Using Marginal Effects for Interpretation in Item Response Theory and for Tests of Differential Item Functioning

Trenton D. Mize

Department of Sociology & Advanced Methodologies
Purdue University

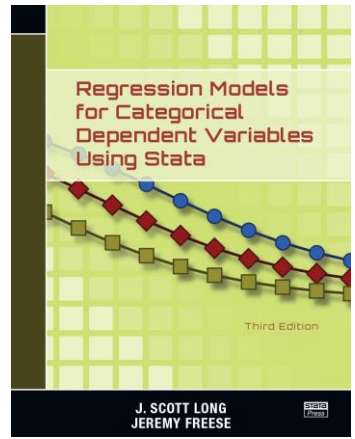
Contents

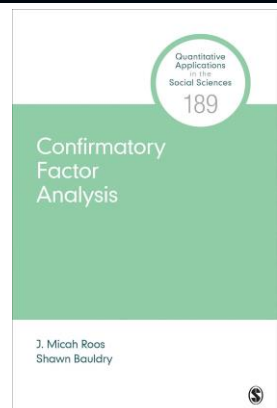
Introduction

1. The field of Categorical Data Analysis has been revolutionized by the use of predictions and marginal effects to understand model results
    a. Thanks in no small part to Long and Freese (**`SPost`**) and Stata's **`margins`** command

2. The latent variable equivalent is Item Response Theory (IRT), implemented by Stata's **`irt`** and **`gsem`** commands

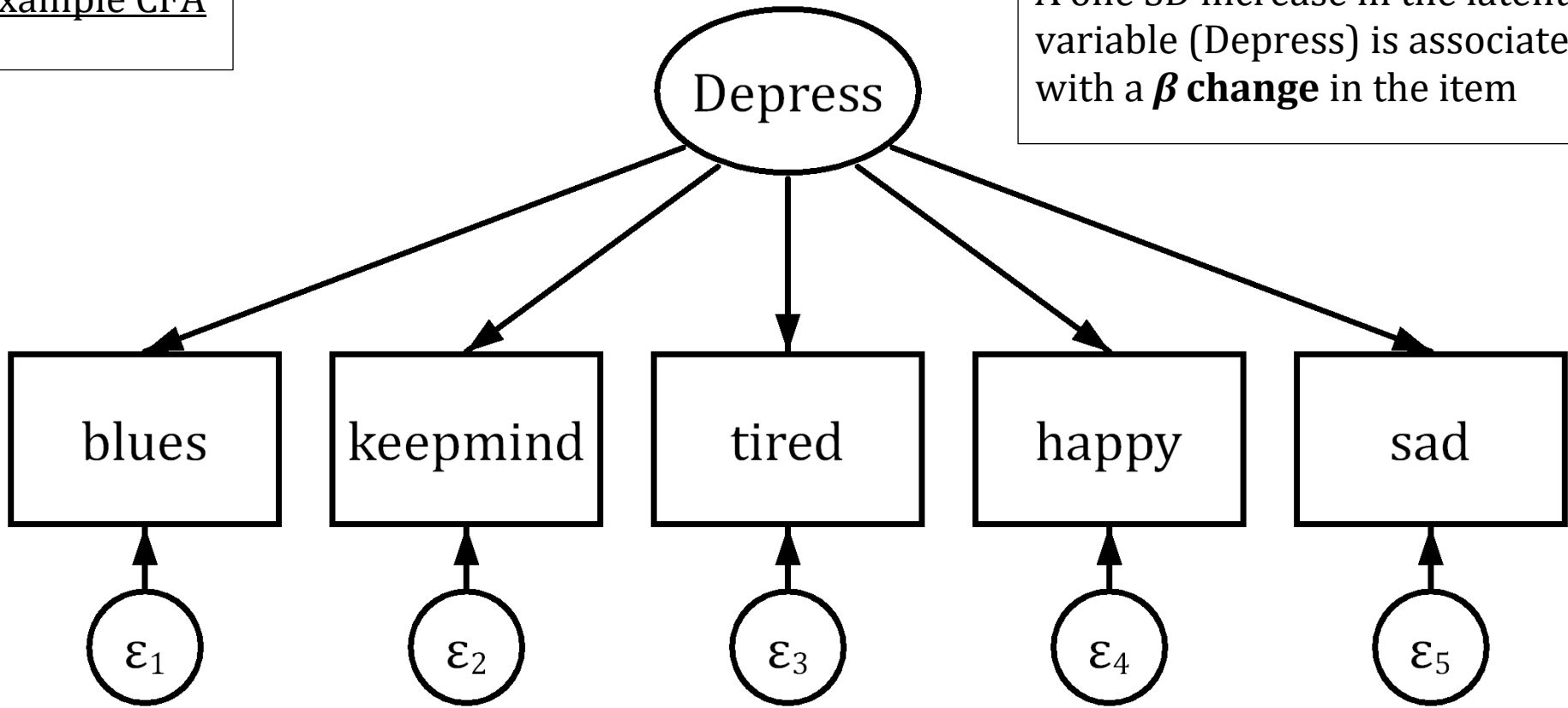
3. We know categorical model coefficients have key limitations—but a similar revolution in interpretation has not yet reached the IRT literature (but see Roos and Bauldry 2022)





