birnbaum is remembered as a deep thinker and dedicated father.

The main concept in IRT is the item characteristic curve (ICC). The ICC describes the probability that a person “succeeds” on a given item (individual test question). In the following graph, we can see an ICC for one item intended to measure ability. Notice that the probability of this broadly defined success increases as ability increases.



ICCs will be different for different items. The probability of success on an item is a function of both the level of the latent trait and the properties of the item. The latent trait is commonly denoted by  $\theta$ . The value of  $\theta$  for a given person is called the person location. The item properties are parameters, commonly known as difficulty and discrimination, that are estimated in the IRT model.

The difficulty parameter, or item location, commonly denoted by  $b$ , represents the location of an item on the ability scale. For example, the following graph plots the ICC for items q1, q2, and q3, with difficulty parameters  $-1$ ,  $0$ , and  $1$ , respectively.



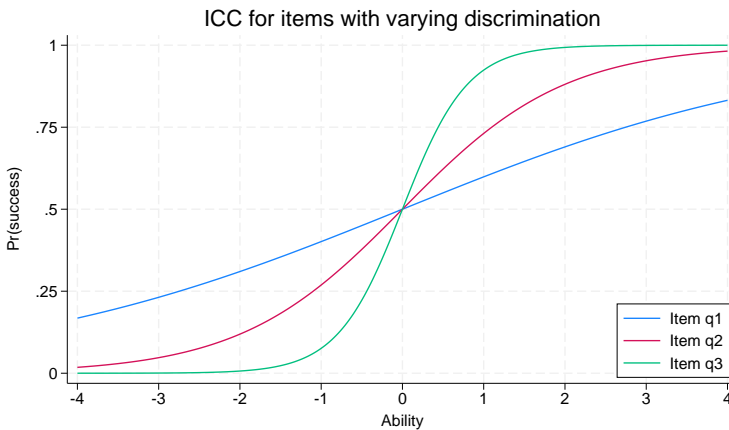
Item q1 is the least difficult, and item q3 is the most difficult. Notice that the change in difficulty shifts the ICC along the ability scale (that is, the horizontal axis or  $x$  axis). The probability of success on item q1 is higher than the probability of success for the other two items at any ability level. We can say item q1 is less difficult than the others because a person would need only an ability level greater than  $-1$

on this ability scale to be expected to succeed on item q1. On the other hand, a person would need an ability level above 0 to be expected to succeed on item q2 and an ability level above 1 to be expected to succeed on item q3.

In designing an instrument intended to differentiate between all levels of a latent trait, a researcher should try to have items with difficulties spread across the full range of the trait.

The second item parameter, discrimination, is related to the slope of the ICC. Discrimination is commonly denoted by  $a$ . This item parameter tells us how fast the probability of success changes with ability near the item difficulty. An item with a large discrimination value has a high correlation between the latent trait and the probability of success on that item. In other words, an item with a large discrimination parameter can distinguish better between low and high levels of the latent trait.

In the graph above, all three items have the same discrimination. In the graph below, all three items have the same difficulty, but they have different discrimination values. A highly discriminating item differentiates better, around its difficulty value, between persons of similar levels of the latent trait.



Imagine two persons, one with ability just below zero, and the other with ability just above zero. According to the ICC for item q1, these persons would have a similar probability of success on this item. According to the ICC for item q3, the person with the higher ability level would have a substantially higher probability of success on this item.

Using an IRT model, we can estimate the discrimination and difficulty parameters,  $a$  and  $b$ , for each item on an instrument designed to measure a particular latent trait. Throughout this manual, we assume that a single latent trait is sufficient to explain a person's response behavior on the group of items. More technically, we assume a unidimensional latent space. We also assume that after we condition on ability, a person's responses to an item are independent of his or her responses to other items. This is called a conditional independence or a local independence assumption.

We can now express a generic functional form of an ICC as

$$\text{Pr}(\text{success}|a, b, \theta) = F\{a(\theta - b)\}$$

The difference term  $(\theta - b)$  tells us that the probability of success is a function of the distance between item location and person location. When  $\theta = b$ , that is, when item difficulty is matched to a person's latent trait level, the individual is equally likely to pass or fail the item. When  $\theta > b$ , the individual is more likely to succeed than to fail. Because we can obtain the same distance with different choices of  $\theta$  and  $b$ , we need to provide a metric for  $\theta$  to identify the model. We do so by assuming  $\theta \sim N(0, 1)$ ,

which also puts the item difficulty parameter on the same scale as the standard normal distribution. With the standard normal scale, items with negative difficulties are considered to be relatively easy, and items with positive difficulties are considered to be relatively hard.

For any IRT model, we assume  $F(\cdot)$  to be of correct functional form and increasing with the value of the latent trait. Because probabilities are bounded between 0 and 1,  $F(\cdot)$  is usually a variation of a cumulative logistic distribution.

