

# **A practical introduction to Item Response Theory (IRT) using Stata 14**

*Malcolm Rosier*

*Survey Design and Analysis Services*

*Email: [mrosier@tpg.com.au](mailto:mrosier@tpg.com.au)*

1

## **Overview**

- *Measurement in the social sciences*
- *Problems with Classical Test Theory (CTT)*
- *Introduction to IRT*
- *Using One-Parameter (1PL) and Two-Parameter (2PL) Logistic Models*
- *Using Rating Scale Model (RSM) and Graded Response Model (GRM)*
- *Merits and uses of IRT*

2

## Measurement in social science

- We use the term “social science” for research where we are measuring characteristics of people in order to develop theories about their behaviour.
- We need to find numbers that validly represent the behaviour or other traits of the people: isomorphic with reality. We can only apply our interpretive statistics once we have the numbers.
- We try to locate each person at a point on some hypothesised underlying dimension or psychological construct or latent trait.
- We try to measure all variables carefully, but not all variables are of equal importance. Criterion variables are more important than predictor variables.
- We especially need high quality variables where the stakes are high, such as large-scale school testing programs like the National Assessment Program – Literacy and Numeracy (NAPLAN), or assessments with financial implications, like a driving test or a residency visa.

3

## Characteristics of scales

- We usually combine a series of individual variables into a single composite variable such as a test or a scale. This has the advantage of:
- **Parsimony:** a single variable can be used for description or in multivariate analyses instead of a series of separate variables.
- **Stability:** the composite variable can give a more reliable estimate of the quality being measured than any of the separate constituent variables.
- **Normality:** a scale is based on a set of variable means and hence will be more normal than the constituent variables.

4

## Levels of measurement

### Categorical (or nominal):

- Mutually exclusive but unordered categories. We can assign numbers but they are arbitrary.

### Ordinal:

- Numbers reflect an order. Often termed Likert scales.

### Interval:

- Differences between pairs of values are meaningful across the scale.

### Ratio:

- Interval level plus a real zero.

5

## Valid statistical procedures for each level

OK to compute	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median, percentiles	No	Yes	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratio, coefficient of variation	No	No	No	Yes

6

## Classical Test Theory: person-item matrix

Person	1	2	3	4	5	6	Total
1	1	1	1	1	1	1	6
2	1	1	1	1	1	0	5
3	1	1	1	1	0	1	5
4	1	1	1	0	1	.	4
5	1	1	1	0	0	0	3
6	1	1	0	0	0	1	3
7	1	1	0	0	0	0	2
8	1	1	0	0	0	0	2
9	1	0	0	1	.	0	2
10	1	0	0	.	0	.	1
Total	10	8	5	4	3	3	

7

## CTT: comments on person-item matrix

- Items are arranged by difficulty. Persons are arranged by ability.
- The persons and items are interdependent. We can't measure persons without items. We can't measure items without persons.
- Persons can obtain the same score with different patterns of responses.
- Missing responses are counted as incorrect. Maybe missing responses at the end of a test just mean the person is a slow worker.
- The top (and bottom) scores do not tell us how much higher (or lower) the person's ability may be.

## Classical Test Theory: test and item statistics

- **Item facility** (difficulty). Percentage or proportion of persons who have the correct response.
- **Point-biserial correlation coefficient**. Capacity of the item to discriminate students on the dimension being measured by the test. Calculated as the correlation across persons between the item score and the test score excluding that item.
- **Reliability**. Quality of the test as a whole. Calculated by repeated measures, or correlating half of the items with the other half.
- **Internal consistency**. Proportion of the total variance among the items that is due to the common factor (latent variable). Assessed by the Cronbach alpha coefficient.
- **Unidimensionality**. All items/variables associated with a common latent trait or dimension. Tested with factor analysis.

9

## Classical Test Theory: some problems

- The key problem is that the evaluation of the quality of the items and the test as a whole depends on the **particular group of persons** and the **particular group of items** used for the evaluation.
- Item difficulty depends on the group.
- Item discrimination depends on the group.
- Reliability measures such as split halves or test-retest are essentially artificial.
- Test scores are bounded between the maximum and minimum values, and hence cannot form an interval scale. It follows that parametric statistics cannot be validly used.
- Assumptions of linearity, homogeneity and normality underlying the use of factor analysis and regression analysis are probably violated.

10

## Item Response Theory (IRT)

- Item Response Theory (IRT) reflects more precisely the relationship between the measurement process and the underlying dimension or latent trait being measured.
- IRT adopts an explicit model for the probability of each possible response to an item. This probability is derived as a function of the latent trait and some item parameters, and then used to obtain the likelihood of ability as a function of the observed responses and item parameters.
- IRT produces an interval level scale. The same scale provides a measure of item difficulty and of person ability.
- The calibration of the scale is carried out by maximum likelihood estimation involving iteration between the item values across persons and the person values across items.
- Each person answering the test is assigned a value on the scale. The standard error of the score for each person each item indicates the fit of each person may be assessed, reflecting the pattern of responses by that person.

11

## Three basic components of IRT

### **Item Response Function (IRF):**

- A mathematical function that relates the latent trait to the probability of endorsing an item.

### **Item Information Function (IIF):**

- An indication of item quality, and the item's ability to differentiate among respondents.

### **Invariance:**

- The position on the latent trait continuum that can be estimated by any items with known IRFs and item characteristics. These are population independent within a linear transformation.

12

## Item Response Function (IRF)

### Definitions

$b_i$  difficulty of item  $i$

$\theta_j$  ability of person  $j$

- Each cell in the person-item matrix represents an encounter between a person of ability  $\theta$  and an item of difficulty  $b$ .
- We can improve the estimates by taking account of item discrimination.

$a_i$  discrimination of item  $i$

13

## Item Response Function (IRF): basic one-parameter model (1)

- Consider an encounter between a person of ability  $\theta$  and an item of difficulty  $b$ . Since a deterministic response is not acceptable, the response must be expressed in terms of probabilities. This gives us the one-parameter logistic model for the probability of a correct response.
- If  $\theta > b$ ,  $p(1) \rightarrow 1.0$
- If  $\theta < b$ ,  $p(1) \rightarrow 0.0$
- If  $\theta = b$ ,  $p(1) \rightarrow 0.5$
- A valid measurement model must accommodate these three possible encounters, which may be prepared in four stages.

14

## Item Response Function (IRF): basic one-parameter model (2)

- The probability of a correct response must be in the range from 0 to 1.
- The possible range of  $(\theta - b)$  is from  $-\infty$  to  $+\infty$ . This is accommodated first by taking  $\exp(\theta - b)$  which can range from 0 to  $+\infty$ .
- The following ratio is formed which can range from 0 to 1.  
$$\frac{\exp(\theta - b)}{1 + \exp(\theta - b)}$$
- The one-parameter logistic model for the probability of a correct response is defined as:  
$$p(x_{ni}) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$
- This basic equation for IRT implies that the probability of a person obtaining the correct answer to an item depends both on the ability of the person and the difficulty of the item.

15

## IRT assumptions

- **Monotonicity.** The probability of a person endorsing an item increases as the person's latent trait level increases.
- **Unidimensionality.** All items are contributing in the same way to the underlying latent trait.
- **Invariance.** Person trait levels do not depend on which items are administered nor on the particular sample of persons (subject to linear transformation). This enables linking of scales measuring the same construct. We can compare persons even if they responded to different items
- **Local independence.** Item responses are independent given a person's ability. They are uncorrelated after controlling for the latent trait.
- **Qualitatively homogeneous populations.** The same IRFs applies to all members of the population.

16



## One-parameter (1PL) and two-parameter (2PL) logistic models

### One-parameter (1PL).

- Test for item difficulty ( $a$ ).
- Constrain discrimination ( $b$ ) to be constant.

### Two-parameter (2PL).

- Test for item difficulty ( $a$ ) and discrimination ( $b$ )

17

## Data for 1PL and 2PL examples

```
. webuse masc1
. format q1 q2 q3 q4 q5 q6 q7 q8 q9 %4.3f
. summarize q1 q2 q3 q4 q5 q6 q7 q8 q9, format sep(0)
```

Variable	Obs	Mean	Std Dev	Min	Max
q1	800	0.627	0.484	0.000	1.000
q2	800	0.522	0.500	0.000	1.000
q3	800	0.795	0.404	0.000	1.000
q4	800	0.441	0.497	0.000	1.000
q5	800	0.226	0.419	0.000	1.000
q6	800	0.375	0.484	0.000	1.000
q7	800	0.273	0.446	0.000	1.000
q8	800	0.860	0.347	0.000	1.000
q9	800	0.708	0.455	0.000	1.000

Note. These are the data cited in the Stata 14 manual [IRT].

