Item Response Theory in Stata 14

Rebecca Pope

rpope@stata.com

Health Econometrician
StataCorp LP

2015 Stata Conference

Thinking About Latent Traits

Defining latent traits

A *latent trait* can be any characteristic that is not directly observed.

Conventionally, this has meant ability.

It can also include feelings, such as satisfaction, mental status, such as anxiety, and health states.

Contrast this with *manifest* variables that can be observed and may be used to measure the latent trait.

For example, we might measure the latent trait "reading proficiency" by using responses on a standardized test.

Item Response Theory

Basic concepts

We measure the latent trait using an *instrument*, which is a collection of *items*.

- Each item has a *difficulty* parameter and a *discrimination* parameter.
- Some models permit items to have a guessing parameter.

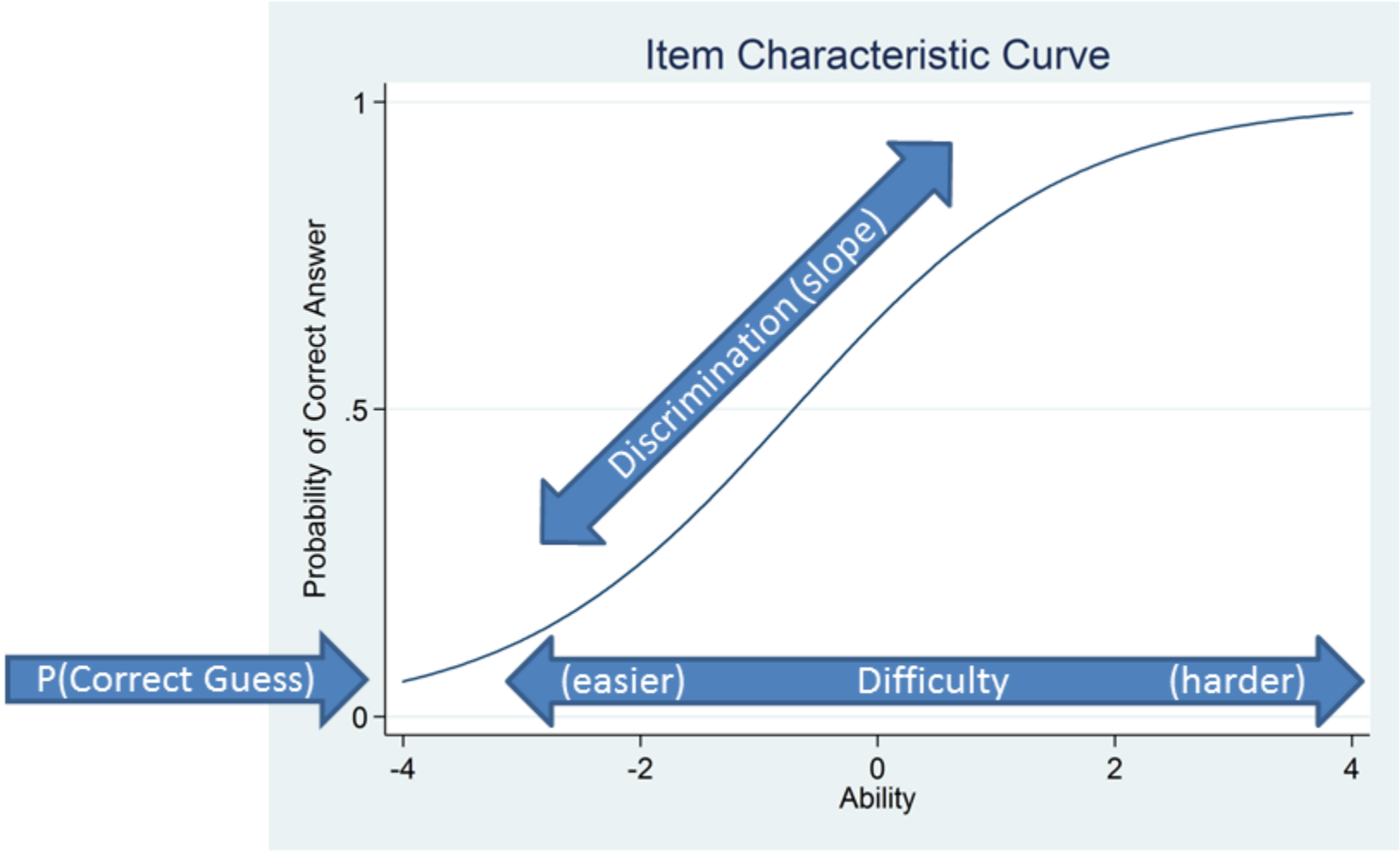
We investigate the probability of positive response to each item separately.

- From this, we can obtain a measure of the level of the latent trait that would be required to have a 50/50 chance of responding correctly.
- We can also aggregate item-level information to instrument-level information.

Measurement

- Binary outcomes
  - Binary responses
    Multiple choice scored "correct" or "incorrect"
- Ordinal outcomes
  - Multiple choice scored "correct", "partially-correct", or "incorrect"
    Likert-type responses
- Nominal outcomes
  - Multiple choice with no correct or incorrect answer

Visualizing IRT



Assumptions

1. The probability of successfully answering a given item i is some known function that we have correctly specified. We usually use some variant of the cumulative logistic distribution.
2. The latent trait is continuous and normally distributed.
3. Conditional on the latent trait, the responses to any two items are independent of each other.
4. The responses of different individuals to the same item are independent of each other.

Estimating IRT Models in Stata

Model choices

- Binary outcome models
  - One-parameter logistic (1PL) model
  - Two-parameter logistic (2PL) model
  - Three-parameter logistic (3PL) model
- Categorical outcome models
  - Partial credit model (PCM)
  - Rating scale model (RSM)
  - Generalized partial credit model (GPCM)
  - Graded response model (GRM)
  - Nominal response model (NRM)
- And *hybrid models* for combinations of any of the above