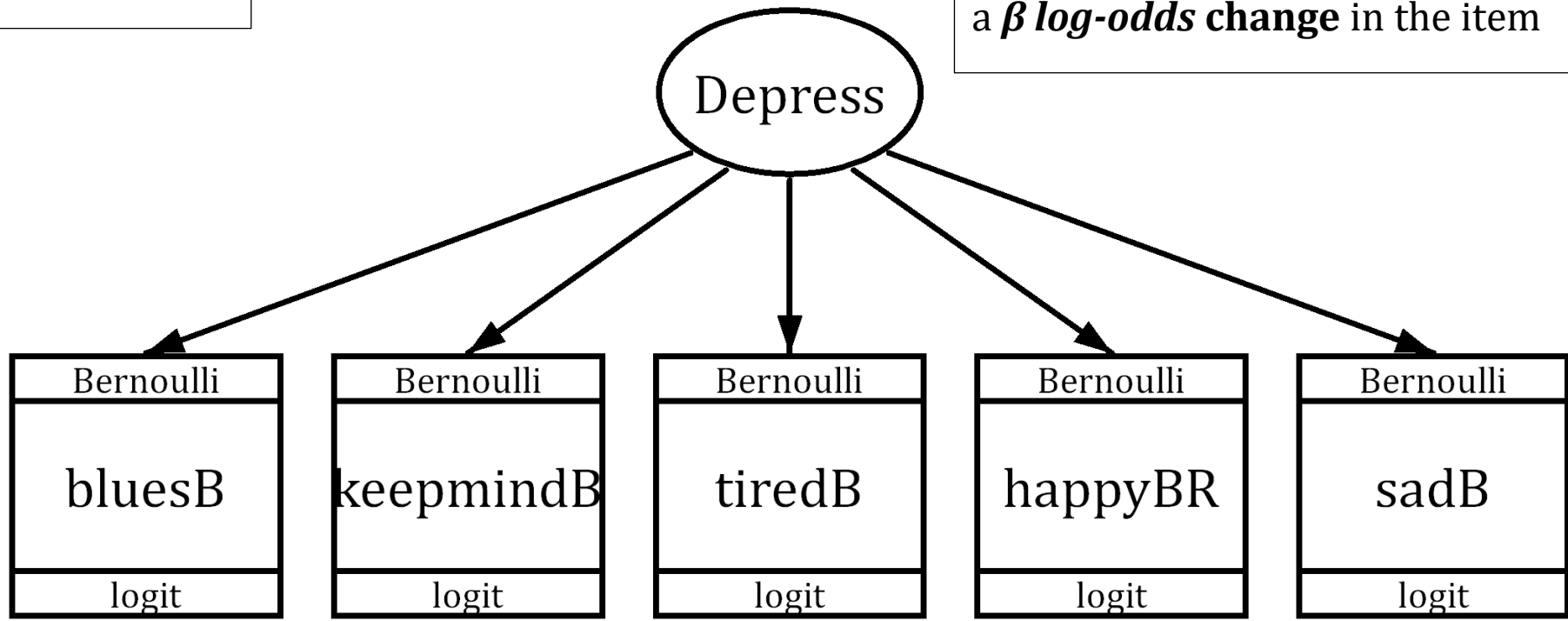
Example CFA

A one SD increase in the latent variable (Depress) is associated with a $\boldsymbol{\beta}$ **change** in the item



Example IRT

A one SD increase in the latent variable (Depress) is associated with a ***β log-odds*** **change** in the item



Item Response Theory

1. In IRT, we treat the paths from the latent variable to observed items as categorical model paths. E.g.,