18

## Summated scores

```
. egen score9_total = rowtotal(q1 q2 q3 q4 q5 q6 q7 q8 q9)
. alpha q1 q2 q3 q4 q5 q6 q7 q8 q9, gen(score9_alpha) detail item
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
q1	800	+	0.5815	0.3684	0.0191	0.4593
q2	800	+	0.4696	0.2229	0.0228	0.5110
q3	800	+	0.4360	0.2371	0.0236	0.5058
q4	800	+	0.4911	0.2500	0.0220	0.5014
q5	800	+	0.4375	0.2308	0.0235	0.5074
q6	800	+	0.5131	0.2834	0.0213	0.4896
q7	800	+	0.3455	0.1131	0.0263	0.5430
q8	800	+	0.4551	0.2893	0.0233	0.4949
q9	800	+	0.4104	0.1797	0.0244	0.5235
Test					0.0229	0.5342

Note. The test alpha would be higher if we dropped item q7.

19

## Principal components analysis

```
. pca q1 q2 q3 q4 q5 q6 q7 q8 q9, components(1)
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.9701	0.9157	0.2189	0.2189
Comp2	1.0544	0.0344	0.1172	0.3361
Comp3	1.0199	0.0230	0.1133	0.4494
Comp4	0.9968	0.1137	0.1108	0.5602
Comp5	0.8831	0.0232	0.0981	0.6583
Comp6	0.8598	0.0694	0.0955	0.7538
Comp7	0.7904	0.0390	0.0878	0.8416
Comp8	0.7513	0.0775	0.0835	0.9251
Comp9	0.6738	.	0.0749	1.0000

Note. The eigenvalue for the first factor is about double that for the second factor, but accounting for only 22% of the variance.

20

## Principal components eigenvectors

Principal components (eigenvectors)

Variable	Comp1	Unexplained
q1	0.4560	0.5903
q2	0.3045	0.8173
q3	0.3300	0.7854
q4	0.3412	0.7706
q5	0.3171	0.8019
q6	0.3615	0.7426
q7	0.1630	0.9476
q8	0.3773	0.7195
q9	0.2716	0.8546

### Post-estimation

We can score the test using the post-estimation `predict` command, which uses the eigenvectors as scoring weights.

```
. predict score9_pca
```

21

## One-parameter logistic model (1PL): commands

```
. irt 1pl q1 q2 q3 q4 q5 q6 q7 q8 q9
```

### Options for estat report

`byparm`: arrange table rows by parameter rather than by item

`sort(a)`: specify that items be sorted according to the estimated discrimination parameters

`sort(b)`: specify that items be sorted according to the estimated difficulty parameters

```
. set cformat %6.3f
```

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

22

## 1PL: results

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Discrim	0.852	0.046	18.59	0.000	0.762	0.942
Diff						
q8	-2.413	0.169	-14.27	0.000	-2.745	-2.082
q3	-1.818	0.140	-12.99	0.000	-2.092	-1.543
q9	-1.193	0.116	-10.27	0.000	-1.421	-0.965
q1	-0.707	0.103	-6.84	0.000	-0.910	-0.504
q2	-0.122	0.096	-1.27	0.205	-0.311	0.067
q4	0.321	0.098	3.29	0.001	0.130	0.512
q6	0.693	0.103	6.72	0.000	0.491	0.895
q7	1.325	0.121	10.99	0.000	1.089	1.561
q5	1.653	0.133	12.43	0.000	1.392	1.913

Note. We provide a metric for  $\theta$  in order to identify the model. We assume a standard normal distribution for ability with mean = 0 and standard deviation = 1. The discrimination is constrained to be equal for all items. The discrimination value of 0.852 is not very good.

23

## 1PL: post-estimation

Prediction of the latent trait in IRT models involves assigning a value to the latent trait. Empirical Bayes combines the prior information about the latent trait with the likelihood to obtain the conditional posterior distribution of the latent trait.

### Options for post-estimation

`latent` specifies that the latent trait is predicted using an empirical Bayes estimator

`se(newvar)` calculates standard errors of the empirical Bayes estimator and stores the result in `newvar`.

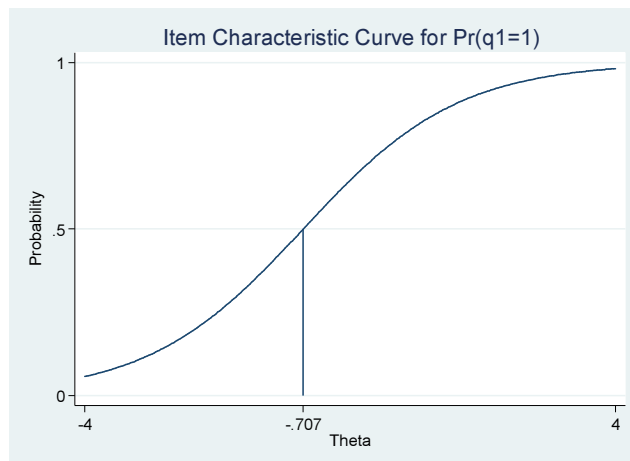
```
. predict score9_pred_1pl, latent se(score9_se_1pl)
```

Note. Stata 14 has a context sensitive post-estimation menu box under the Statistics tab. Or use the command `postest`.

24

## Item Characteristic Curve (ICC): Item q1

```
. irtgraph icc q1, blocation
```



25

## Comment on ICC for q1

For this ICC of item q1 the **blocation** option adds a vertical line at the estimated item difficulty of -0.707. That is, the ability of a person with a probability of 0.5 of making a correct response. The probability of success increases as ability increases.

The ability range  $\theta$  is shown from -4.0 to +4.0.

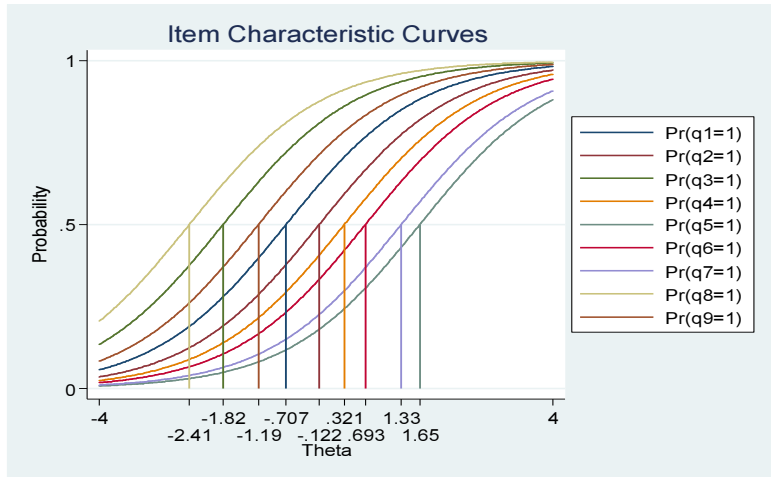
Where  $\theta_j$  is greater than  $b_p$ , the probability of a correct response is greater than 0.50.

Where  $\theta_j$  is less than  $b_p$ , the probability of a correct response is less than 0.50.

Where  $\theta_j$  is equal to  $b_p$ , the probability of a correct response equals 0.50.

## Item Characteristic Curve (ICC): All items

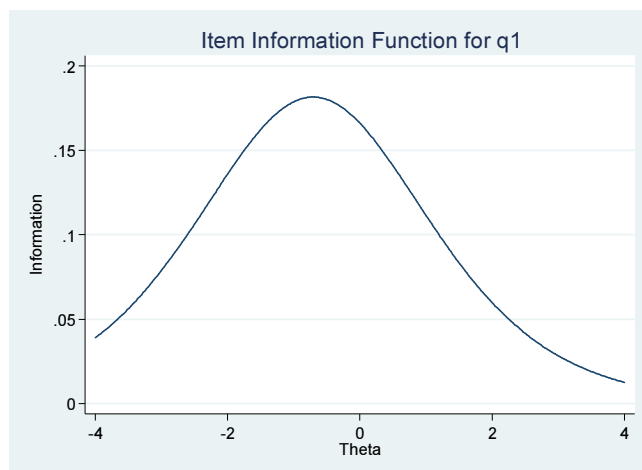
```
. irtgraph icc q1 q2 q3 q4 q5 q6 q7 q8 q9, blocation legend(pos(3)  
col(1) ring(1) xlabel(,alt)
```



27

## Item Information Function (IIF): Item q1

```
. irtgraph iif q1
```



28

## Comment on IIF for item q1

Item Information Function (IIF) replaces item reliability as used in CTT

The IIF describes how well or precisely an item measures at each level of the trait that is being measured by a given test ( $\theta$ ). In IRT, the term "information" is used to describe reliability or precision of an item or a whole instrument.

A major advantage of IRT is that both items and people are placed on the same scale (usually a standard score scale, with mean = 0.0, and standard deviation = 1.0) so that people can be compared to items and vice-versa.

The standard error of measurement (SEM) is the variance of the latent trait. It is the reciprocal of information, so that more information means less error. Measurement error is expressed on the same metric as the latent trait, so can be used to build confidence levels.

Difficulty: the location on the ability dimension of the highest information point.