Through choices of  $F(\cdot)$  and specification of certain constraints on the estimated parameters, we can fit a variety of different types of IRT models. Using the `irt` commands, we can fit IRT models to binary, ordinal, and nominal items. Below we demonstrate an IRT model with binary items and an IRT model with ordinal items. For additional information and examples of the models available for binary items, see [IRT] [irt 1pl](#), [IRT] [irt 2pl](#), and [IRT] [irt 3pl](#). For models with ordinal items, see [IRT] [irt grm](#), [IRT] [irt rsm](#), and [IRT] [irt pcm](#). For models with nominal items, see [IRT] [irt nrm](#). Each of these models can allow parameters to differ across groups such as males and females or age categories; see [IRT] [irt, group\(\)](#). In addition to fitting the models, we can better understand each item and its relationship to the latent trait through a variety of graphs, as demonstrated in the examples below.

The `irt` commands fit IRT models via maximum likelihood estimation. See [Item response theory](#) in [BAYES] [bayesm](#) and Balov (2016) for examples of fitting IRT models to binary items using a Bayesian approach.

From a broader statistical perspective, IRT models can be viewed as extensions of (unidimensional) confirmatory factor analysis (CFA) models to binary and categorical outcomes and as special cases of generalized linear mixed-effects models; see chapter 1 in De Boeck and Wilson (2004) and chapter 3 in Skrondal and Rabe-Hesketh (2004) for a theoretical discussion and Zheng and Rabe-Hesketh (2007) for applied examples.

## ► Example 1: Binary IRT models

In this example, we present IRT analysis of binary data and highlight some postestimation features of `irt`. We use an abridged version of the mathematics and science data from De Boeck and Wilson (2004). Student responses to test items are coded 1 for correct and 0 for incorrect. Here we list the first five observations.

```
. use https://www.stata-press.com/data/r19/masc1
(Data from De Boeck & Wilson (2004))
. list in 1/5
```

	q1	q2	q3	q4	q5	q6	q7	q8	q9
1.	1	1	1	0	0	0	0	1	0
2.	0	0	1	0	0	0	0	1	1
3.	0	0	0	1	0	0	1	0	0
4.	0	0	1	0	0	0	0	0	1
5.	0	1	1	0	0	0	0	1	0

Looking across the rows, we see that the first student correctly answered items q1, q2, q3, and q8, the second student correctly answered items q3, q8, and q9, and so on.

Let's say the goal of the test is to assess students' mathematical ability and perhaps classify the students into groups, for example, gifted, average, and remedial. We could look at the total test score for each student, but the problem is that the total score depends on the composition of the test. If the test comprises easy items, most students will appear to be gifted, and if the test comprises hard items, most students

will be assigned to the remedial group. When the model fits the data, an attractive property of IRT is that, except for measurement error, parameter estimates are invariant; that is, examinee ability estimates are not test dependent, and item parameter estimates are not group dependent.