    a. Binary items = binary logit/probit
    b. Ordinal items = ordinal logit
    c. Etc.

2. To interpret IRT models, we use:
    a. Intercept (difficulty)
    b. Coefficient (discrimination)
    c. Predictions (item characteristics curves)
    d. Etc.

Running Example: CES-D Depression Scale

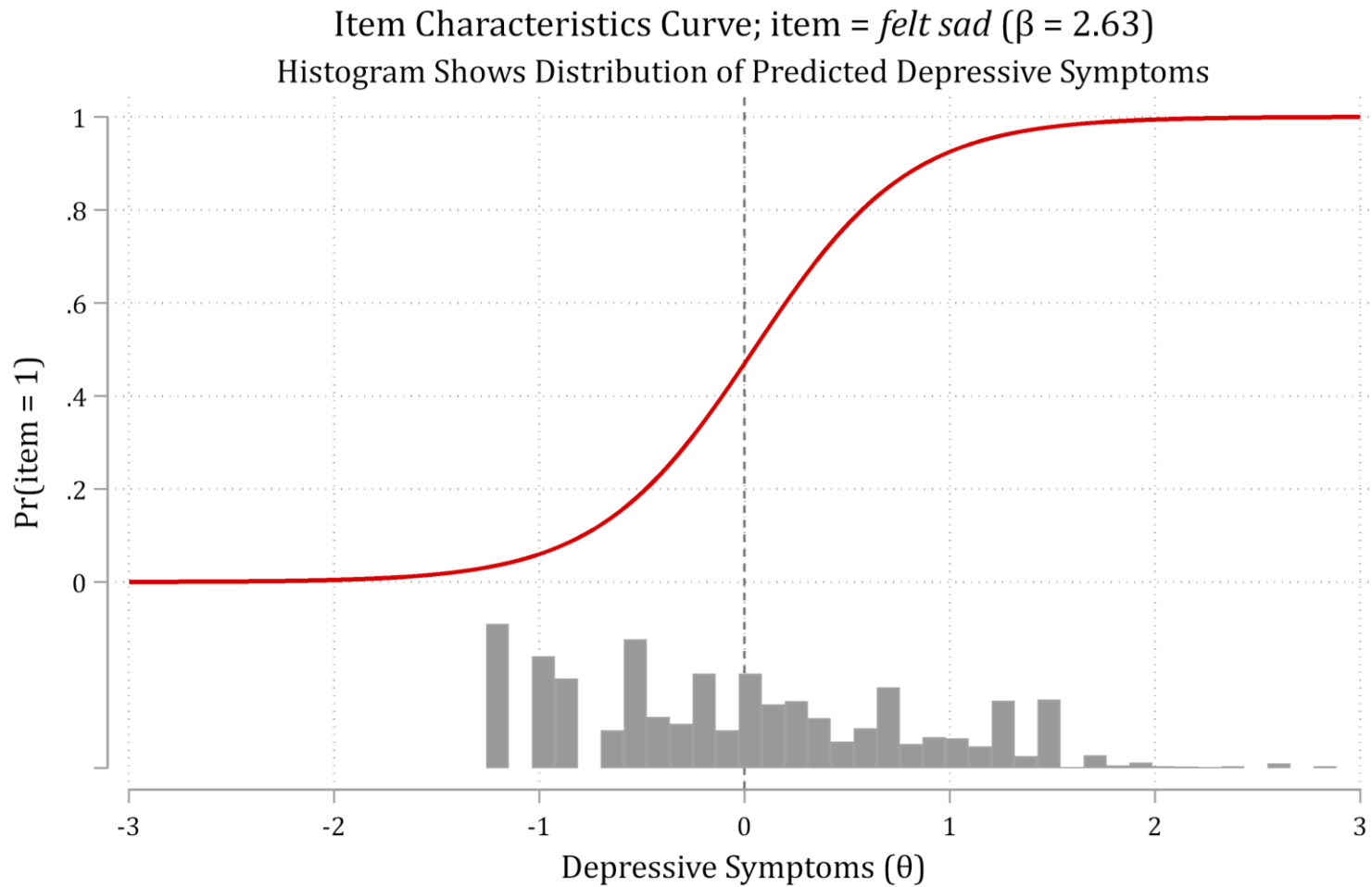
1. Data: Add Health Wave 4 (N = 5,114)
   a. $\overline{age} \approx 28.5$