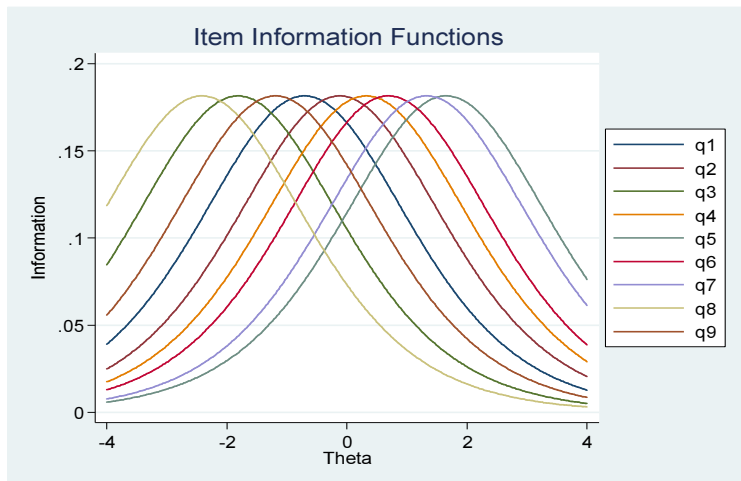
Discrimination: the height of the information.

High discriminations: tall, narrow IIFs: high precision, narrow range.

Low discriminations: short, wide IIFs: low precision, broad range.

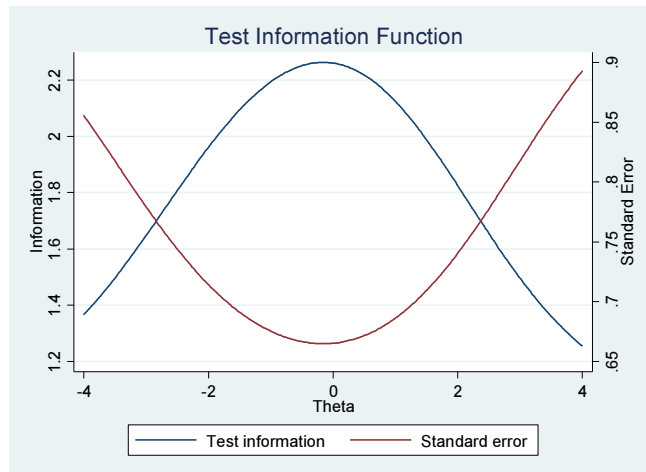
## Item Information Function (IIF): All items

```
. irtgraph iif, legend(pos(3) col(1) ring(1))
```



## Test Information Function (TIF)

```
. irtgraph tif, se
```



31

## Comment on TIF

The IIFs are additive and can be combined into the Test Information Function (TIF). We can judge the test as a whole, and see which parts of the trait range are working best.

The scale of test information is on the left, plotted in blue. We add the `se` option to include the standard error. The scale of the standard error is on the right, plotted in red. The standard error is lower where we have more information from more items in the centre of the ability range. The standard error increases at the extremes of the range where we have less information.

The TIF is useful in test development where, depending on the specific needs, the test can be chosen to cover the whole spectrum or to focus on a particular range of the ability scale. For tests with alternate formats, TIFs are used to ensure the formats carry the same information across the targeted latent trait range. If we want to develop alternative versions of tests to measure the same trait, we can use TIFs to balance the items across the versions.



## Two-parameter logistic model (2PL): results (1)

```
. irt 2pl q1 q2 q3 q4 q5 q6 q7 q8 q9
. estat report, byparm sort(b)
```

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

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<b>Discrim</b>					
q8	1.400	0.234	5.98	0.000	0.941 1.858
q3	0.925	0.157	5.89	0.000	0.617 1.232
q9	0.638	0.122	5.21	0.000	0.398 0.878
q1	1.615	0.244	6.63	0.000	1.138 2.093
q2	0.658	0.116	5.66	0.000	0.430 0.885
q4	0.819	0.128	6.37	0.000	0.567 1.070
q6	0.983	0.148	6.65	0.000	0.693 1.273
q5	0.896	0.154	5.83	0.000	0.595 1.197
q7	0.356	0.111	3.19	0.001	0.137 0.574

33

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

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<b>Diff</b>					
q8	-1.714	0.193	-8.90	0.000	-2.092 -1.337
q3	-1.709	0.242	-7.05	0.000	-2.184 -1.234
q9	-1.508	0.279	-5.41	0.000	-2.055 -0.962
q1	-0.475	0.075	-6.36	0.000	-0.621 -0.328
q2	-0.151	0.120	-1.26	0.208	-0.387 0.084
q4	0.330	0.108	3.06	0.002	0.119 0.541
q6	0.623	0.111	5.59	0.000	0.404 0.841
q5	1.591	0.233	6.84	0.000	1.135 2.047
q7	2.840	0.872	3.26	0.001	1.132 4.549

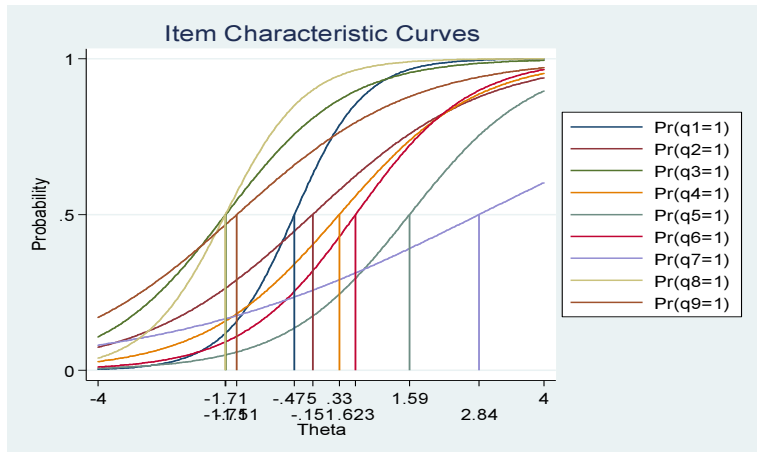
### Post-estimation

```
. predict score9_pred_2pl, latent se(score9_se_2pl)
```

34

## Item Characteristic Curve (ICC): All items

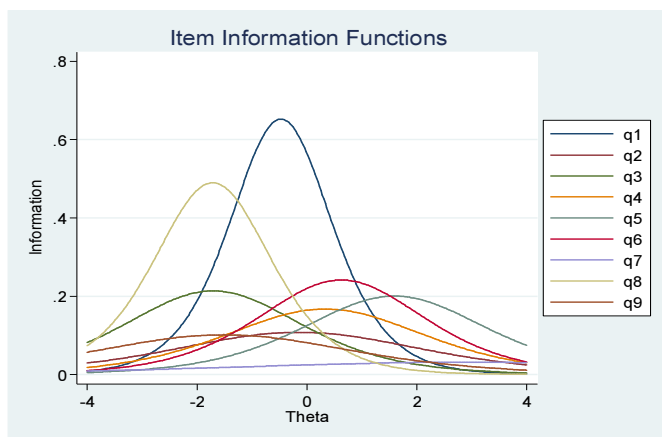
```
. irtgraph icc q1 q2 q3 q4 q5 q6 q7 q8 q9, blocation  
legend(pos(3) col(1) ring(1) xlabel(,alt))
```



35

## Item Information Function (IIF): All items

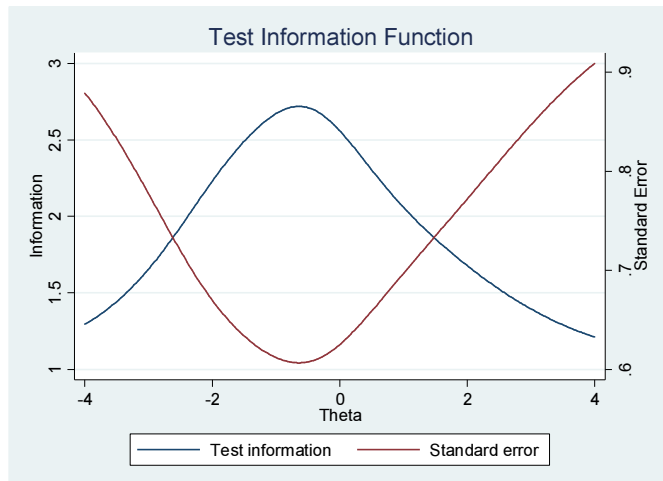
```
. irtgraph iif
```



36

## Test Information Function (TIF)

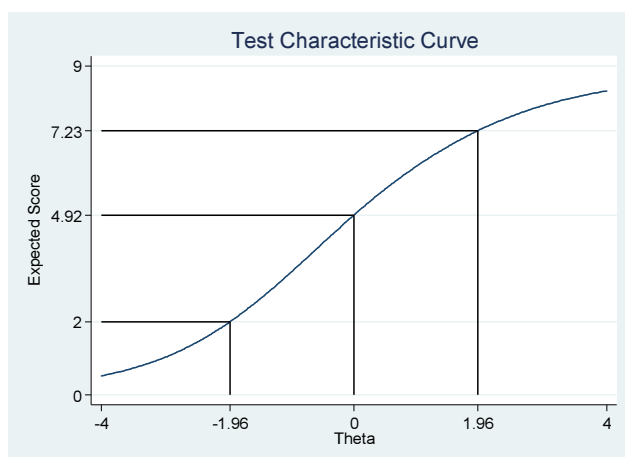
```
. irtgraph tif
```



37

## Test Characteristic Curve (TCC): Version 1

```
. irtgraph tcc, thetalines(-1.96 0 1.96)
```



38

## Comment on 2PL graphs

The ICC for all items shows the discrimination varying across items. Steeper curves better discrimination. The results clearly show that item q7 has a very poor discrimination of 0.356.