The basic command structure

To fit a single model you type, for example

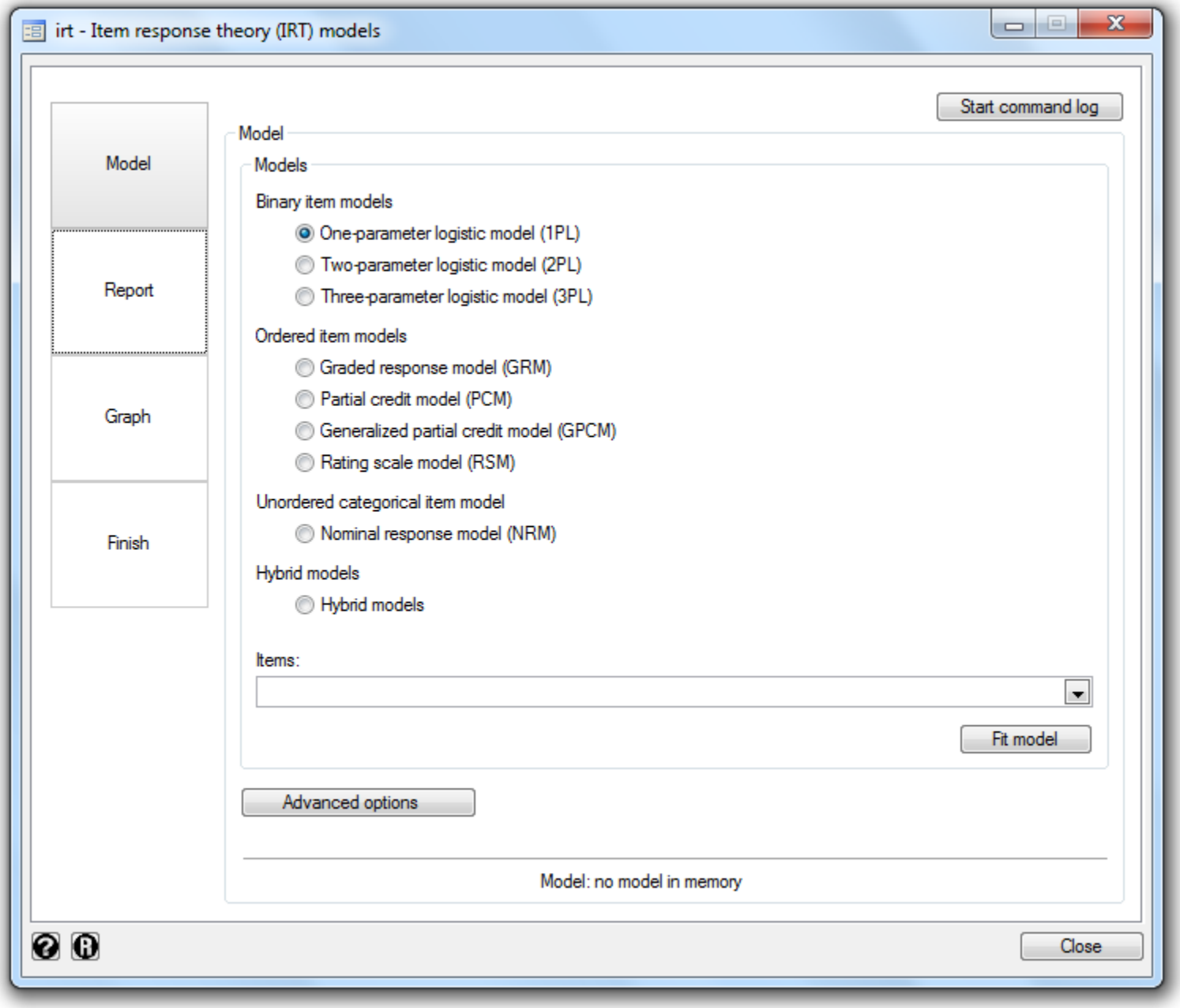
```
irt 1pl varlist
```

For the subcommand, you can select from `1pl`, `2pl`, `3pl`, `grm`, `nrm`, `pcm`, `gpcm`, and `rsm`.

To fit a hybrid model you type, for example

```
irt hybrid (2pl varlist1) (grm varlist2)
```

The IRT control panel



Common options

options	Description
Model	
listwise	drop observations with any missing items
SE/Robust	
vce(*vcetype***)**	*vcetype* may be **oim**, **robust**, **cluster** *clustvar*, **bootstrap**, or **jackknife**
Reporting	
level(*#***)**	set confidence level; default is **level(95)**
notable	suppress coefficient table
noheader	suppress output header
display_options	control columns and column formats
Integration	
intmethod(*intmethod***)**	integration method
intpoints(*#***)**	set the number of integration points; default is **intpoints(7)**
Maximization	
maximize_options	control the maximization process; seldom used
startvalues(*svmethod***)**	method for obtaining starting values
noestimate	do not fit the model; show starting values instead
dnumerical	use numerical derivative techniques
coeflegend	display legend instead of statistics

Find out more in *Options* of **[IRT] irt 1pl** or the manual entry for any of the other `irt` subcommands.

Postestimation

- Reorganize estimation output
  - Group results by item or parameter
  - Sort by discrimination, difficulty, or guessing parameter
- Graphics
  - Plot item characteristic curves for binary items
  - Plot category characteristic curves for categorical items
  - Plot item information functions
  - Plot the test characteristic curve
  - Plot the test information function
- Predictions
  - Latent trait
  - Conditional and marginal probability of a given response

IRT in Action

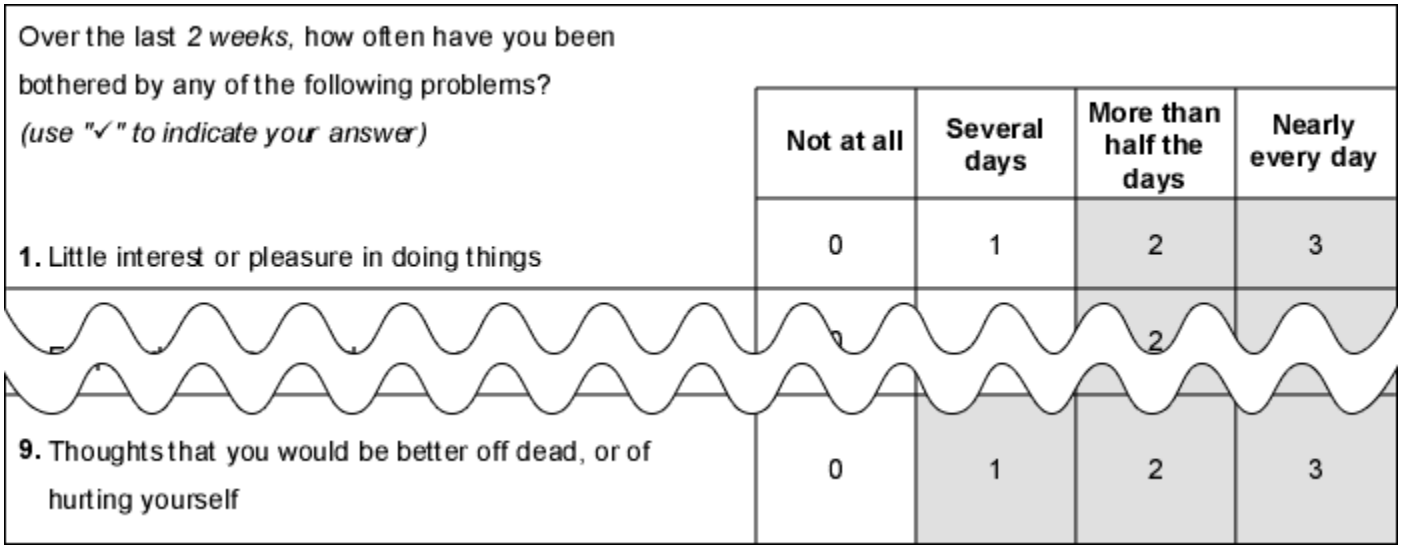
Background

Depression is a risk factor for poor outcomes for individuals with health conditions that are commonly treated in the emergency department (ED). But, administering classic diagnostic questionnaires or conducting formal clinical evaluations in the ED is not feasible due to time constraints and the lack of trained mental health professionals.

Suppose we wish to develop a new method of screening patients in the emergency department for depression. This instrument will help us determine who should be referred to mental health services for follow-up.

Background, continued

We create a series of binary-response items based on the PHQ-9, a tool designed for use in primary care.



Over the last *2 weeks*, how often have you been bothered by any of the following problems? *(use "✓" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
9. Thoughts that you would be better off dead, or of hurting yourself	0	1	2	3

In the past week, did you have little interest in doing things on 4 or more days?

⋮

In the past week, did you think you would be better off dead or think of hurting yourself?