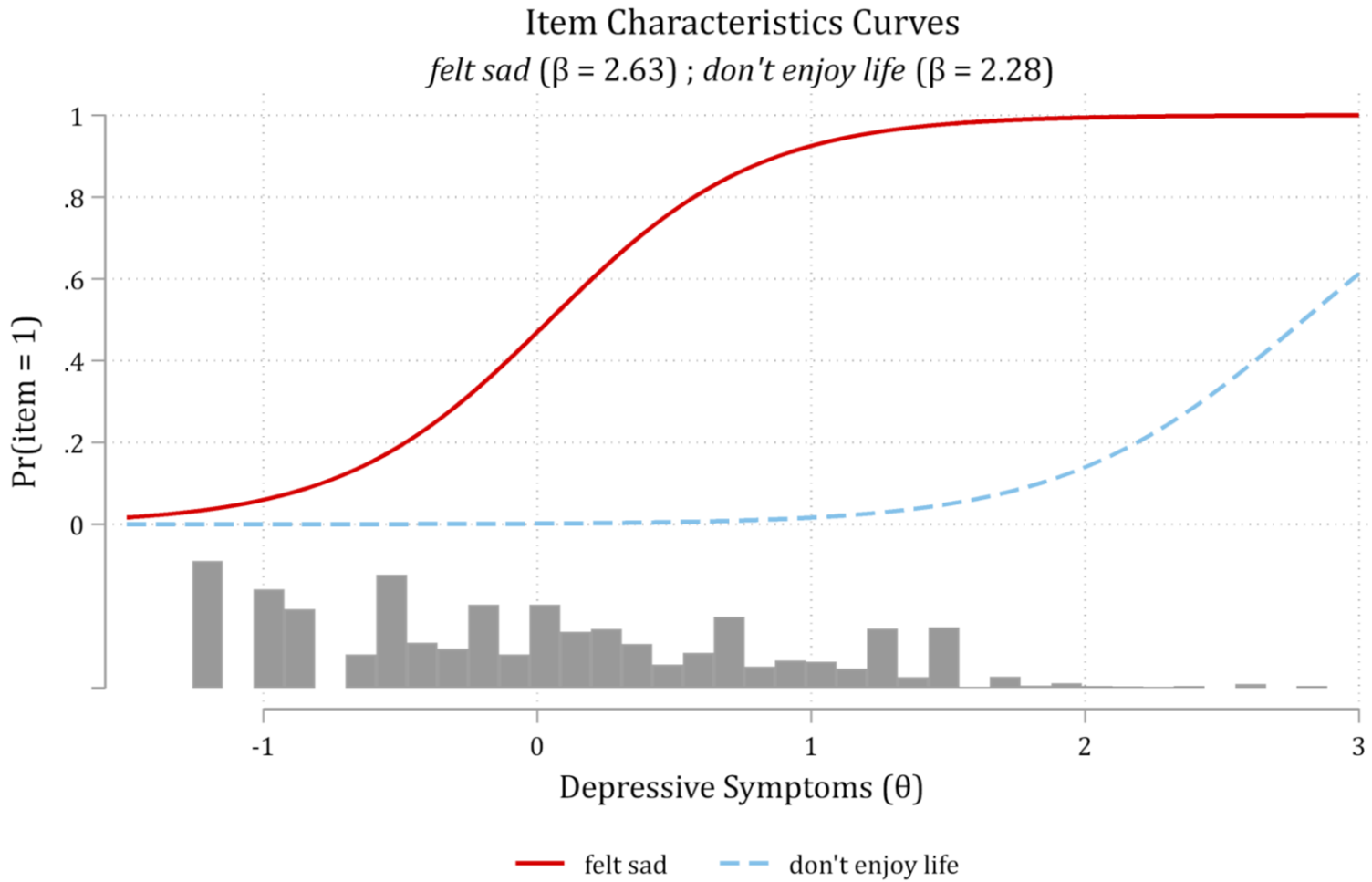
2. Latent variable (θ): *depressive symptoms*
   a. Mean = 0
   b. SD = 1