The IIFs for the 2pl model combine the two item parameters. The location of the centre of the IIF reflects the difficulty of the item, the height of the IIF reflects the item discrimination. Item q8 is easiest, so that the peak of its IIF is on the left end of the theta continuum. Item q7 is the most difficult, and also the least discriminating. On the basis of all this information, from both the traditional and IRT measures, it appears that we should re-run the analyses dropping q7.

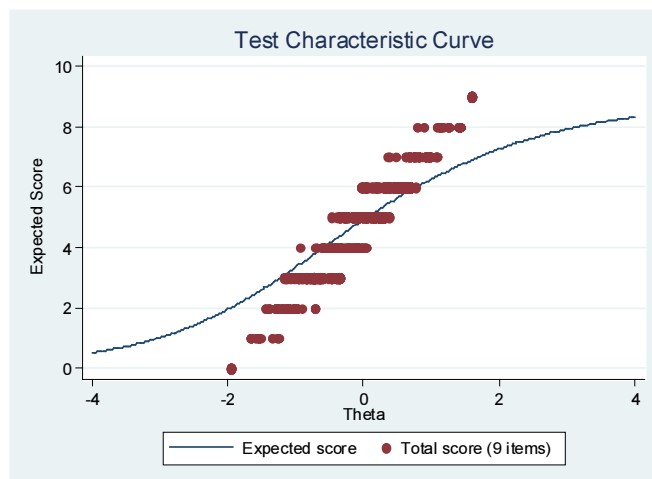
The TIF shows where we have most information about the set of items in the test. The graph shows the standard errors, which are higher where we have less information.

The TCC shows the relationship between the ability estimated by IRT 2PL and the expected score. The **thetalines** option lets us see the expected score corresponding to specified ability locations.

This plot tells us what kind of scores to expect from individuals with different levels of the latent trait. For example, we can expect above-average individuals to score 4.92 or above. Using the 95% critical values from the standard normal distribution (-1.96 and +1.96), this plot also tells us that we can expect 95% of randomly selected people to score between about 2 and 7.

## Test Characteristic Curve (TCC): Version 2

```
. irtgraph tcc, addplot(scatter score9_total score9_pred_2pl)
```



## 1PL and 2PL item responses

```
. list q3 q8 q3 q9 q1 q2 q4 q6 q5 q7 score9_pca score9_pred_1pl
  score9_pred_2pl if score9_total == 5 in 1/125, sep(0) noobs
```

q8	q3	q9	q1	q2	q4	q6	q5	q7	score pca	score pred_1pl	score pred_2pl
1	1	1	0	0	0	0	1	1	-0.146	0.062	-0.254
1	1	1	0	1	1	0	0	0	0.026	0.062	-0.166
1	1	1	0	1	0	1	0	0	0.086	0.062	-0.100
1	1	1	0	0	1	0	1	0	0.175	0.062	-0.071
0	1	1	1	1	0	1	0	0	-0.058	0.062	-0.014
1	0	1	1	1	1	0	0	0	0.152	0.062	0.113
1	1	1	1	1	0	0	0	0	0.282	0.062	0.156
1	1	0	1	0	0	1	0	1	0.188	0.062	0.174
1	1	1	1	0	1	0	0	0	0.360	0.062	0.223
1	1	0	1	1	1	0	0	0	0.372	0.062	0.231
1	1	1	1	0	0	0	1	0	0.430	0.062	0.255
1	1	0	1	1	0	0	1	0	0.443	0.062	0.264
1	1	1	1	0	0	1	0	0	0.419	0.062	0.292
1	1	0	1	1	0	1	0	0	0.432	0.062	0.301

41

## Comment on 2PL TCC item responses

We specified a TCC with an added a plot (red spots) of the summated score versus ability (predicted score). This shows the different patterns of responses that constitute each summated score. For example, the ability corresponding to a summated score of 5 ranges across the ability continuum from about  $-0.5$  to  $+0.5$ .

We list a sample of respondents with summated score = 5. The constituent items are arranged in order of difficulty and the persons are arranged in order of their IRT 2pl predicted score. The IRT scores form an interval scale for the underlying latent trait theta of mathematical ability that is independent of the specific test; that is, of the specific set of items and persons.

We see that the summated score = 5 can be made up in many different ways. Most, but not all members of this sample, make the correct response on the two easiest items q3 and q8. Only a few make the correct response on the two hardest items q5 and q7. Both the pca and the 2pl scores reflect this. They are correlated 0.99. Note that the one-parameter estimates are the same for all respondents. The 1pl model (Rasch) does not take account of the different discriminations or the pattern of item responses.

## 1PL and 2PL likelihood-ratio test

We compare the 1PL and 2PL models by performing a likelihood-ratio test. The likelihood ratio expresses how many times more likely the data are under one model than the other. The logarithm of the likelihood ratio is used to compute a chi-square to measure goodness of fit. This is compared to a critical value to calculate a p-value to decide whether to reject the null model in favour of the alternative model.

```
. estimates store score9_masc_1pl  
. estimates store score9_masc_2pl  
. lrtest score9_masc_1pl score9_masc_2pl, stats
```

```
Likelihood-ratio test   LR chi2(8)= 47.76  
(Assumption: score9_masc_pred_1pl nested in  
score9_masc_pred_2pl)  
Prob > chi2 = 0.0000
```

```
-----  
      Model      Obs  ll(model)  df  
-----  
score9_masc_1pl  800  -4142.35  10  
score9_masc_2pl  800  -4118.47  18  
-----
```

43

## 1PL and 2PL likelihood-ratio test calculations

The likelihood ratio test compares the goodness of fit of two models. The null or unrestricted model is 1PL. It is a special case of the alternative model or restricted model 2PL.

$D = 2 \ln(\text{likelihood null model} / \text{likelihood alternative model})$

$D = [2 \ln(\text{likelihood 1PL model})] - [2 \ln(\text{likelihood 2PL model})]$

$D = 2 * (\log\text{-likelihood 1PL model} - \log\text{-likelihood 2PL model})$

$D = 2 * (4142.35 - 4118.47) = 47.76$

Degrees of freedom:  $18 (2\text{PL}) - 10 (1\text{PL}) = 8 \text{ df}$

D follows a chi-squared distribution. The probability associated with this chi-squared value is 0.0000.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are measures of the relative quality of statistical models for a given set of data. They are based on the likelihood function.

44

## Ordered response models

Scales may include categorical items; that is, two or more response categories for each item. Examples: descriptive scales, attitude scales, and performance measures.

The IRT models estimate the probability associated with each category of each response.

Stata 14 handles three types of IRT models for ordered responses:

- Graded Response Model (GRM)
- Rating Scale Model (RSM)
- Partial Credit Model (PCM)

Stata 14 also handles unordered categorical models:

- Nominal Response Model (NRM).

45

## Data for RSM and GRM examples: describe

Trait: distrust in charity organizations

```
. webuse charity.dta  
. describe
```

```
-----  
variable  storage display value  
name      type    format  label  variable label  
-----  
ta1       byte    %17.0g  agree  Charitable Organizations More Effective  
ta2       byte    %17.0g  agree  Degree of Trust  
ta3       byte    %17.0g  agree  Charitable Organizations Honest/Ethical  
ta4       byte    %17.0g  agree  Role Improving Communities  
ta5       byte    %17.0g  agree  Job Delivering Services  
-----
```

```
. label list
```

```
agree:  
  0 strongly agree  
  1 agree  
  2 disagree  
  3 strongly disagree
```

46

### Data for RSM and GRM examples: summarize

```
. format ta1 ta2 ta3 ta4 ta5 %4.3f  
. summarize ta1 ta2 ta3 ta4 ta5, format
```

Variable	Obs	Mean	Std. Dev.	min	Max
ta1	885	1.102	0.826	0.000	3.000
ta2	912	1.418	0.935	0.000	3.000
ta3	929	1.068	0.790	0.000	3.000
ta4	934	0.773	0.772	0.000	3.000
ta5	923	1.138	0.939	0.000	3.000

47

### Data for RSM and GRM examples: tabulation

```
. tabm ta1 ta2 ta3 ta4 ta5
```

variable	strongly agree	agree	disagree	strongly disagree	total
ta1	203	447	177	58	885
ta2	185	263	362	102	912
ta3	205	511	158	55	929
ta4	372	438	88	36	934
ta5	266	350	221	86	923

Note. We have used here the `tabm` function that you need to download and install.