A look at our data

```
. describe

Contains data from depr.dta
  obs:         1,000                          Depression screening data
 vars:            11                          23 Jul 2015 12:27
 size:        11,000                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
d1              byte    %9.0g      yn         No interest in things
d2              byte    %9.0g      yn         Feels hopeless
d3              byte    %9.0g      yn         Trouble sleeping
d4              byte    %9.0g      yn         Feels tired
d5              byte    %9.0g      yn         Unusual eating
d6              byte    %9.0g      yn         Feels like a failure
d7              byte    %9.0g      yn         Trouble concentrating
d8              byte    %9.0g      yn         Moves slow or is fidgety
d9              byte    %9.0g      yn         Suicidal ideation
female          byte    %9.0g      female     Patient sex
-------------------------------------------------------------------------------
Sorted by:

. notes

_dta:
  1.  Simulated for Stata Conference 2015.
  2.  Items based on PHQ-9, (c) 1999 by Pfizer Inc.
  3.  PHQ-9 available at
      http://www.integration.samhsa.gov/images/res/PHQ%20-%20Questions.pdf
```

A 1PL model

```
Fitting fixed-effects model:

Iteration 0:   log likelihood = -4569.9596
Iteration 1:   log likelihood = -4544.2515
Iteration 2:   log likelihood =  -4543.815
Iteration 3:   log likelihood = -4543.8143
Iteration 4:   log likelihood = -4543.8143

Fitting full model:

Iteration 0:   log likelihood = -4167.2584
Iteration 1:   log likelihood = -4077.8709
Iteration 2:   log likelihood = -4077.3488
Iteration 3:   log likelihood = -4077.3487

One-parameter logistic model                    Number of obs     =      1,000
Log likelihood = -4077.3487
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Discrim |   1.541299   .0598738    25.74   0.000     1.423949     1.65865
-------------+----------------------------------------------------------------
d1           |
        Diff |   .8577526   .0673467    12.74   0.000     .7257555    .9897496
-------------+----------------------------------------------------------------
d2           |
        Diff |   1.572861   .0860161    18.29   0.000     1.404272    1.741449
-------------+----------------------------------------------------------------
d3           |
        Diff |  -1.297328   .0769829   -16.85   0.000    -1.448212   -1.146445
-------------+----------------------------------------------------------------
d4           |
        Diff |  -.656133    .063346   -10.36   0.000    -.7802889   -.5319772
```

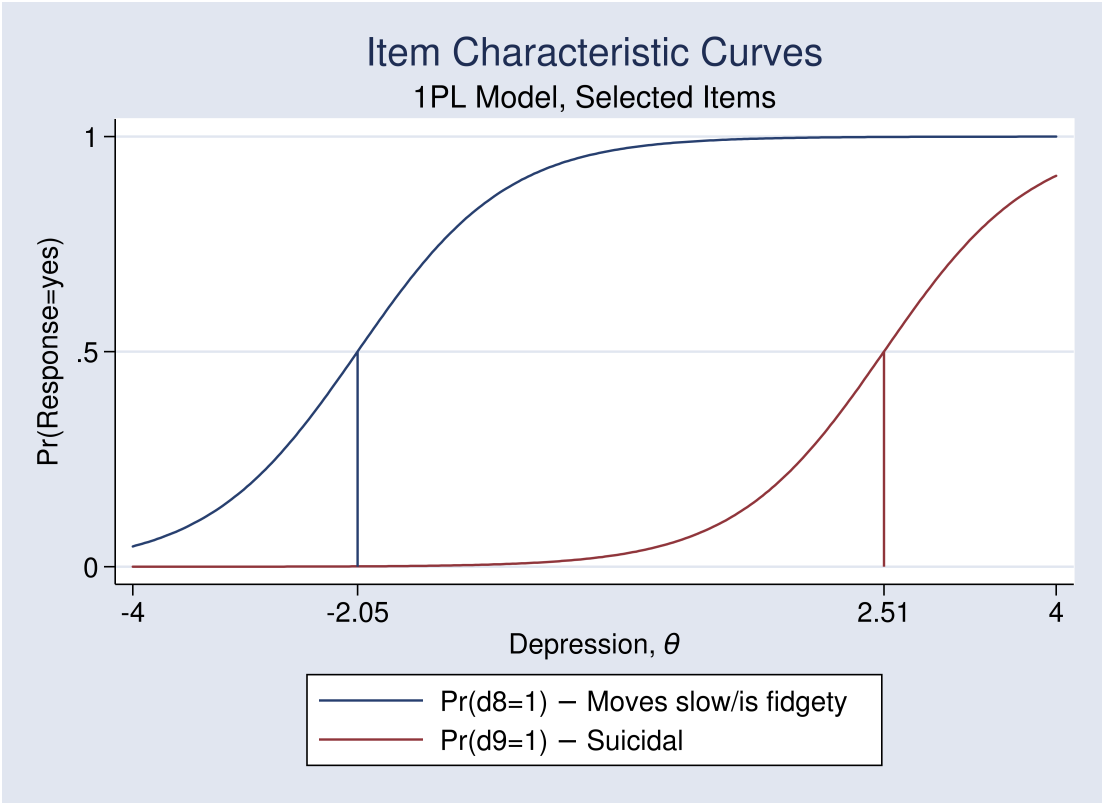
Reorganizing the results

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

One-parameter logistic model                    Number of obs     =      1,000
Log likelihood = -4077.3487
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Discrim |   1.541299   .0598738    25.74   0.000     1.423949     1.65865
-------------+----------------------------------------------------------------
Diff         |
          d8 |  -2.052764   .1045475   -19.63   0.000    -2.257674   -1.847855
          d3 |  -1.297328   .0769829   -16.85   0.000    -1.448212   -1.146445
          d4 |   -.656133    .063346   -10.36   0.000    -.7802889   -.5319772
          d7 |  -.2776815   .0594574    -4.67   0.000    -.3942159   -.1611471
          d5 |   .1212162   .0589232     2.06   0.040     .0057289    .2367035
          d1 |   .8577526   .0673467    12.74   0.000     .7257555    .9897496
          d6 |   1.129998   .0732831    15.42   0.000     .9863661    1.273631
          d2 |   1.572861   .0860161    18.29   0.000     1.404272    1.741449
          d9 |   2.507561   .1278655    19.61   0.000     2.256949    2.758173
------------------------------------------------------------------------------
```

Item characteristic curves (1PL)