We fit a 1PL model to binary items q1–q9 as follows.

```
. irt 1pl q1-q9
Fitting fixed-effects model:
Iteration 0:  Log likelihood = -4275.6606
Iteration 1:  Log likelihood = -4269.7861
Iteration 2:  Log likelihood = -4269.7825
Iteration 3:  Log likelihood = -4269.7825
Fitting full model:
Iteration 0:  Log likelihood = -4153.3609
Iteration 1:  Log likelihood = -4142.374
Iteration 2:  Log likelihood = -4142.3516
Iteration 3:  Log likelihood = -4142.3516
One-parameter logistic model                                Number of obs = 800
Log likelihood = -4142.3516
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
	Discrim	.852123	.0458445	18.59	0.000	.7622695	.9419765
q1	Diff	-.7071339	.1034574	-6.84	0.000	-.9099066	-.5043612
q2	Diff	-.1222008	.0963349	-1.27	0.205	-.3110138	.0666122
q3	Diff	-1.817693	.1399523	-12.99	0.000	-2.091994	-1.543391
q4	Diff	.3209596	.0976599	3.29	0.001	.1295498	.5123695
q5	Diff	1.652719	.1329494	12.43	0.000	1.392144	1.913295
q6	Diff	.6930617	.1031842	6.72	0.000	.4908243	.8952991
q7	Diff	1.325001	.1205805	10.99	0.000	1.088668	1.561335
q8	Diff	-2.413443	.1691832	-14.27	0.000	-2.745036	-2.08185
q9	Diff	-1.193206	.1162054	-10.27	0.000	-1.420965	-.965448

Looking at the output table, we see that the first row reports the estimate of the item discrimination parameter, labeled Discrim. In a 1PL model, this parameter is shared by all items. The estimate of 0.85 suggests the items are not particularly discriminating; that is, in the vicinity of a given difficulty estimate, any two students with distinct abilities would have similar predicted probabilities of responding correctly



to an item. The remaining rows report the estimates of the difficulty parameters, labeled `Diff`, for each item. The items appear to cover a wide range of the item difficulty spectrum, with item q8 being the lowest ( $\hat{b}_8 = -2.41$ ) and item q5 being the highest ( $\hat{b}_5 = 1.65$ ).

We use `estat report` to arrange the output in a particular sort order, which, in our example, makes it easy to see which items are easy and which are hard; see [\[IRT\] estat report](#) for details.

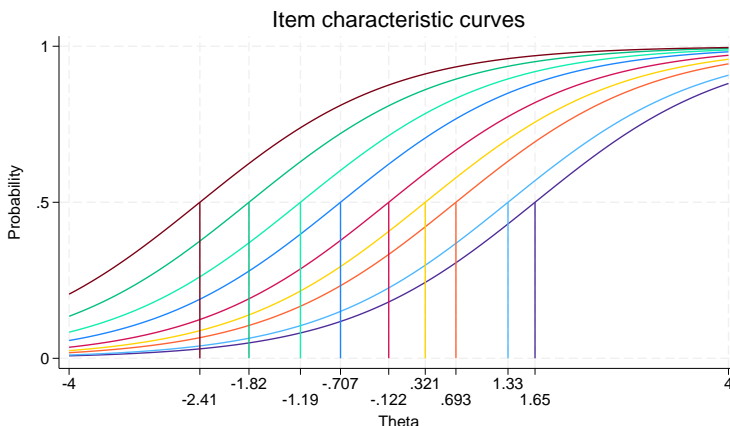
```
. estat report, sort(b) byparm
```

One-parameter logistic model Number of obs = 800  
Log likelihood = -4142.3516

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
Discrim	.852123	.0458445	18.59	0.000	.7622695	.9419765
Diff						
q8	-2.413443	.1691832	-14.27	0.000	-2.745036	-2.08185
q3	-1.817693	.1399523	-12.99	0.000	-2.091994	-1.543391
q9	-1.193206	.1162054	-10.27	0.000	-1.420965	-.965448
q1	-.7071339	.1034574	-6.84	0.000	-.9099066	-.5043612
q2	-.1222008	.0963349	-1.27	0.205	-.3110138	.0666122
q4	.3209596	.0976599	3.29	0.001	.1295498	.5123695
q6	.6930617	.1031842	6.72	0.000	.4908243	.8952991
q7	1.325001	.1205805	10.99	0.000	1.088668	1.561335
q5	1.652719	.1329494	12.43	0.000	1.392144	1.913295

To visualize the item locations on the difficulty spectrum, we plot the ICCs for all items using `irtgraph icc`; see [\[IRT\] irtgraph icc](#) for details.

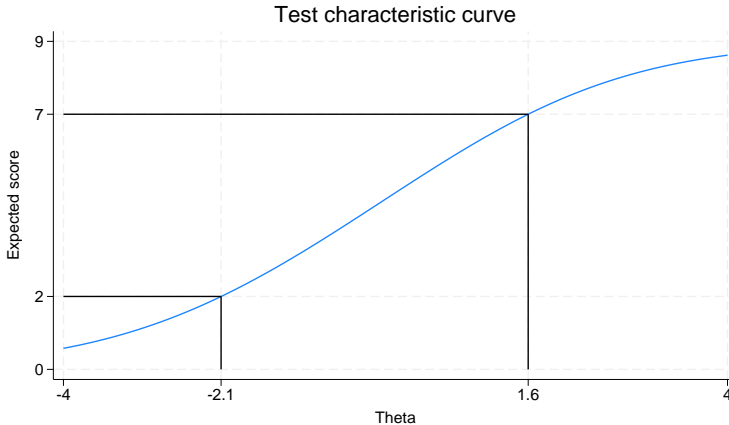
```
. irtgraph icc, blocation legend(off) xlabel(,alt)
```



The probabilities represent the expected scores for each item along the latent trait continuum. For the IPL model, the midpoint probability for each item corresponds with the estimated difficulty parameter.

The sum of the probabilities gives us the expected score on the whole test. A plot of the expected score against the latent trait is called a test characteristic curve (TCC). Below we plot the TCC for our model using `irtgraph tcc`; see [\[IRT\] irtgraph tcc](#) for details. The `scorelines(2 7)` option specifies that droplines corresponding to the expected scores of 2 and 7 also be plotted. According to the estimated TCC, these expected scores correspond with the latent trait locations  $-2.1$  and  $1.6$ , respectively.

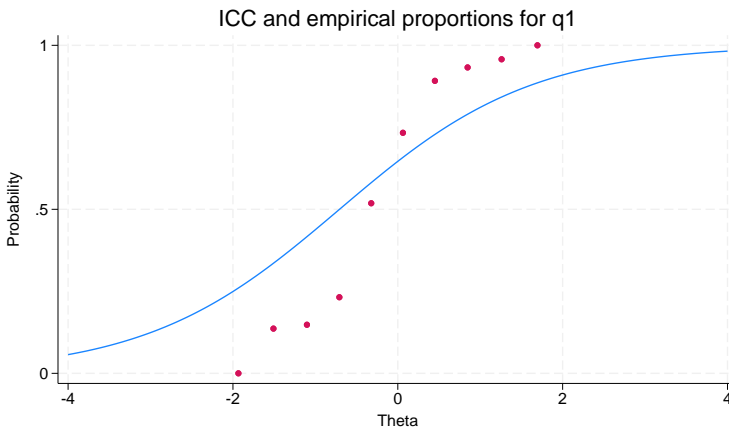
```
. irtgraph tcc, scorelines(2 7)
```



The invariance property of IRT holds only if the model fits the data. One informal method to check item fit is to superimpose empirical proportions on an ICC. If the predicted ICC follows closely the empirical trace line implied by the proportions, an item is assumed to have a satisfactory fit.

To calculate the empirical proportions, we predict the latent trait and collapse the items by the latent trait. We then call `irtgraph icc` with option `addplot()` to superimpose the proportions on the ICC.