48



## Data for RSM and GRM examples: skewness

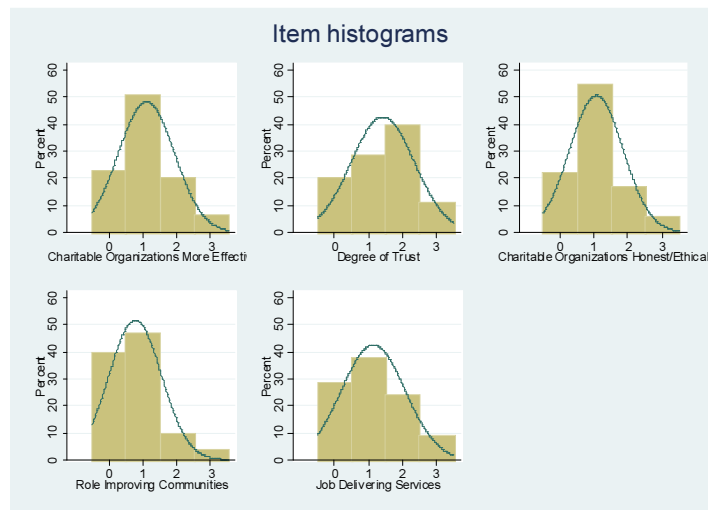
```
. sktest ta1 ta2 ta3 ta4 ta5
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	Prob>chi2
ta1	885	0.0000	0.3558	29.90	0.0000
ta2	912	0.2369	0.0000	.	0.0000
ta3	929	0.0000	0.2121	41.64	0.0000
ta4	934	0.0000	0.0007	.	0.0000
ta5	923	0.0000	0.0000	69.08	0.0000

Note. All items except **ta2** are significantly skewed. This is common for Likert items.

49

## RSM and GRM skewness histograms



50

## Cronbach alpha coefficient

```
. alpha ta1 ta2 ta3 ta4 ta5, gen(score_ta_alpha) item
```

```
Test scale = mean(unstandardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
ta1	885	+	0.5829	0.3307	0.2336	0.6473
ta2	912	+	0.6177	0.3303	0.2225	0.6529
ta3	929	+	0.6903	0.4820	0.1937	0.5785
ta4	934	+	0.6898	0.4833	0.1970	0.5808
ta5	923	+	0.7011	0.4437	0.1840	0.5913
Test					0.2063	0.6629

51

## Principal components (factor) analysis

```
. pca ta1 ta2 ta3 ta4 ta5, components(1)
```

```
Principal components/correlation      Number of obs      =      832
                                      Number of comp.    =         1
                                      Trace               =         5
Rotation: (unrotated = principal)     Rho                 = 0.4368
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.1841	1.2752	0.4368	0.4368
Comp2	0.9088	0.1589	0.1818	0.6186
Comp3	0.7499	0.1018	0.1500	0.7686
Comp4	0.6480	0.1391	0.1296	0.8982
Comp5	0.5089	.	0.1018	1.0000

```
. predict score_ta_pca
```

52

## RSM: results (1)

```
. irt rsm ta1 ta2 ta3 ta4 ta5
. estat report, byparm
```

```
Rating scale model           Number of obs = 945
Log likelihood = -5401.6068
```

The procedure `irt rsm` fits rating scale models (RSMs) to ordered categorical responses. This model is an extension of the 1PL model to ordered categorical items. In this case we have more than two categories.

The RSM constrains the difference between the difficulty parameters between adjacent categories to be equal across the items. Due to these constraints, the RSM requires that all items have the same number of responses.

The responses are assumed to be functionally equivalent; that is, the responses should have the same meaning across all items. This applies to our data, which have the same response set for agreement for all items.

We then `predict` each person's score; that is, their location on the "ability" range. In this case their ability  $\theta$  refers to their "distrust in charity organizations".

53

## RSM: results (2)

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Discrim	0.804	0.039	20.46	0.000	0.727	0.881
Diff						
ta1=1 vs 0	-1.258	0.121	-10.37	0.000	-1.496	-1.020
ta1=2 vs 1	-0.163	0.132	-1.23	0.218	-0.422	0.096
ta1=3 vs 2	0.895	0.153	5.84	0.000	0.595	1.196
ta2=1 vs 0	-1.483	0.127	-11.69	0.000	-1.731	-1.234
ta2=2 vs 1	-0.388	0.135	-2.88	0.004	-0.652	-0.124
ta2=3 vs 2	0.671	0.154	4.35	0.000	0.368	0.973
ta3=1 vs 0	-1.339	0.121	-11.07	0.000	-1.576	-1.102
ta3=2 vs 1	-0.244	0.131	-1.86	0.063	-0.501	0.013
ta3=3 vs 2	0.814	0.152	5.36	0.000	0.517	1.112
ta4=1 vs 0	-0.061	0.100	-0.62	0.538	-0.257	0.134
ta4=2 vs 1	1.033	0.128	8.10	0.000	0.783	1.283
ta4=3 vs 2	2.092	0.157	13.34	0.000	1.784	2.399
ta5=1 vs 0	-0.809	0.108	-7.48	0.000	-1.021	-0.597
ta5=2 vs 1	0.286	0.126	2.27	0.023	0.039	0.533
ta5=3 vs 2	1.344	0.151	8.91	0.000	1.049	1.640

54

## Comment on RSM results

The discrimination is held constant across all items at 0.804.

The difficulty is shown as the probability for each response category compared to the previous category.

The following abbreviated table shows only the coefficients with the differences between adjacent pairs for each item. The pattern of differences for rsm is the same for all items.

## RSM: difference between categories

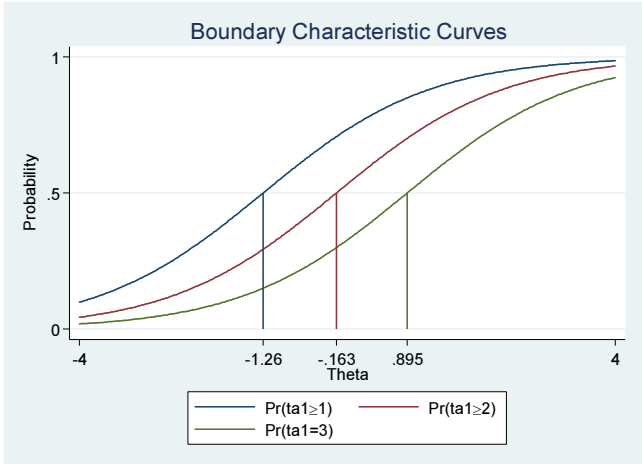
	Coef.	Difference
ta1=1 vs 0	-1.258	
ta1=2 vs 1	-0.163	1.095
ta1=3 vs 2	0.895	1.058
ta2=1 vs 0	-1.483	
ta2=2 vs 1	-0.388	1.095
ta2=3 vs 2	0.671	1.059
ta3=1 vs 0	-1.339	
ta3=2 vs 1	-0.244	1.095
ta3=3 vs 2	0.814	1.058
ta4=1 vs 0	-0.061	
ta4=2 vs 1	1.033	1.094
ta4=3 vs 2	2.092	1.059
ta5=1 vs 0	-0.809	
ta5=2 vs 1	0.286	1.095
ta5=3 vs 2	1.344	1.058

### Post-estimation

```
. predict score_ta_rsm_pred, latent se(score_ta_rsm_se)
```

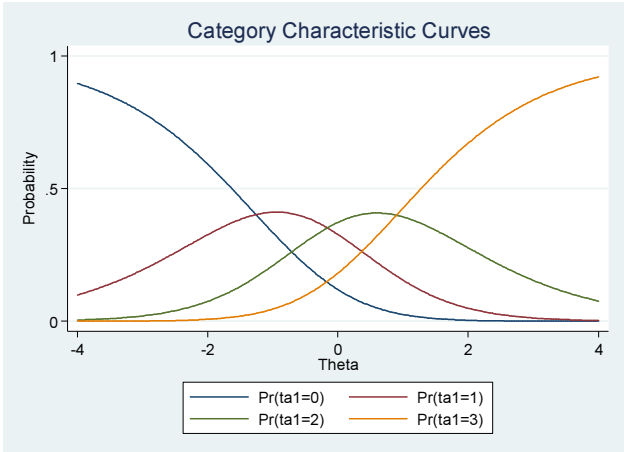
# Boundary Characteristic Curve (BCC): Item ta1

```
. irtgraph icc ta1, blocation
```

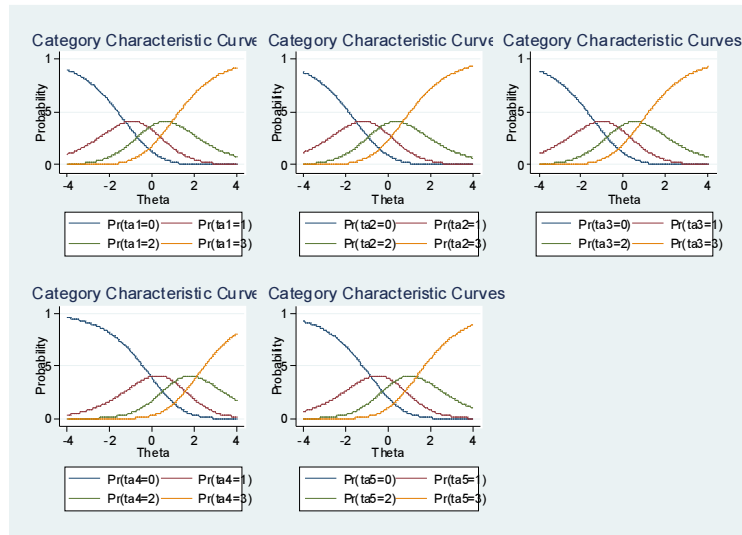


# Category Characteristic Curve (CCC): Item ta1