```
irtgraph icc d8 d9, blocation
```



`blocation` draws lines from the estimated difficulty up to the curve.

A 2PL model

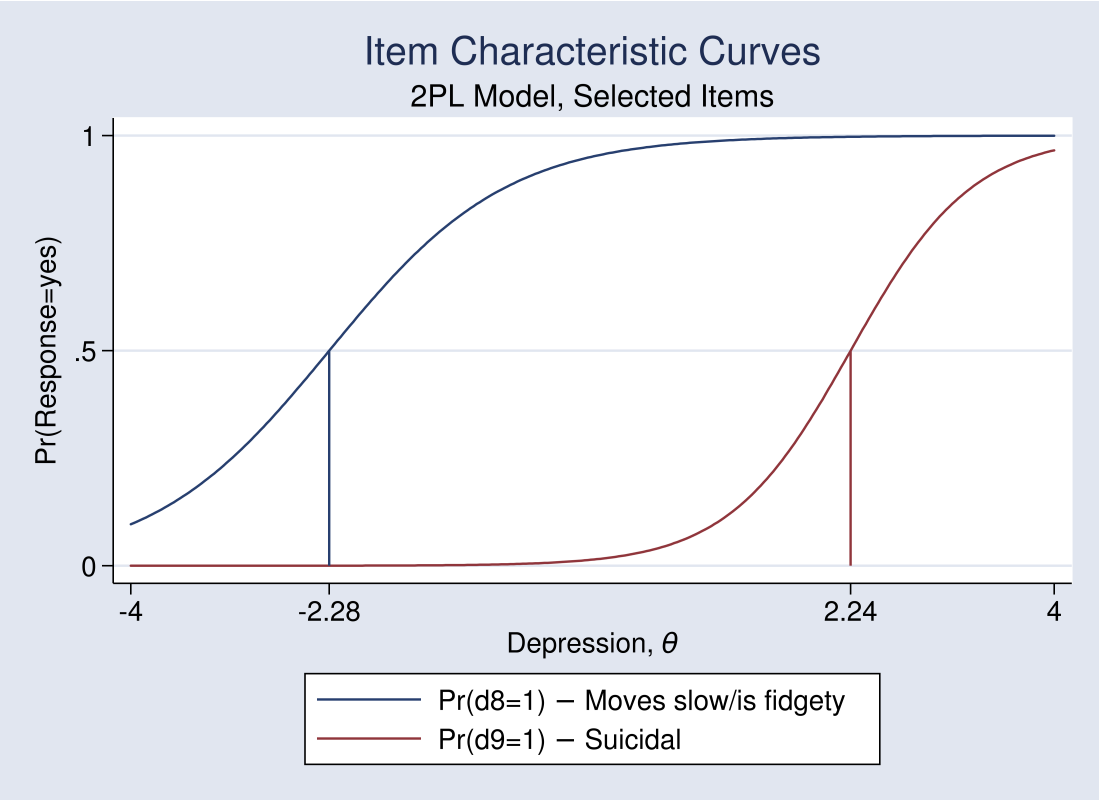
```
. quietly irt 2pl d*

. estat report, byparm sort(b)

Two-parameter logistic model                    Number of obs     =      1,000
Log likelihood = -4058.2333
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Discrim      |
          d8 |    1.30211   .1981992     6.57   0.000     .9136467    1.690573
          d3 |   .9739991   .1266772     7.69   0.000     .7257163    1.222282
          d4 |   1.218435   .1292939     9.42   0.000     .9650236    1.471846
          d7 |   1.864502   .1880718     9.91   0.000     1.495888    2.233116
          d5 |   1.550059   .1523476    10.17   0.000     1.251464    1.848655
          d1 |   2.070205   .2215149     9.35   0.000     1.636044    2.504366
          d6 |   1.634441    .173512     9.42   0.000     1.294364    1.974518
          d2 |   1.844864   .2135546     8.64   0.000     1.426305    2.263423
          d9 |   1.893582   .2957141     6.40   0.000     1.313993    2.473171
-------------+----------------------------------------------------------------
Diff         |
          d8 |  -2.280343    .243705    -9.36   0.000    -2.757996    -1.80269
          d3 |  -1.733487   .1904042    -9.10   0.000    -2.106672   -1.360301
          d4 |  -.7425246   .0865185    -8.58   0.000    -.9120977   -.5729515
          d7 |  -.2439826   .0553367    -4.41   0.000    -.3524405   -.1355247
          d5 |   .1317081   .0590492     2.23   0.026     .0159738    .2474425
          d1 |   .7559761   .0643166    11.75   0.000     .6299178    .8820344
          d6 |   1.099713   .0887349    12.39   0.000     .9257959    1.273631
          d2 |   1.435696   .1045111    13.74   0.000     1.230858    1.640534
          d9 |   2.236693   .1952693    11.45   0.000     1.853972    2.619414
------------------------------------------------------------------------------
```

Item characteristic curves (2PL)