```
. predict Theta, latent
(option ebmeans assumed)
(using 7 quadrature points)
. collapse q*, by(Theta)
. irtgraph icc q1, addplot(scatter q1 Theta)
> title("ICC and empirical proportions for q1")
```



We see that the fit of the ICC to the implied empirical trace line is poor. This is true for all items in the model. It is possible that a 2PL model may be more appropriate for this item. Before we fit a 2PL model, we store our estimates for later use.

```
. estimates store onep
```

To fit a 2PL model to the data, we type

```
. use https://www.stata-press.com/data/r19/masc1, clear
(Data from De Boeck & Wilson (2004))
. irt 2pl q1-q9
```

Fitting fixed-effects model:

```
Iteration 0: Log likelihood = -4275.6606
Iteration 1: Log likelihood = -4269.7861
Iteration 2: Log likelihood = -4269.7825
Iteration 3: Log likelihood = -4269.7825
```

Fitting full model:

```
Iteration 0: Log likelihood = -4146.9386
Iteration 1: Log likelihood = -4119.3568
Iteration 2: Log likelihood = -4118.4716
Iteration 3: Log likelihood = -4118.4697
Iteration 4: Log likelihood = -4118.4697
```

```
Two-parameter logistic model                                Number of obs = 800
Log likelihood = -4118.4697
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
q1	Discrim	1.615292	.2436467	6.63	0.000	1.137754	2.092831
	Diff	-.4745635	.074638	-6.36	0.000	-.6208513	-.3282757
q2	Discrim	.6576171	.1161756	5.66	0.000	.4299171	.885317
	Diff	-.1513023	.1202807	-1.26	0.208	-.3870481	.0844435
q3	Discrim	.9245051	.1569806	5.89	0.000	.6168289	1.232181
	Diff	-1.70918	.242266	-7.05	0.000	-2.184012	-1.234347
q4	Discrim	.8186403	.1284832	6.37	0.000	.5668179	1.070463
	Diff	.3296791	.1076105	3.06	0.002	.1187663	.5405919
q5	Discrim	.8956621	.1535128	5.83	0.000	.5947825	1.196542
	Diff	1.591164	.2325918	6.84	0.000	1.135293	2.047036
q6	Discrim	.9828441	.147888	6.65	0.000	.6929889	1.272699
	Diff	.622954	.1114902	5.59	0.000	.4044373	.8414708
q7	Discrim	.3556064	.1113146	3.19	0.001	.1374337	.5737791
	Diff	2.840278	.8717471	3.26	0.001	1.131685	4.548871
q8	Discrim	1.399926	.233963	5.98	0.000	.9413668	1.858485
	Diff	-1.714416	.1925531	-8.90	0.000	-2.091814	-1.337019
q9	Discrim	.6378452	.1223972	5.21	0.000	.3979512	.8777392
	Diff	-1.508254	.2787386	-5.41	0.000	-2.054571	-.9619361

Now each item has its own discrimination parameter that models the slope of the ICC for that item. In a 1PL model, the discrimination for all items was estimated to be 0.85. Looking at item q1 in the output table above, we see that its discrimination is estimated to be 1.62, which corresponds to a steeper slope and should result in a better item fit.