```
irtgraph icc ta1, ccc
```



## Category Characteristic Curve (CCC): All items



59

## Comment on RSM graphs

The ICC for RSM produces a Boundary Characteristic Curve (BCC) if the option **blocation** is specified.

The ICC for RSM produces a Category Characteristic Curve (CCC) by default or if the option **ccc** is specified. The CCC shows the probability of each response versus the theta values that measure the location of the latent trait of distrust in charity organizations.

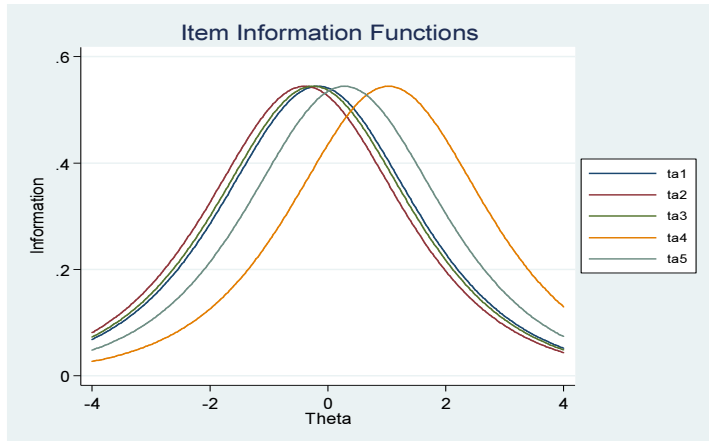
The first line  $Pr(ta1 = 0)$  shows the probability of selecting a 0 response for persons of different ability (theta).

The second line  $Pr(ta1 = 1)$  shows the probability of selecting a 1 response for persons of different ability (theta).

The points where the adjacent categories cross represent transitions from one category to the next. Thus, respondents with low levels of distrust, below approximately  $-1.4$ , are most likely to choose the first category on item ta1 (strongly agree), respondents located approximately between  $-1.4$  and  $0.0$  are most likely to choose the second category on item ta1 (agree), and so on.

## Item Information Function (IIF): All items

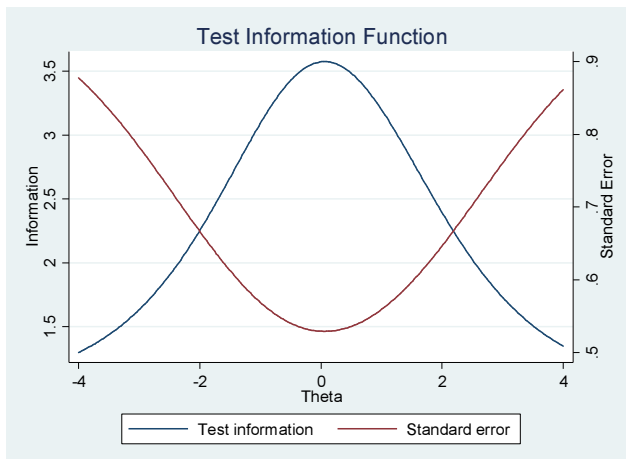
```
. irtgraph iif
```



61

## Test Information Function (TIF)

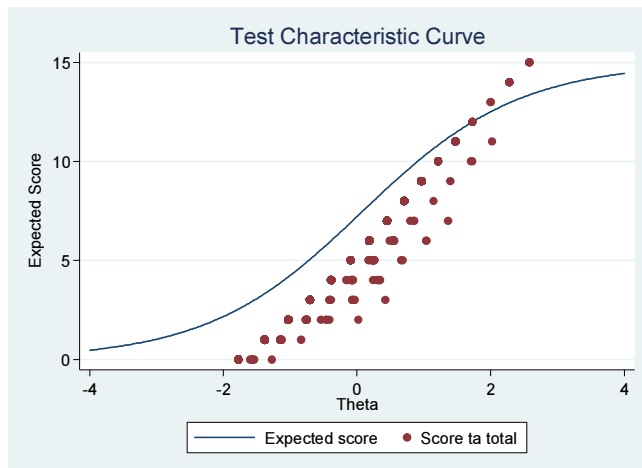
```
. irtgraph tif
```



62

## Test Characteristic Curve (TCC)

```
. irtgraph tcc, addplot(scatter score_ta_total score_ta_rsm_pred)
```



63

## GRM: results (1)

```
. irt grm ta1 ta2 ta3 ta4 ta5
. estat report, byparm
```

Graded response model                      Number of obs = 945  
Log likelihood = -5159.2791

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<b>Discrim</b>					
ta1	0.908	0.096	9.50	0.000	0.720 1.095
ta2	0.943	0.097	9.75	0.000	0.754 1.133
ta3	1.734	0.155	11.16	0.000	1.430 2.039
ta4	1.933	0.186	10.41	0.000	1.569 2.298
ta5	1.428	0.126	11.29	0.000	1.180 1.675

64



## GRM: results (2)

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----						
Diff						
ta1>=1	-1.540	0.164	-9.39	0.000	-1.861 -1.219	
ta1>=2	1.296	0.143	9.08	0.000	1.016 1.576	
ta1=3	3.305	0.325	10.17	0.000	2.668 3.942	
ta2>=1	-1.661	0.168	-9.90	0.000	-1.990 -1.332	
ta2>=2	0.007	0.082	0.08	0.934	-0.154 0.168	
ta2=3	2.531	0.241	10.49	0.000	2.058 3.004	
ta3>=1	-1.080	0.084	-12.93	0.000	-1.244 -0.916	
ta3>=2	1.017	0.080	12.76	0.000	0.860 1.173	
ta3=3	2.233	0.150	14.91	0.000	1.939 2.526	
ta4>=1	-0.345	0.058	-5.96	0.000	-0.458 -0.231	
ta4>=2	1.466	0.098	14.90	0.000	1.273 1.659	
ta4=3	2.419	0.162	14.90	0.000	2.101 2.737	
ta5>=1	-0.855	0.083	-10.26	0.000	-1.019 -0.692	
ta5>=2	0.681	0.075	9.11	0.000	0.534 0.827	
ta5=3	2.074	0.154	13.48	0.000	1.773 2.376	

65

## GRM: difference between categories

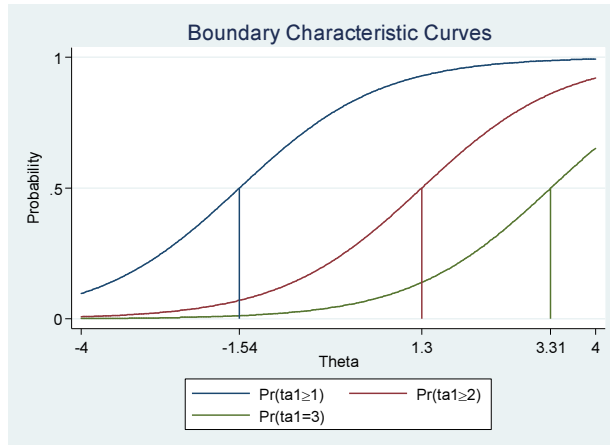
	Coef.	Difference
-----		
ta1>=1	-1.540	
ta1>=2	1.296	2.836
ta1=3	3.305	2.009
ta2>=1	-1.661	
ta2>=2	0.007	1.668
ta2=3	2.531	2.524
ta3>=1	-1.080	
ta3>=2	1.017	2.097
ta3=3	2.233	1.216
ta4>=1	-0.345	
ta4>=2	1.466	1.811
ta4=3	2.419	0.953
ta5>=1	-0.855	
ta5>=2	0.681	1.536
ta5=3	2.074	1.393

. predict score\_ta\_grm\_pred, latent se(score\_ta\_grm\_se)

66

## Boundary Characteristic Curve (BCC): Item ta1

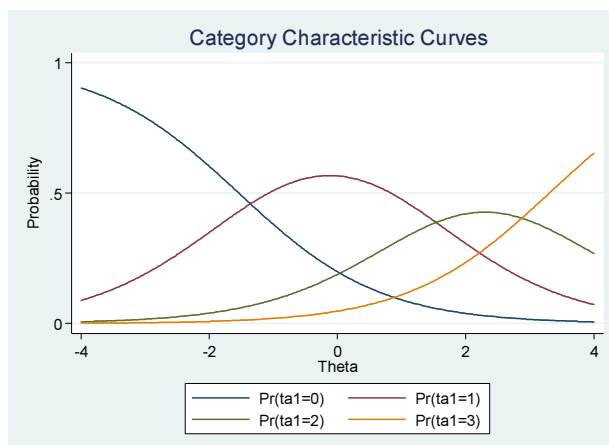
```
.irtgraph icc ta1, blocation
```



67

## Category Characteristic Curve (CCC): Item ta1