Set these latent variable means and SDs by convention and convenience

3. 10 items; originally ordinal; binarized to *"never or rarely"* or other
   0 = Not depressive symptomatic
   1 = Depressive symptomatic

```
irt  2pl       botherB bluesB keepmindB depressB tiredB dislikedB ///
               sadB feltgoodBR happyBR enjoylifeBR

Two-parameter logistic model                    Number of obs     =      5,105
Log likelihood = -21248.571
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
bluesB       |
     Discrim |      3.199      0.163   19.640   0.000        2.880       3.518
        Diff |      0.788      0.024   32.554   0.000        0.740       0.835
-------------+----------------------------------------------------------------
feltgoodBR   |
     Discrim |      0.811      0.079   10.235   0.000        0.655       0.966
        Diff |      4.035      0.338   11.920   0.000        3.371       4.698
-------------+----------------------------------------------------------------
sadB         |
     Discrim |      2.626      0.122   21.588   0.000        2.388       2.864
        Diff |      0.045      0.021    2.180   0.029        0.005       0.086
-------------+----------------------------------------------------------------
enjoylifeBR  |
     Discrim |      2.274      0.214   10.605   0.000        1.854       2.695
        Diff |      2.798      0.128   21.924   0.000        2.548       3.049
------------------------------------------------------------------------------
```


Item Characteristics Curve; item = *felt sad* (β = 2.63)
Histogram Shows Distribution of Predicted Depressive Symptoms
Pr(item = 1)
1
.8
.6
.4
.2
0
-3
-2
-1
0
1
2
3
Depressive Symptoms (θ)



Effect sizes very different, despite similar coefficient (discrimination) estimates

Marginal Effects for Interpretation in IRT

1. For a binary DV/item, predicted probability is prediction (η) of interest:

$$\eta = \Pr(item_k = 1) = \frac{\exp(\mathbf{x\beta})}{1 + \exp(\mathbf{x\beta})}$$

2. Consider the formula for a marginal effect (ME) with all observed variables in a regression:

$$ME = \eta(x_k = end, \mathbf{x}_{-k} = \mathbf{x}^*) - \eta(x_k = start, \mathbf{x}_{-k} = \mathbf{x}^*)$$