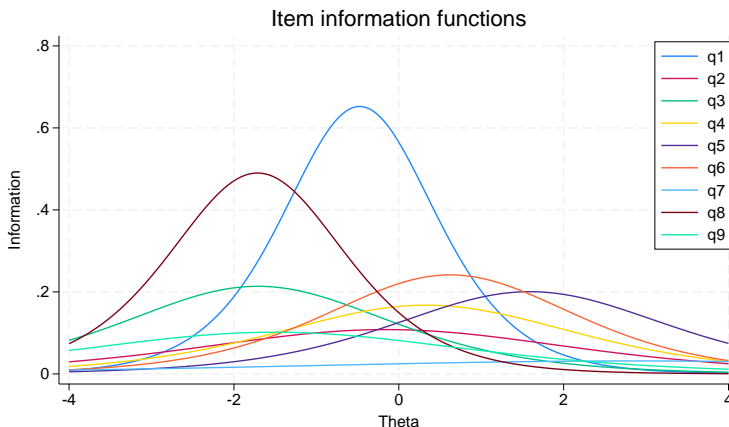
Because the 1PL model is nested in a 2PL model, we can perform a likelihood-ratio test to see which model is preferred.

```
. lrtest onep .
Likelihood-ratio test
Assumption: onep nested within .
LR chi2(8) = 47.76
Prob > chi2 = 0.0000
```

The near-zero significance level favors the model that allows for a separate discrimination parameter for each item.

Continuing with the 2PL model, we can also plot the amount of information an item provides for estimating the latent trait. A plot of item information against the latent trait is called an item information function (IIF). We use `irtgraph iif` to obtain the IIFs for all items in the model; see [\[IRT\] irtgraph iif](#) for details.

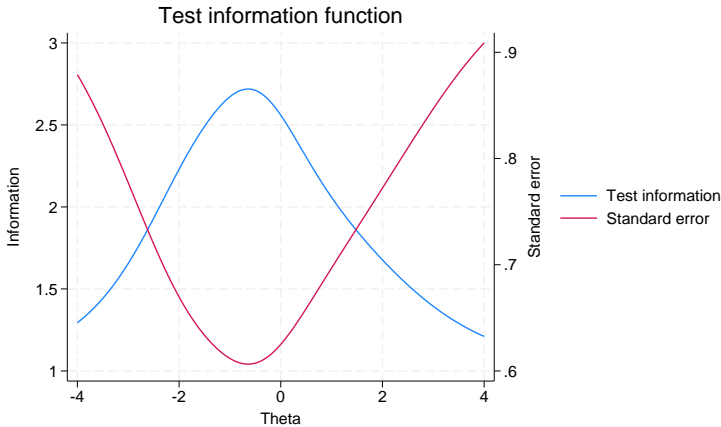
```
. irtgraph iif, legend(pos(1) ring(0) region(lcolor(black)))
```



For a 2PL model, IIFs are unimodal and symmetric, and each item provides the maximum amount of information at its estimated difficulty parameter. The height of an IIF and therefore the amount of information an item provides around the difficulty parameter is proportional to the item's estimated discrimination. Items q1 and q8 are most discriminating and have the steepest IIFs.

We can sum up all the IIFs to obtain a test information function (TIF). The TIF plot tells us how well the instrument can estimate person locations; see [\[IRT\] irtgraph tif](#) for details.

```
. irtgraph tif, se
```



The test provides maximum information for persons approximately located at  $\theta = -0.5$ . As we move away from that point in either direction, the standard error of the TIF increases, and the instrument provides less and less information about  $\theta$ .

The TIF is useful in designing instruments targeted at obtaining precise estimates of a person's latent trait level at specified intervals. If our interest lies in identifying gifted and remedial students, we would like the instrument to be more precise at the extrema of the ability range. If we wish to have a similar precision of ability estimate across the entire ability range, we would like to see a relatively flat TIF. Because the TIF is a sum of IIFs, we can obtain the desired shape of the TIF by incorporating items targeted at a specified ability interval.

◀

The last binary model, not shown here, is a 3PL model. This model adds to the 2PL model by accommodating the possibility of guessing. We discuss this model in the [\[IRT\] irt 3pl](#) entry.

## ► Example 2: Categorical IRT models

Categorical IRT models include models for ordered and unordered responses. Here we present a graded response model (GRM) for ordered responses.

The GRM is an extension of the 2PL model to categorical outcomes. To illustrate the model, we use the data from [Zheng and Rabe-Hesketh \(2007\)](#). `charity.dta` contains five survey questions, `ta1` through `ta5`, measuring faith and trust in charity organizations. Responses are strongly agree (0), agree (1), disagree (2), and strongly disagree (3). Higher scores indicate higher levels of distrust. Here we list the first five observations.

```
. use https://www.stata-press.com/data/r19/charity
(Data from Zheng & Rabe-Hesketh (2007))
. list in 1/5, nlabel
```

	ta1	ta2	ta3	ta4	ta5
1.	.	2	1	1	.
2.	0	0	0	0	0
3.	1	1	2	0	2
4.	1	2	2	0	1
5.	.	1	1	1	1

Looking across the first row, we see that the first respondent did not provide an answer to items `ta1` and `ta5`, answered 2 on item `ta2`, and answered 1 on items `ta3` and `ta4`. All `irt` commands exclude missing items for a given observation from the likelihood calculation but keep the nonmissing items for that observation. If you wish to remove the entire observation from the model, add the `listwise` option at estimation time.