```
.irtgraph icc ta1, ccc
```



68

## Comment on GRM graphs

The ICC of a GRM produces a Boundary Characteristic Curve (BCC) if the **blocation** option is specified. For item ta1 this shows the probability of each response relative to prior responses versus the theta values that measure the location of the latent trait of distrust in charity organizations.

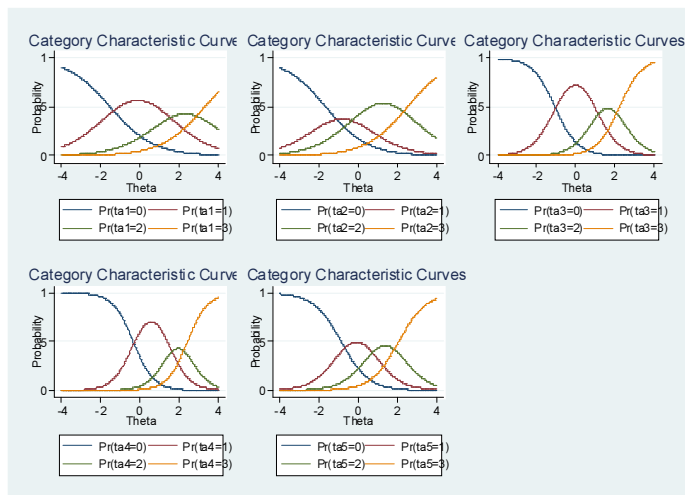
a person with  $\theta = -1.54$  has a 50% chance of responding 0 rather than responding greater or equal to 1

a person with  $\theta = 1.30$  has a 50% chance of responding 0 or 1 rather than responding greater than or equal to 2

a person with  $\theta = 3.31$  has a 50% chance of responding 0, 1, or 2 rather than responding equal to 3.

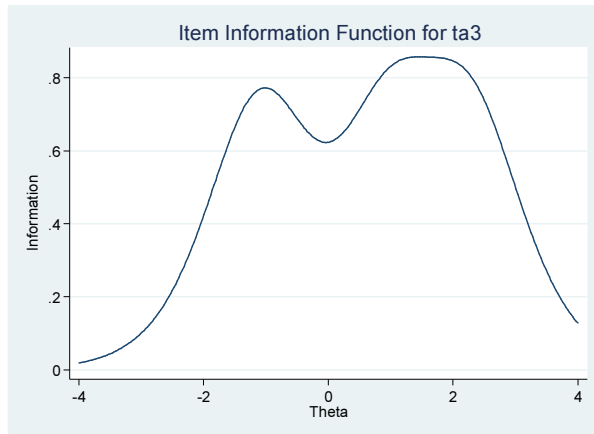
The ICC of a GRM produces a Category Characteristic Curve (CCC) if the option **ccc** is specified (although it may be omitted since this is the default). The CCC for item ta1 shows the probability of each response versus the theta values that measure the location of the latent trait of distrust in charity organizations.

## Category Characteristic Curve (CCC): All items



## Item Information Function (IIF): Item ta3

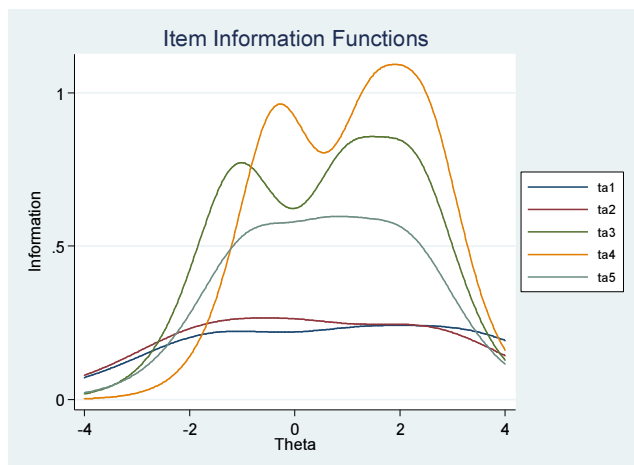
```
.irtgraph iif ta3
```



71

## Item Information Function (IIF): All items

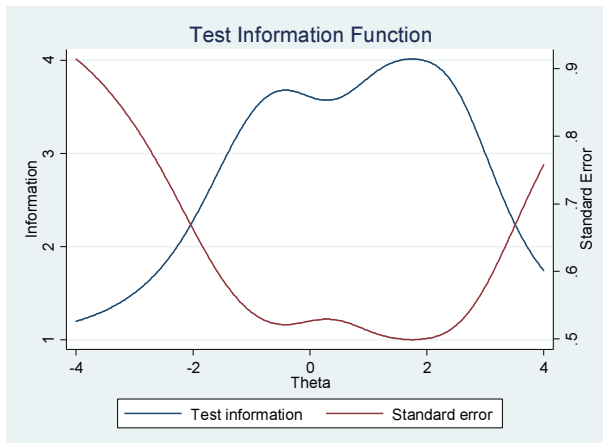
```
.irtgraph iif
```



72

## Test Information Function (TIF)

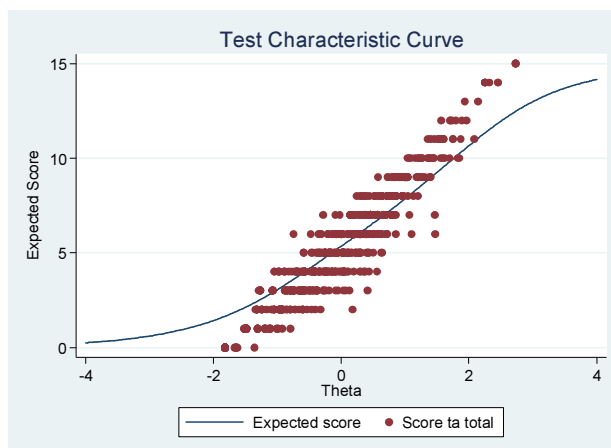
```
.irtgraph iif
```



73

## Test Characteristic Curve (TCC)

```
.irtgraph tcc
```



74

## RSM and GRM item responses

ta1	ta2	ta4	ta3	ta5	rsm_pred	rsc_se	grm_pred	grm_se
1	3	0	0	1	-0.096	0.596	-0.583	0.550
0	3	0	1	1	-0.096	0.596	-0.340	0.538
.	2	0	1	1	-0.073	0.637	-0.264	0.533
1	2	0	1	1	-0.096	0.596	-0.253	0.512
0	2	0	1	2	-0.096	0.596	-0.197	0.545
.	2	1	0	.	-0.075	0.692	-0.191	0.617
1	1	0	1	2	-0.096	0.596	-0.185	0.528
1	2	1	0	1	-0.096	0.596	-0.166	0.522
1	1	1	0	2	-0.096	0.596	-0.090	0.538
1	2	1	1	0	-0.096	0.596	-0.052	0.523
1	1	1	1	1	-0.096	0.596	0.032	0.494
.	1	1	1	1	-0.073	0.637	0.044	0.512
2	0	1	1	1	-0.096	0.596	0.087	0.514
0	1	1	1	2	-0.096	0.596	0.116	0.518
.	.	1	1	1	-0.029	0.685	0.133	0.537
1	1	1	2	0	-0.096	0.596	0.152	0.538
0	2	1	2	0	-0.096	0.596	0.160	0.554
1	0	1	1	2	-0.096	0.596	0.175	0.520
1	0	1	2	1	-0.096	0.596	0.289	0.520
0	0	1	2	2	-0.096	0.596	0.402	0.537

75

## Comment on RSM and GRM responses

For each item the respondents can select 0, 1, 2 or 3. Or they may have a missing response.

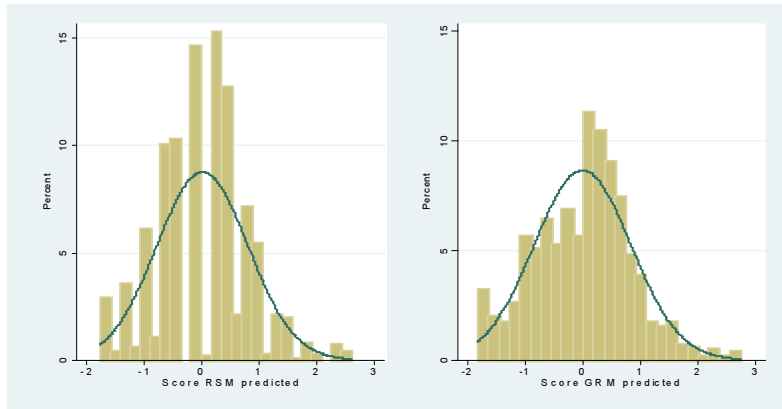
The items are arranged in order of difficulty in selecting response = 3 (strongly disagree).

There are many possible patterns that can produce a summated score = 5.

The RSM scores are that same for all persons with 5 valid responses. Ditto for those with 4 or 3 valid responses. The standard errors are also the same, being higher for cases with more missing data.

The GRM scores reflect the variability of the pattern of responses. The standard errors also vary, still being higher for cases with more missing data.

## Compare RSM and GRM: distribution of scores



77

## Compare RSM and GRM: skewness

```
. sktest score_ta_rsm_pred score_ta_grm_pred
```

Skewness/Kurtosis tests for Normality				----- joint -----	
Variable	Obs	Pr(Skew)	Pr(Kurt)	adj chi2(2)	Prob>chi2
score_ta_rsm_pred	945	0.3697	0.0960	3.57	0.1679
score_ta_grm_pred	945	0.1475	0.2661	3.33	0.1892

78

## RSM and GRM likelihood-ratio test