3. In IRT, no control variables ($\mathbf{x}_{-k}$); focal IV is the latent variable (θ):

$$ME_{IRT} = \eta(\theta = end) - \eta(\theta = start)$$

Choosing start and end values for predictions

1. Where do you make predictions for the marginal effect (ME)?
    a. *Start* could be: θ = its mean (= 0)
    b. *End* could be: θ = its mean + 1
        i. Or mean + SD
    c. Can center the changes on the mean. E.g.,
        i. *Start* = -0.5
        ii. *End* = 0.5

2. E.g., If we set the latent variable to $\theta \sim N(0,1)$, a centered +1 (= SD) marginal effect in IRT:

$$ME_{IRT} = \eta(\theta = -0.5) - \eta(\theta = 0.5)$$

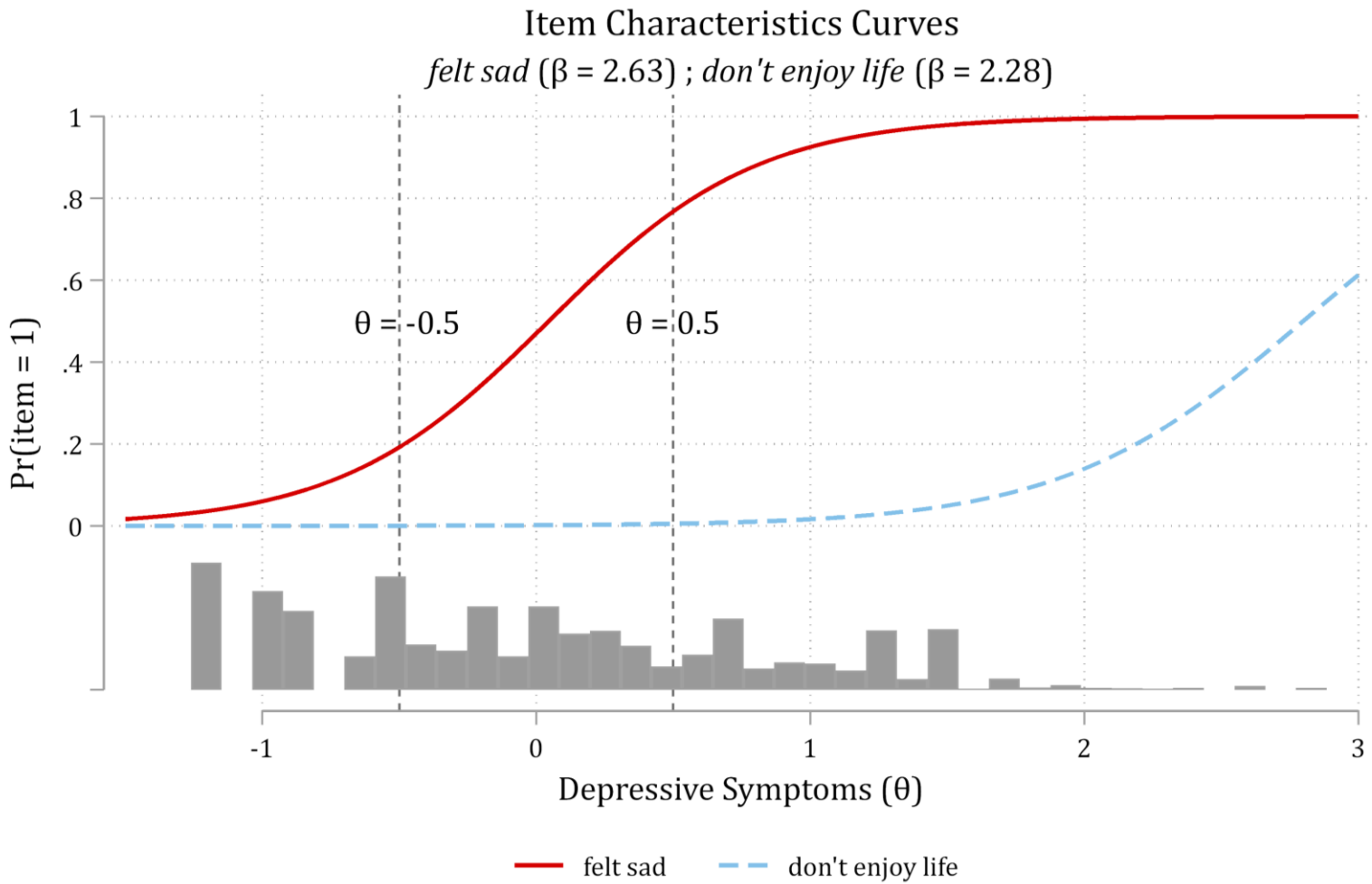
Some technical details

1. Need the "coefficient/constant" regression form parameters to calculate MEs; rather than the traditional IRT "discrimination/difficulty" parameters

$$IRT\ difficulty = \frac{-1 * constant}{disrimination}$$

2. SEs/significance tests of MEs require post-estimation tools. Options:
    a. Delta-method
    b. Bootstrapping

3. MEs based on predictions at observed values (e.g. for "average marginal effects") difficult as the IV (θ) is unobserved…
    a. So, use MEs akin to "marginal effects at the mean"



Average Person

For an average person, a standard deviation increase in depressive symptoms:

$ME_{sad} = 0.58$

$ME_{don't\ enjoy\ life} = 0.01$

Comparing discrimination (coefs) with MEs

Log-odds coefficients and MEs;
sorted by size of coefficient

	coef	ME
tiredB	0.738	0.162
feltgoodBR	0.811	0.029
dislikedB	0.990	0.162
keepmindB	1.028	0.231
botherB	1.454	0.335
happyBR	1.893	0.010
enjoylifeBR	2.274	0.005
sadB	2.626	0.575
bluesB	3.199	0.269
depressB	3.503	0.391

The `irt_me` Command

1. Can fit the model with **`irt`** or **`gsem`**