We fit a GRM as follows:

```
. irt grm ta1-ta5
Fitting fixed-effects model:
Iteration 0: Log likelihood = -5559.6414
Iteration 1: Log likelihood = -5473.9434
Iteration 2: Log likelihood = -5467.4082
Iteration 3: Log likelihood = -5467.3926
Iteration 4: Log likelihood = -5467.3926
Fitting full model:
Iteration 0: Log likelihood = -5271.0634
Iteration 1: Log likelihood = -5162.5917
Iteration 2: Log likelihood = -5159.2947
Iteration 3: Log likelihood = -5159.2791
Iteration 4: Log likelihood = -5159.2791
Graded response model
Log likelihood = -5159.2791
Number of obs = 945
```

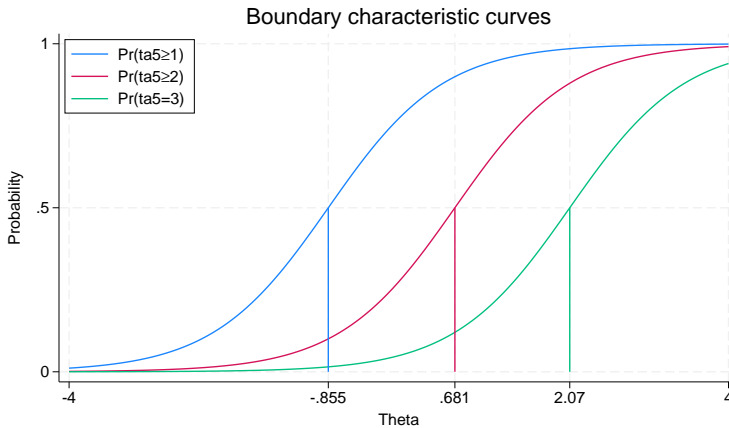
		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
ta1	Discrim	.907542	.0955772	9.50	0.000	.7202142	1.09487
	Diff						
	>=1	-1.540098	.1639425			-1.861419	-1.218776
	>=2	1.296135	.1427535			1.016343	1.575927
	=3	3.305059	.3248468			2.668371	3.941747
ta2	Discrim	.9434675	.0967483	9.75	0.000	.7538444	1.133091
	Diff						
	>=1	-1.661331	.167878			-1.990366	-1.332296
	>=2	.0068314	.082222			-.1543208	.1679836
	=3	2.531091	.2412513			2.058247	3.003935
ta3	Discrim	1.734201	.1554383	11.16	0.000	1.429548	2.038855
	Diff						
	>=1	-1.080079	.0835119			-1.243759	-.9163983
	>=2	1.016567	.0796635			.8604297	1.172705
	=3	2.232606	.1497814			1.93904	2.526172
ta4	Discrim	1.93344	.1857629	10.41	0.000	1.569351	2.297528
	Diff						
	>=1	-.3445057	.0578468			-.4578833	-.2311282
	>=2	1.466254	.0983823			1.273428	1.65908
	=3	2.418954	.162392			2.100672	2.737237
ta5	Discrim	1.42753	.1263962	11.29	0.000	1.179798	1.675262
	Diff						
	>=1	-.8552358	.0833158			-1.018532	-.6919399
	>=2	.6805315	.07469			.5341418	.8269211
	=3	2.074243	.1538858			1.772632	2.375853

Because the GRM is derived in terms of cumulative probabilities, the estimated category difficulties represent a point at which a person with ability equal to a given difficulty has a 50% chance of responding in a category equal to or higher than the difficulty designates; see [\[IRT\] irt grm](#) for details. For example, looking at the estimated parameters of item ta5, we see that a person with  $\theta = -0.86$  has a 50% chance

of answering 0 versus greater than or equal to 1, a person with  $\theta = 0.68$  has a 50% chance of answering 0 or 1 versus greater than or equal to 2, and a person with  $\theta = 2.07$  has a 50% chance of answering 0, 1, or 2 versus 3.

We can use `irtgraph icc` to plot these probabilities; here we show them for item `ta5` together with the estimated category difficulties. In a GRM, the midpoint probability for each category is located at the estimated category difficulty.