```
irtgraph icc d8 d9, blocation
```



With a 2PL model, the slope of the ICCs for different items can vary.

A moment for methods

For 1PL and 2PL models, let i index the item, j index the person, and $y_{ij} = 1$ indicate a positive or "correct" response.

Stata estimates the probability of providing a positive response as

$$\Pr(y_{ij} = 1|\alpha_i, \beta_i, \theta_j) = \frac{e^{(\alpha_i\theta_j+\beta_i)}}{1 + e^{(\alpha_i\theta_j+\beta_i)}}$$

However, the α_i are constrained to be equal for all items when estimating the 1PL model. We report the discrimination, a_i, and difficulty, b_i using the IRT parameterization; the transformation is $a_i = \alpha_i$ and $b_i = -\frac{\beta_i}{\alpha_i}$.

Estimation is by maximum-likelihood and uses Gauss–Hermite quadrature to approximate the log likelihood by default.

Letting $p_{ij} = \Pr(y_{ij} = 1|\alpha_i, \beta_i, \theta_j)$ and $q_{ij} = 1 - p_{ij}$, the conditional density for person j is

$$f(\mathbf{y}_j|\mathbf{B}, \theta_j) = \prod_{i=1}^{I} p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}}$$

where $\mathbf{y}_j = (y_{1j}, \ldots, y_{Ij})$, $\mathbf{B} = (\alpha, \beta_1, \ldots, \beta_I)$ in the 1PL model or $\mathbf{B} = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_I)$ in the 2PL model, and I is the number of items.

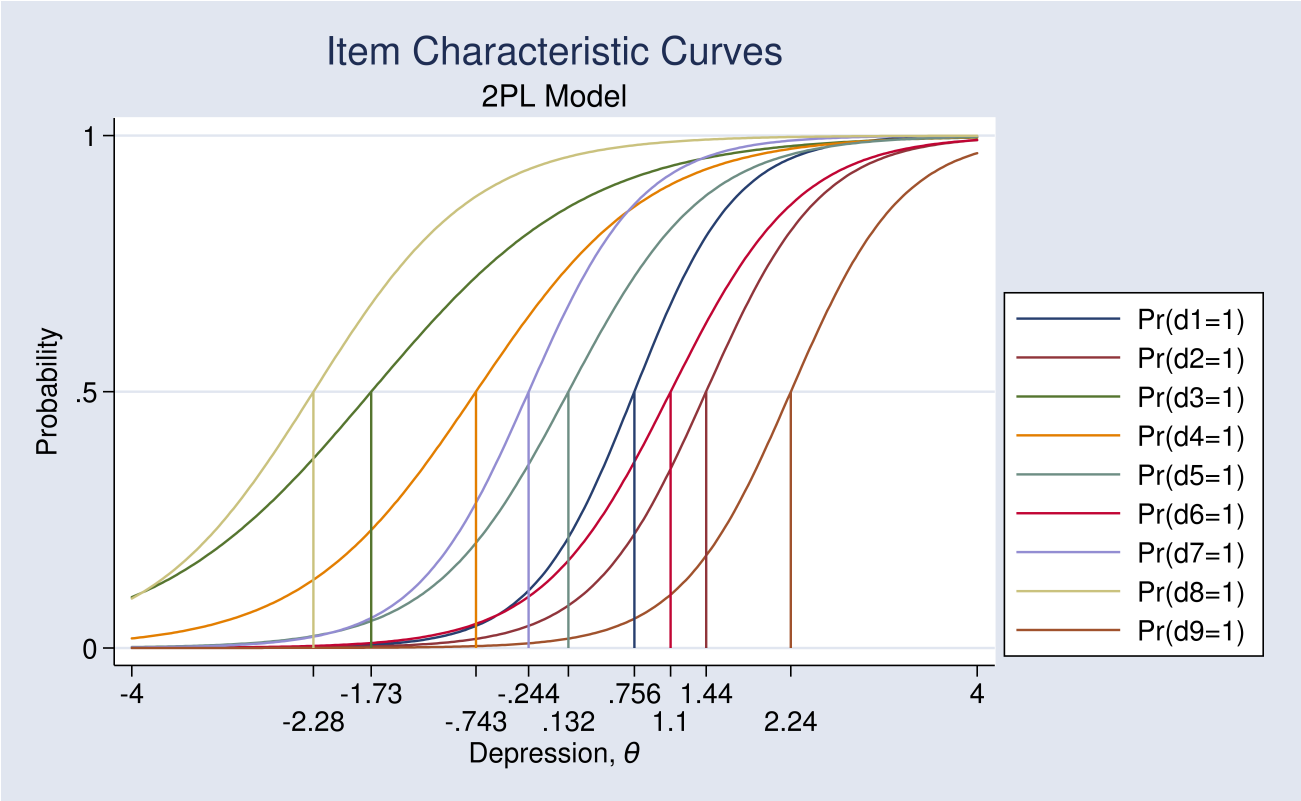
The log likelihood is the sum of the N individual log likelihoods, $\log \mathcal{L}_j(\mathbf{B})$ where

$$\mathcal{L}_j(\mathbf{B}) = \int_{-\infty}^{\infty} f(\mathbf{y}_j|\mathbf{B}, \theta_j)\phi(\theta_j)d\theta_j$$

and $\phi(\cdot)$ is the standard normal density function. For details, see *Methods and formulas* in **[IRT] irt hybrid**.

IRT in Action

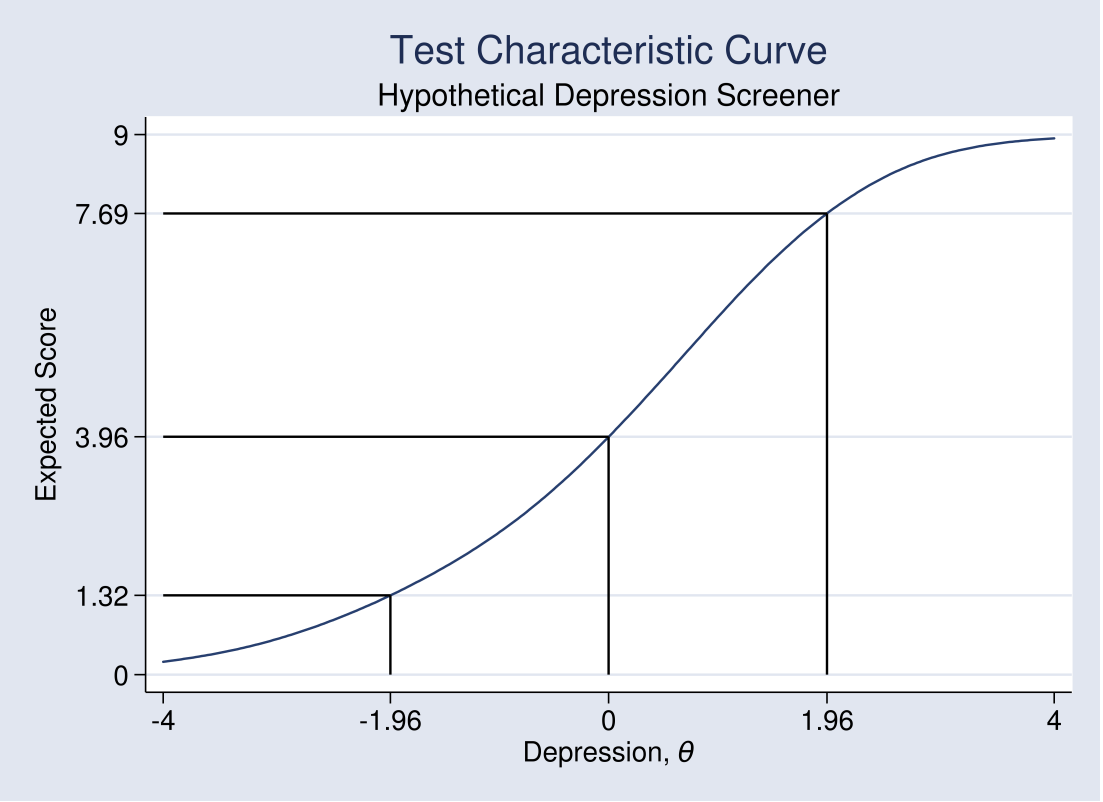
Aggregating item information



We need to aggregate the scores from each individual item into a total score that we base our referral decision on.

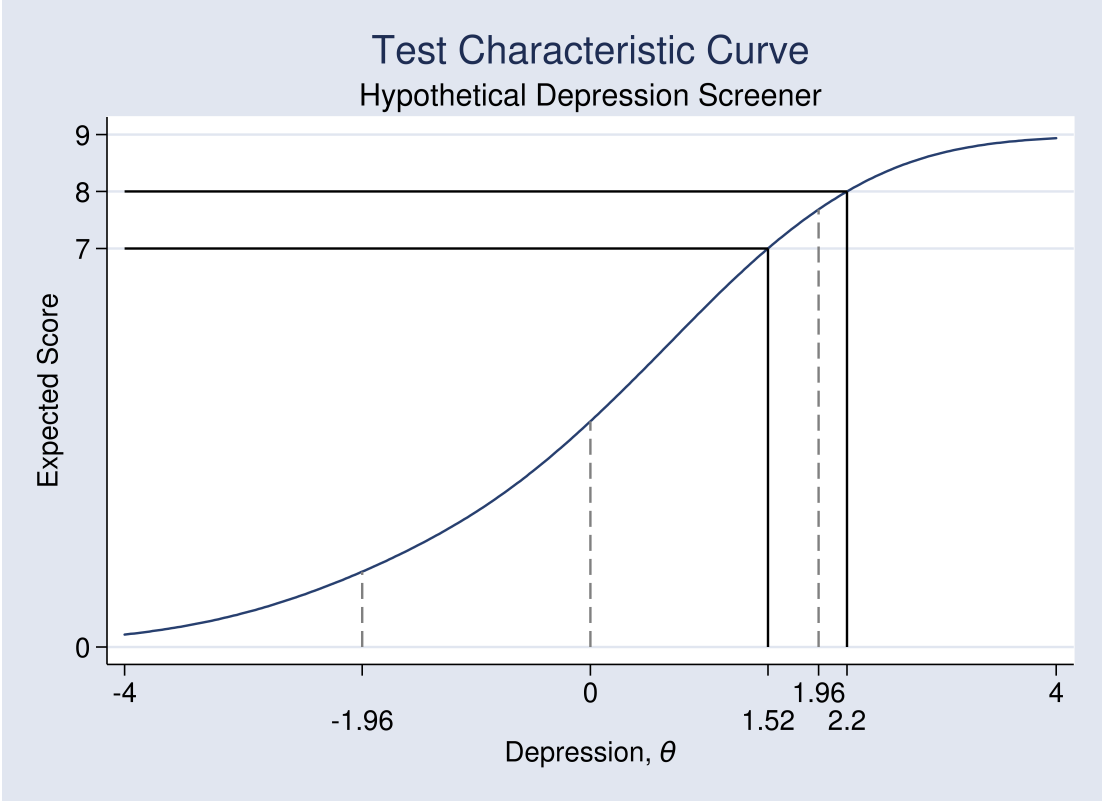
Expected scores and the test characteristic curve

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



Obtaining values of the latent trait at an expected score