```
. estimates store score_rsm
. estimates store score_grm
. lrtest score_rsm score_grm, stats

Likelihood-ratio test
  LR chi2(12) =    484.66
  Prob > chi2 =    0.0000
(Assumption: score_rsm nested in score_grm)
```

```
-----
      Model  Obs  ll(model)  df
-----
  score_rsm  945  -5401.607    8
  score_grm  945  -5159.279   20
-----
```

79

## Summary comparison of IRT with CTT (1)

IRT models have unique features that complement CTT-based measures.

CTT item and person characteristics are group dependent. IRT models are invariant: the scores for subjects from that population can be compared directly even if they answer different subsets of the items.

CTT reliability is based on parallel tests, which are difficult to achieve in practice. IRT reliability is a function conditional on the scores of the measured latent construct.

Missing values are difficult to handle in CTT. The IRT maximum likelihood estimation enables analysis of items with data missing at random.

80



## Summary comparison of IRT with CTT (2)

CTT focuses on the test. IRT focuses on individual items. In the estimation process, the influence of the particular group used to calibrate the scale is minimised by adjusting for the mean ability of the calibration group and the spread of ability within the group.

The CTT score for a given person depends on a particular set of items. The IRT score for a given person does not depend on a particular set of items.

The score (ability) of a person is given by the location of that person on the calibrated scale. The accuracy of the person's ability may be improved by selecting items whose difficulty is close to that ability.

81

## Uses of IRT: Item banks and adaptive testing

Items calibrated by means of IRT can be used to construct item banks available for the selection of sets of items for groups of persons with defined ranges of ability. Adaptive or tailored testing can then be employed.

In this process, a person is initially given a small set of items covering a range of difficulty. The person's responses are then used to select items closer in difficulty to the person's ability, until an accurate measure of that ability is determined. There is no need to give easy items to persons of high ability (and vice versa) so that the total number of items needed is less than for a conventional test.

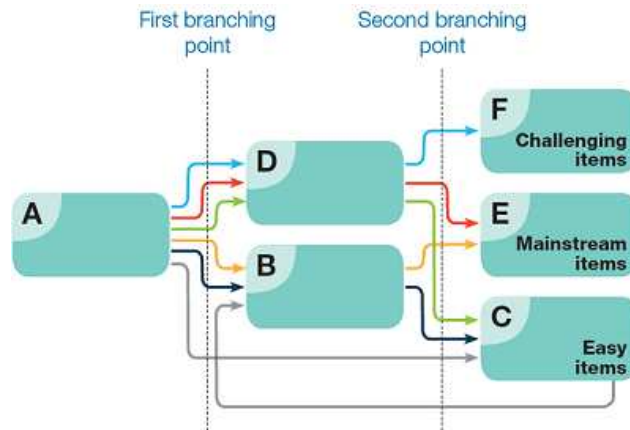
Students at the same ability ( $\theta$ ) level can be given different items so there is less chance of cheating or memorising the answers, unless the person memorises the correct responses to all items in the item bank.

Adaptive testing is useful in cases where a licence or award is at stake, such as a quiz for potential immigrants, or for a driver's licence.

NAPLAN is proposing to use adaptive testing.

82

## NAPLAN proposed adaptive testing



[http://www.nap.edu.au/verve/\\_resources/Moving\\_NAPLAN\\_online\\_-\\_Understanding\\_the\\_tailored\\_test\\_design.pdf](http://www.nap.edu.au/verve/_resources/Moving_NAPLAN_online_-_Understanding_the_tailored_test_design.pdf)

## Uses of IRT: Item and test equating

If we wish to equate tests across two age groups, we need to ensure that we have a set of items common to both groups. For the younger age group, these would represent harder items. For the older age group, these would represent easier items.

We need to check that the link items are working satisfactorily across both groups. This does not mean that the item difficulties should be equal. It does mean that the ICC for a link item should have the same shape/profile for each group, albeit displaced along the person continuum.

The process of differential item functioning (DIF) is used to check if the items apply satisfactorily across different types of groups: gender, age, SES, culture, language, etc.

`difmh` is a Stata command that performs a chi-squared test for uniform differential item functioning in the context of IRT analysis. The Mantel-Haenszel test for DIF is used to determine whether the responses to an item are independent of the group to which an individual belongs.

## Uses of IRT: Examine cases with missing values

We may wish to investigate persons who have omitted or declined to answer certain items. Since the IRT procedures use MLE, we can obtain estimates of the ability of persons on items that they did not answer.

Consider a test where some respondents did not answer the questions towards the end of a test. IRT does not penalise a slow worker who simply did not manage to complete all items. That person's ability is assessed on the basis of items that were answered. The researcher can also investigate characteristics of missed items.

Some items in a scale may be rather sensitive. For example, in a scale for anti-social behaviour, persons may choose to avoid items asking about anti-social activities in which they might have engaged. With IRT we can look at the ability and other characteristics of persons who omitted such items.

85

## Uses of IRT: Examine alternatives

IRT can be used to examine the mean ability associated with the different responses to an item.

In the case of a multi-choice performance item, this lets us see the ability associated with each of the incorrect alternatives. We can assess which alternatives are more plausible, and also see which are not working as expected.

Multi-choice performance items designed to reward responses that are partially correct, would be scored with the Partial Credit Model (PCM). With IRT we can look at the ability associated with members of the response set. Some responses may be combined or deleted.

For attitude scales, we can examine how well the categories of the response set are working. We may be able to reduce the number of categories.

86

## **Uses of IRT: Large-scale testing programs**

Australia: National Assessment Program – Literacy and Numeracy (NAPLAN)

US: National Assessment of Educational Progress (NAEP), College Testing Board (CTB) and many other national and state testing programs.

International: Programme for International Student Assessment (PISA),

International: Studies from the International Association for Educational Achievement (IEA).

87

## **Uses of IRT: Comparison across groups (1)**

Most of the large-scale assessment programs operate at several levels.

NAPLAN is conducted at Years 3, 5, 7 and 9. These programs use IRT procedures to link results to a common scale across the levels. Minimum standards and common scales for NAPLAN results across all year levels.

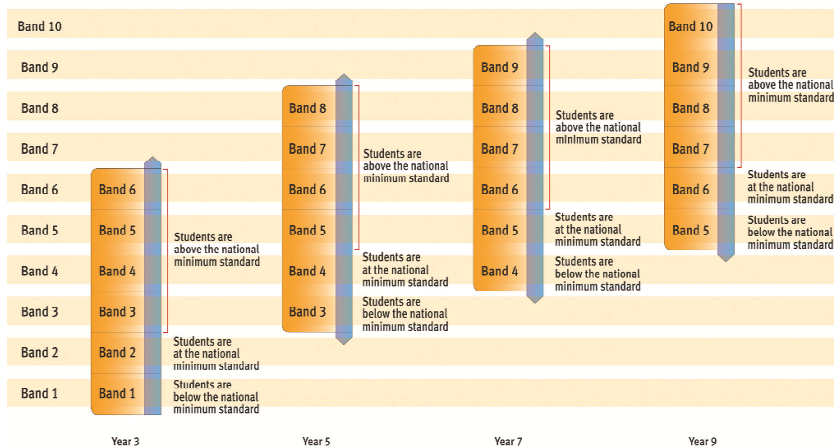
Linking items across levels requires the estimation of some items in common. For example, there may be a 20-item test for Year 3, and a 20-item test for Year 5, with 10 of the items common across year levels, and so on.

One method of linking is to run the IRT estimation using all items and all students across the four testing levels.

88

## Uses of IRT: Comparison across groups (2)

National Assessment Program—Literacy and Numeracy National Assessment Scale



89

## Uses of IRT: Comparison across time

Linking items across time requires that a set of link items from the two occasions should be administered to a single group of students.

For public testing programs, it is not feasible to do this in the country where the tests are administered. So for an Australian program it would be necessary to administer the link items from the two occasions in another country with similar culture and achievement standards, such as New Zealand or Singapore.

90

## Summary

- IRT is now available in Stata 14 as a useful complement to its range of CTT procedures.
- In terms of valid measurement, the assumptions underlying IRT are more justifiable than those for CTT.
- The IRT procedures and associated graphs are very easy to use.
- Most smaller research projects should consider using IRT, especially for criterion variables.
- Most larger programs, where the stakes are higher, should adopt IRT procedures.

91

## Data and do-files

Two do-files are available if users wish to replicate the analyses shown in this presentation. Please contact [sales@surveydesign.com.au](mailto:sales@surveydesign.com.au), or contact the author at [mrosier@tpg.com.au](mailto:mrosier@tpg.com.au).

- Do-file `masc01.do` is used to show the `1p1` and `2p1` procedures.
- Do-file `charity01.do` is used to show the `rsm` and `grm` procedures.

The datasets cited are available from the web from within Stata.

The Stata 14 IRT manual (pdf) describes further `irt` procedures, and provides formulae and other statistical details.

92