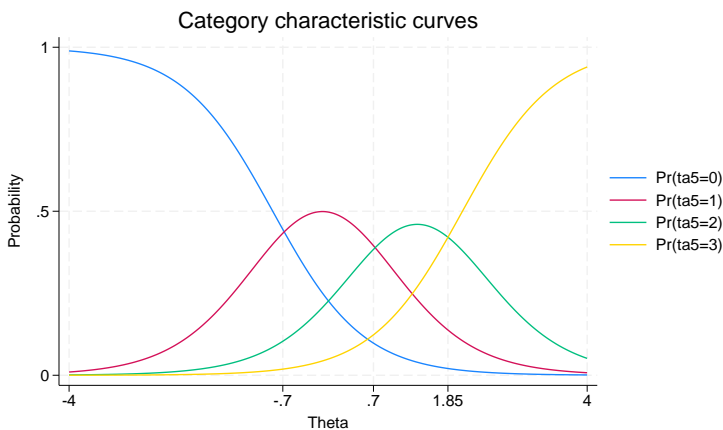
```
. irtgraph icc ta5, blocation legend(pos(11) ring(0) region(lcolor(black)))
```



When we plot characteristic curves for categorical items in ways reminiscent of ICCs for binary items, the resulting curves are called boundary characteristic curves (BCCs).

We can also plot the probabilities of respondents choosing exactly category  $k$ . For categorical items, the resulting curves are called category characteristic curves (CCCs). In fact, this is the default behavior of `irtgraph icc`.

```
. irtgraph icc ta5, xlabel(-4 -.7 .7 1.85 4)
```

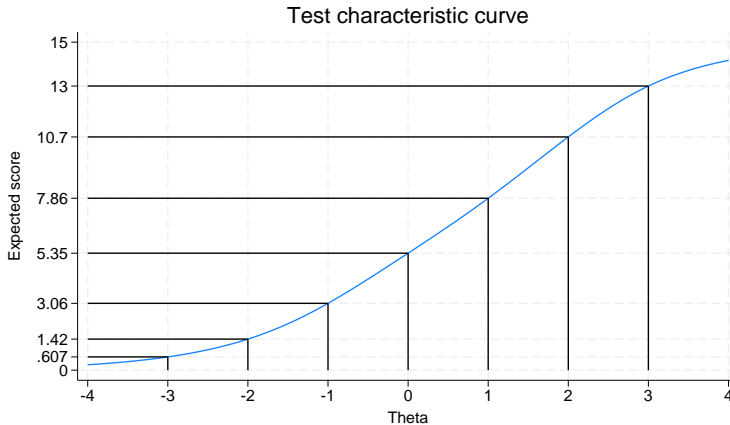




The points where the adjacent categories cross represent transitions from one category to the next. Thus, respondents with low levels of distrust, below approximately  $\theta = -0.7$ , are most likely to choose the first category on item ta5 (strongly agree), respondents located approximately between  $-0.7$  and  $0.7$  are most likely to choose the second category on item ta5 (agree), and so on.

As in the first example, we can plot the test characteristic function for the whole instrument.

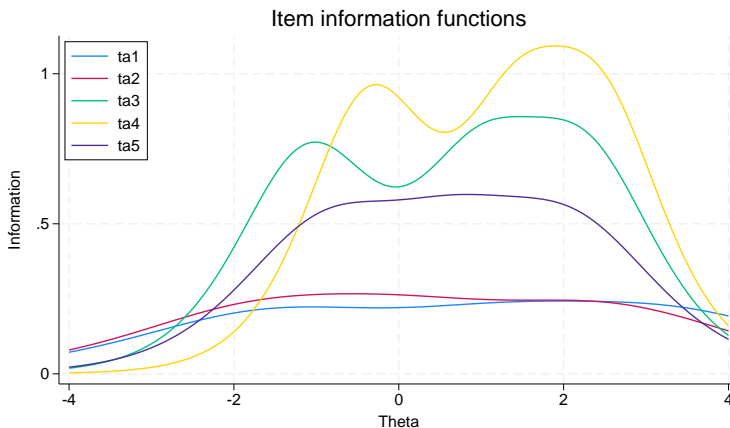
```
. irtgraph tcc, thetalines(-3/3)
```



Because we have 5 items, each with a minimum score of 0 and a maximum score of 3, the expected score ranges from 0 to 15. We also asked `irtgraph icc` to plot the expected scores for different values of  $\theta$ . For respondents located at  $\theta = -3$  and below, the expected score is less than 1, which means those respondents are most likely to choose the answer coded 0 on each and every item.