```
irtgraph tcc, scorelines(7 8) thetalines(-1.96 0 1.96)
```



We still have to interpret this as the expected score at a particular value of θ.

Checking model specification

The invariance of IRT allows us to make predictions about the expected score for individuals outside of the sample the instrument was developed in.

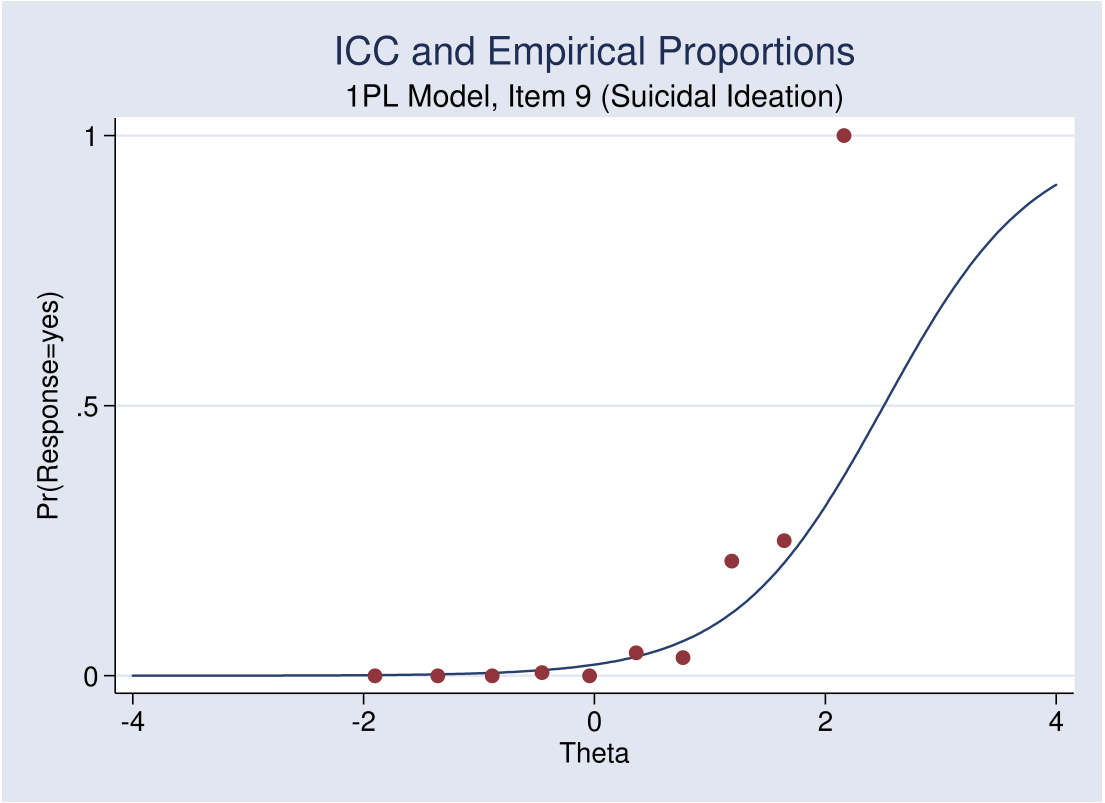
The invariance property only holds for correctly-specified models.

To check model specification, we can

- visually inspect the distribution of the observed responses against the predicted latent trait and ICC.
- test more restrictive models against a model that allows more parameters to vary by item.

Visual inspection

```
quietly irt 1pl d*
predict Theta, latent
collapse d*, by(Theta)
irtgraph icc d9, addplot(scatter d9 Theta)
```



If we were checking fit for a 2PL or 3PL model, we would need to create bins for the latent scores before using `collapse`.

Test for model fit

To conduct the test, we need to make sure we use `estimates store`

```
quietly irt 1pl d*
estimates store onepl
quietly irt 2pl d*
estimates store twopl
```

We can then use `lrtest` to test the null hypothesis that the more restrictive model is preferred.