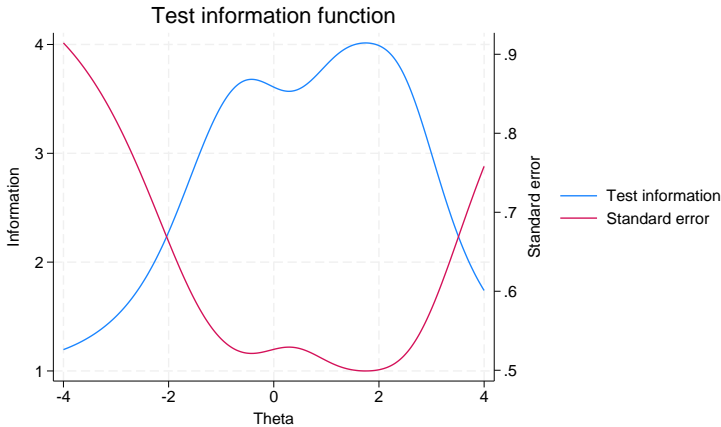
For categorical items, the item information function is no longer unimodal or symmetric, because each category contributes its own information, which may peak over a different ability range. We see this in the graph below.

```
. irtgraph iif, legend(pos(11) ring(0) region(lcolor(black)))
```



Because the test information function is the sum of the individual IIFs, its plot will also exhibit peaks and valleys.

```
. irtgraph tif, se
```



◀

In the above example, we presented the GRM. The `irt` command also supports other models for categorical responses; see [IRT] `irt nrm` for a discussion of the nominal response model (NRM), [IRT] `irt pcm` for a discussion of the partial credit model (PCM), and [IRT] `irt rsm` for a discussion of the rating scale model (RSM).

In addition to binary and categorical IRT models, the `irt` command allows you to apply different models to subsets of items and perform a single calibration for the whole instrument. We call such models hybrid IRT models; see [IRT] `irt hybrid` for a further discussion and examples.

## References

- Baker, F. B., and S.-H. Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. 2nd ed, revised and expanded. Boca Raton, FL: CRC Press.
- Balov, N. 2016. Bayesian binary item response theory models using bayesmh. *The Stata Blog: Not Elsewhere Classified*. <https://blog.stata.com/2016/01/18/bayesian-binary-item-response-theory-models-using-bayesmh/>.
- Birnbaum, A. 1968. “Some latent trait models and their use in inferring an examinee’s ability”. In *Statistical Theories of Mental Test Scores*, edited by F. M. Lord and M. R. Novick, 395–479. Reading, MA: Addison–Wesley.
- Boardley, D., C. M. Fox, and K. L. Robinson. 1999. Public policy involvement of nutrition professionals. *Journal of Nutrition Education* 31: 248–254. [https://doi.org/10.1016/S0022-3182\(99\)70460-7](https://doi.org/10.1016/S0022-3182(99)70460-7).
- Bond, T. G., and C. M. Fox. 2015. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 3rd ed. New York: Routledge.
- de Ayala, R. J. 2022. *The Theory and Practice of Item Response Theory*. 2nd ed. New York: Guilford Press.
- De Boeck, P., and M. Wilson, eds. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer. <https://doi.org/10.1007/978-1-4757-3990-9>.
- Embretson, S. E., and S. P. Reise. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fischer, G. H., and I. W. Molenaar, eds. 1995. *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer.
- Hambleton, R. K., H. Swaminathan, and H. J. Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

- King, J., and T. G. Bond. 1996. A Rasch analysis of a measure of computer anxiety. *Journal of Educational Computing Research* 14: 49–65. <https://doi.org/10.2190/URRN-X4N9-V74C-U621>.
- Kondratak, B. 2022. `uirt`: A command for unidimensional IRT modeling. *Stata Journal* 22: 243–268.
- Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ: Lawrence Erlbaum.
- McDonald, R. P. 1999. *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum.
- Perrot, B., E. Bataille, and J.-B. Hardouin. 2018. `validscale`: A command to validate measurement scales. *Stata Journal* 18: 29–50.
- Raciborski, R. 2015. Spotlight on irt. *The Stata Blog: Not Elsewhere Classified*. <https://blog.stata.com/2015/07/31/spotlight-on-irt/>.
- Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute of Educational Research.
- Raykov, T., and G. A. Marcoulides. 2018. *A Course in Item Response Theory and Modeling with Stata*. College Station, TX: Stata Press.
- Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC.
- van der Linden, W. J., and R. K. Hambleton, eds. 1997. *Handbook of Modern Item Response Theory*. New York: Springer.
- Wright, B. D., and M. H. Stone. 1979. *Best Test Design: Rasch Measurement*. Chicago: MESA Press.
- Wu, A. W., R. D. Hays, S. Kelly, F. Malitz, and S. A. Bozzette. 1997. Applications of the Medical Outcomes Study health-related quality of life measures in HIV/AIDS. *Quality of Life Research* 6: 531–554. <https://doi.org/10.1023/A:1018460132567>.
- Zheng, X., and S. Rabe-Hesketh. 2007. Estimating parameters of dichotomous and ordinal item response models with `gllamm`. *Stata Journal* 7: 313–333.

## Also see

[IRT] [Glossary](#)

[IRT] [DIF](#) — Introduction to differential item functioning

[SEM] [gsem](#) — Generalized structural equation model estimation command

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow and NetCourseNow are trademarks of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.

For suggested citations, see the FAQ on [citing Stata documentation](#).