```
. lrtest onepl twopl

Likelihood-ratio test                                 LR chi2(8)  =     38.23
(Assumption: onepl nested in twopl)                   Prob > chi2 =    0.0000
```

Differential Item Functioning

What is DIF?

When differential item functioning (DIF) occurs, members of different groups have different probability profiles for an item.

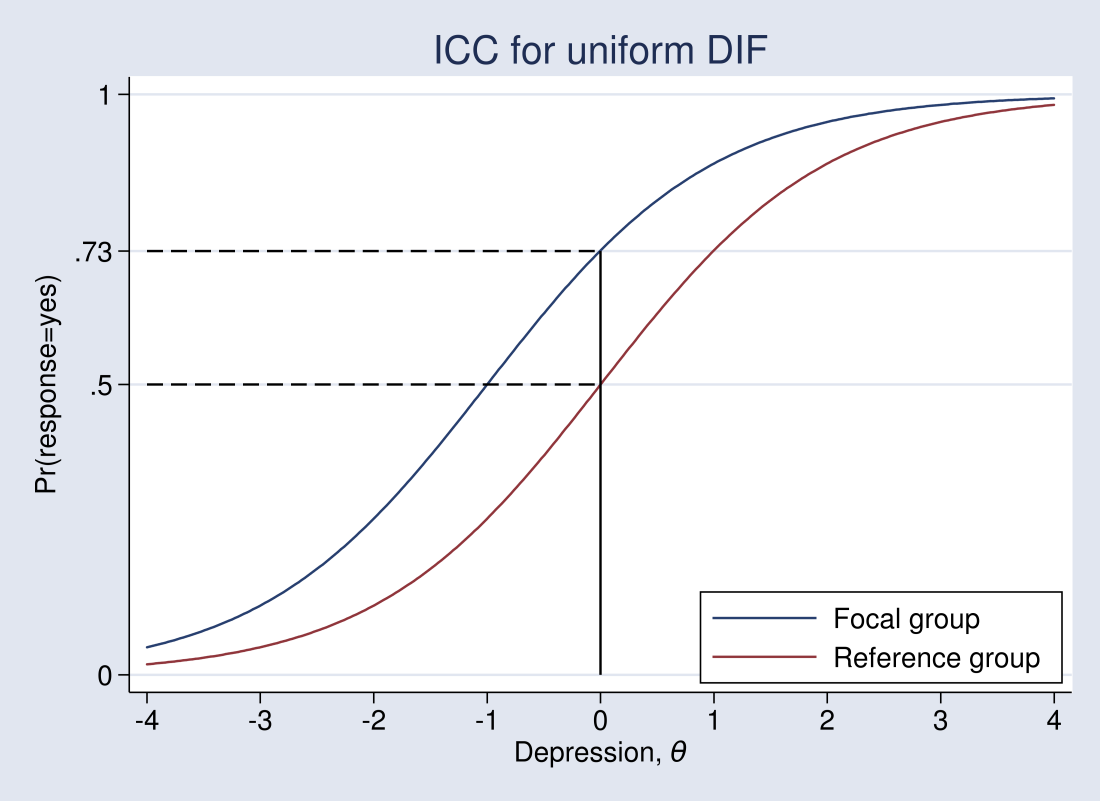
The *focal group* is the group with the characteristc we think may influence performance and is compared to the *reference group*.

There are two types of DIF — uniform and nonuniform.

For uniform DIF, the focal group always underperforms (or outperforms) the reference group.

For nonuniform DIF, the focal group outperforms the reference group over some range of the latent trait and underperforms over a different range.

Visualizing uniform DIF



Testing for uniform DIF

We conduct the Mantel–Haenszel test for uniform DIF using the `difmh` command.

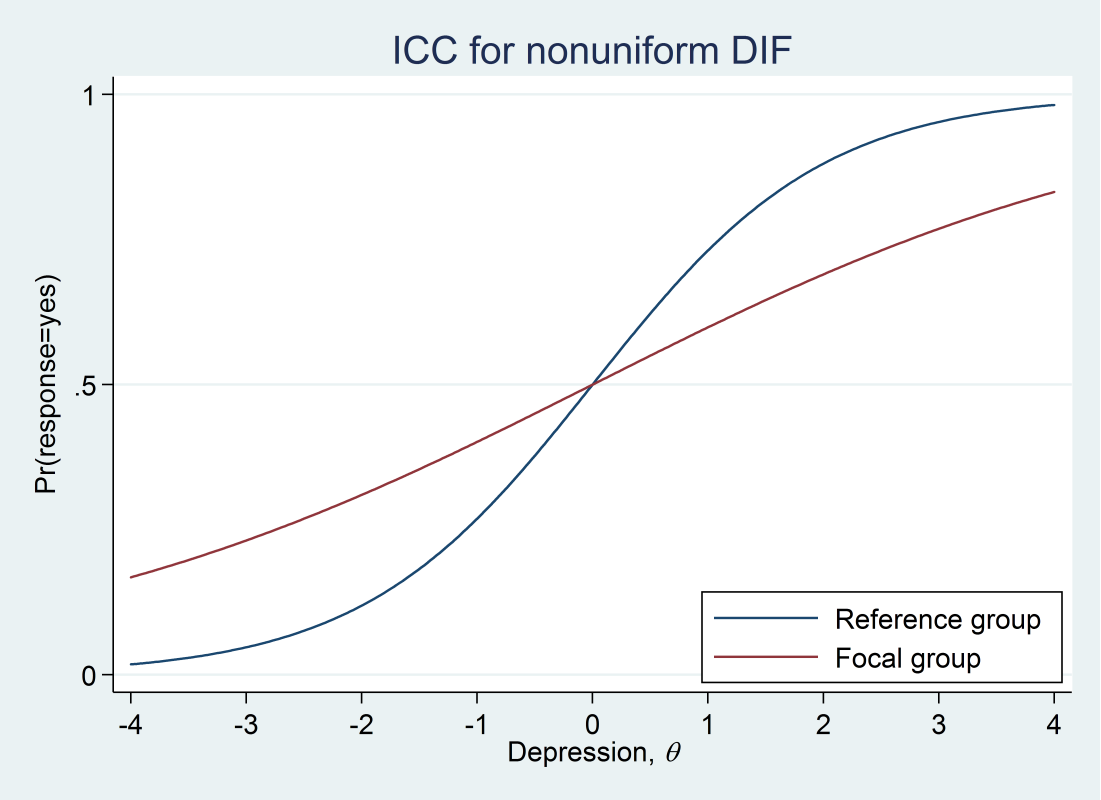
```
. difmh d*, gr(female)

Mantel-Haenszel DIF Analysis

 Item |      Chi2        Prob. | Odds Ratio      [95% Conf. Interval]
------+----------------------+-----------------------------------
   d1 |    0.1530       0.6957 |     0.9038       0.6040       1.3523
   d2 |    6.5677       0.0104 |     0.4715       0.2703       0.8225
   d3 |    5.2762       0.0216 |     0.6238       0.4234       0.9189
   d4 |   21.0087       0.0000 |     2.3235       1.6169       3.3389
   d5 |    0.0071       0.9328 |     0.9707       0.6908       1.3642
   d6 |    1.2743       0.2590 |     0.7696       0.5067       1.1691
   d7 |    0.0062       0.9371 |     0.9973       0.6871       1.4475
   d8 |    0.3241       0.5692 |     1.2244       0.7009       2.1387
   d9 |    0.6836       0.4083 |     1.4453       0.7073       2.9531
-------------------------------------------------------------------
```

For the same level of depression, women have higher odds of a postive response to item 4 (feeling tired) and lower odds of a positive response to item 2 (feeling hopeless) and item 3 (trouble sleeping).

Visualizing nonuniform DIF



For more information about DIF, see **[IRT] difmh**.

Extras

Fitting categorical IRT models

Instead of forcing respondants to choose between "Yes" and "No" as responses, we may have also let them pick "Don't know". At this point, we would have a nominal response model (NRM).

We can estimate an NRM in Stata by typing

```
. irt nrm d*
```

However, because the NRM allows the discrimination and difficulty parameters to vary by item, just like the 2PL, we may want to fit the model `quietly` and then use `estat report` to group responses by difficulty.

```
. quietly irt nrm d*
. estat report, byparm sort(b)
```

Postestimation graphs for categorical IRT models

After NRM and other categorical models, we obtain category characteristic curves rather than item characteristic curves. However, we still use `irtgraph icc`; Stata makes the adjustment for us.

Because there are multiple responses for each question, we should look at each item separately. So, we might have a do-file that looks like this.

```
irtgraph icc d1, name(item1, replace)
irtgraph icc d2, name(item2, replace)
/* ... */
irtgraph icc d9, name(item9, replace)
```

Postestimation predictions for categorical IRT models

When we switch to categorical models, we the nature of our predictions changes. Rather than calculating an overall probability of a positive response, we have to specify the response we are interested in.

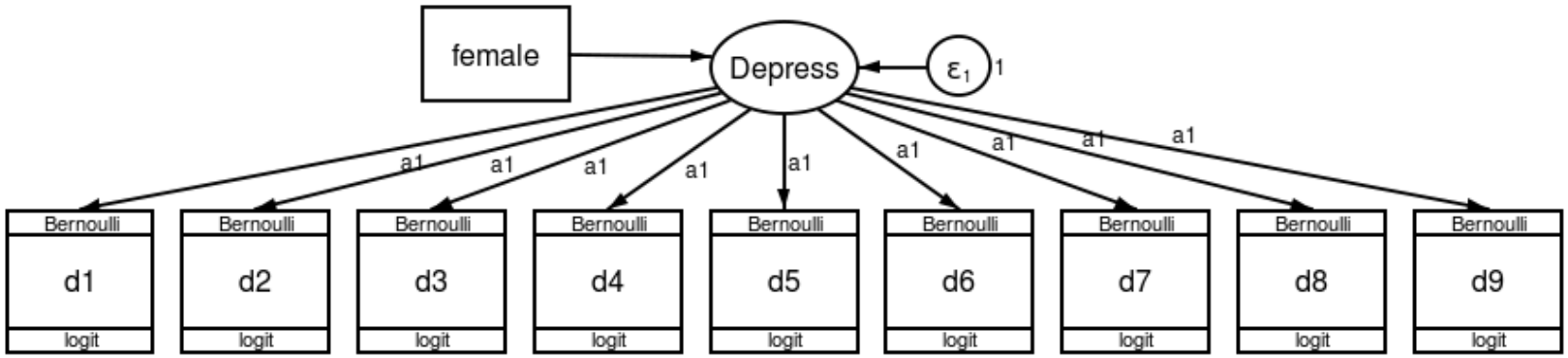
In our context, we are probably still interested in predicting the probability of "yes" responses. So, for example, if we want to calculate the predicted probability of responding "yes" to item 9, conditional on the value of the latent trait, we can submit

```
. predict pr d9_yes, pr outcome(d9 1)
```

But we could also calculate marginal, or population-averaged, probabilities of a positive response by adding option `marginal` to our `predict` command.

Fitting an IRT model with covariates

To fit an IRT model with covariates, we can use `gsem`. The path diagram in the SEM Builder for a 1PL model looks like this:

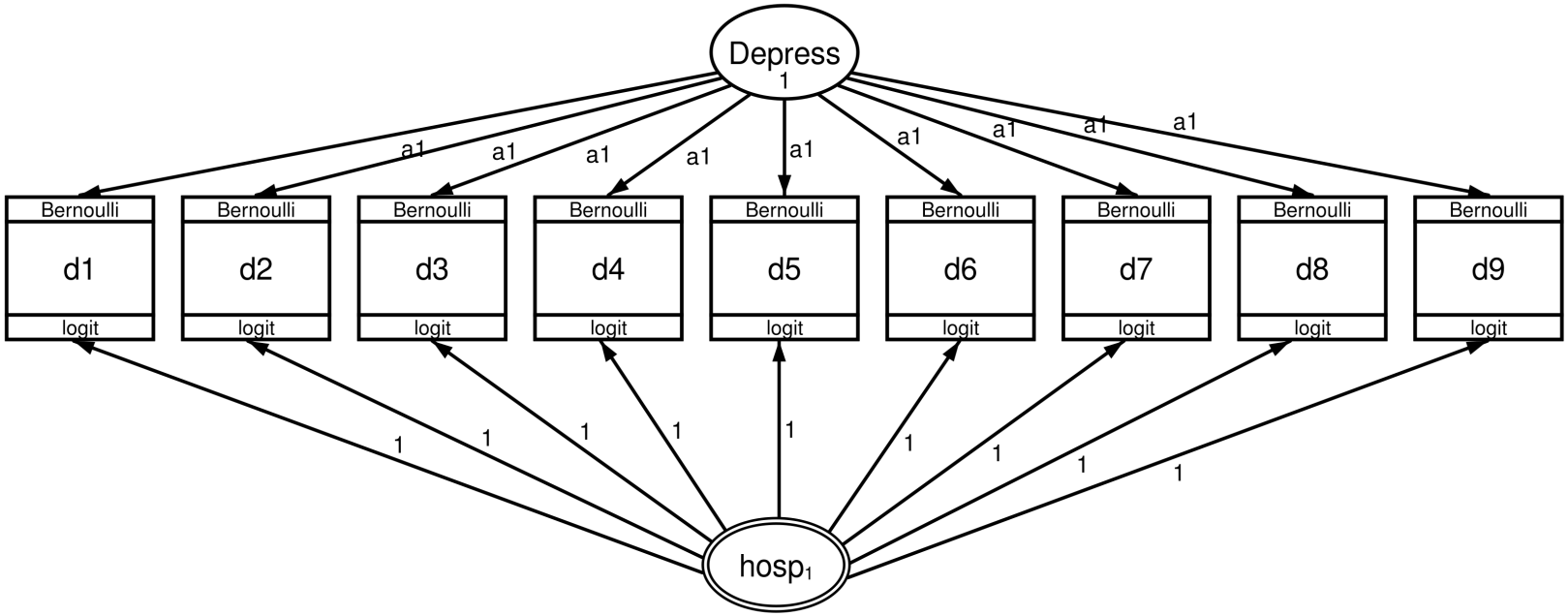


or, to submit the command from the do-file editor, we would type

```
gsem (Theta@a -> d1 d2 d3 d4 d5 d6 d7 d8 d9, logit)          ///
   (female -> Theta), variance(e.Theta@1) latent(Theta)
```

Fitting a multilevel IRT model: The Path diagram

The path diagram in the SEM Builder for a multilevel 1PL looks like this:



Fitting a multilevel IRT model: The code

To submit the command from the do-file editor, we would type

```
gsem (Theta@a -> d* H[hospital]@1, logit),    ///
   variance(Theta@1) latents(Theta H) intpoints(3)
```

Tips for successfully estimating multilevel IRT models:

Gradually increase the number of integration points. Specifying more integration points increases accuacy at the price of computational time.

Consider fitting a one-level model first and using the stored estimates as starting values for the multilevel model.

Full code for graphs

Replicate ICC for selected items, 1PL

```
#delimit ;
irtgraph icc d8 d9, blocation subtitle("1PL Model, Selected Items")
  ytitle("Pr(Response=yes)") xtitle("Depression, {it:{&theta}}")
  order(1 "Pr(d8=1) {&minus} Moves slow/is fidgety"
  2 "Pr(d9=1) {&minus} Suicidal"))
;
#delimit cr
```

Replicate ICC for all items, 2PL

```
#delimit ;
irtgraph icc, blocation subtitle("2PL Model")
  ytitle("Pr(Response=yes)") ysize(4)
  xtitle("Depression, {it:{&theta}}") xlabel(, alt) xsize(6.5)
  legend(pos(4) col(1) ring(1))
;
#delimit cr
```

Replicate TCC with specified values of the latent trait

```
#delimit ;
irtgraph tcc, thetalines(-1.96 0 1.96)
   subtitle("Hypothetical Depression Screener")
   xtitle("Depression, {it:{&theta}}")
;
#delimit cr
```

Replicate TCC with specified values of the expected score

```
#delimit ;
irtgraph tcc, scorelines(7 8)
   thetalines(-1.96 0 1.96, lcolor(gray) lpattern(dash) noylines)
   subtitle("Hypothetical Depression Screener")
   xtitle("Depression, {it:{&theta}}") xlabel(, alt)
;
#delimit cr
```

Replicate ICC and empirical proportions

```
#delimit ;
irtgraph icc d9, title("ICC and Empirical Proportions")
   subtitle("1PL Model, Item 9 (Suicidal Ideation)")
   ytitle("Pr(Response=yes)")
   addplot(scatter d9 Theta)
;
#delimit cr
```

References

From this presentation

The Patient Health Questionnaire (PHQ-9) Quick Depression Assessment is part of the Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MD PHQ) and is copyright 1999 to Pfizer Inc.

Atzema C.L, Schull, M.J., and Tu, J.V. (2011). The effect of a charted history of depression on emergency department triage and outcomes in patients with acute myocardial infarction.*CMAJ: Canadian Medical Association Journal, 183* (6), 663-669.

Guck, T.P., Elasser, G.N., Kavan, M.G., Barone, E.J. (2003). Depression and congestive heart failure. *Congestive Heart Failure, 9*(3):163-9.

Katon, W.J. (2008). The comorbidity of diabetes mellitus and depression. *The American Journal of Medicine, 121*(11 Suppl 2), S8-15.

For good textbook treatments

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

de Boeck, P. and Wilson, M., Eds. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York, Springer.

This presentation was created using reveal.js