```
irt 2pl      botherB bluesB keepmindB depressB tiredB dislikedB ///
             sadB feltgoodBR happyBR enjoylifeBR



*equivalent model with gsem

gsem         (Depress -> botherB bluesB keepmindB depressB tiredB dislikedB ///
                         sadB feltgoodBR happyBR enjoylifeBR) ///
             , logit var(Depress@1)
```

Calculating MEs by hand

1. Can use **`nlcom`** to get delta method SEs for the MEs

```
*Calculate 1st (start) prediction
nlcom       exp(_b[bluesB:_cons] + _b[bluesB:Depress]*-0.5) ///
            / (1 + exp(_b[bluesB:_cons] + _b[bluesB:Depress]*-0.5))

*Calculate 2nd (end) prediction
nlcom       exp(_b[bluesB:_cons] + _b[bluesB:Depress]*0.5) ///
            / (1 + exp(_b[bluesB:_cons] + _b[bluesB:Depress]*0.5))

*Calculate difference in predictions (MEM)
nlcom       [exp(_b[bluesB:_cons] + _b[bluesB:Depress]*0.5) ///
            / (1 + exp(_b[bluesB:_cons] + _b[bluesB:Depress]*0.5))] ///
            - ///
            [exp(_b[bluesB:_cons] + _b[bluesB:Depress]*-0.5) ///
            / (1 + exp(_b[bluesB:_cons] + _b[bluesB:Depress]*-0.5))]
```

`irt_me`

1. Calculating MEs by hand is cumbersome and error prone, especially for many items
    a. Formulas for some models (e.g., ordinal, nominal) are even more cumbersome than for binary items

2. **`irt_me`** automates calculation and supports IRT models with items that are:
    a. Binary
    b. Ordinal
    c. Nominal
    d. Count
    e. Continuous
    f. A mix of the above

Title

irt_me — Calculates marginal effects for the latent variable (Theta) after IRT models

General syntax

irt_me [*varlist*] **,** [options]

*Read the help file to see options here

irt_me command after irt commands

```
irt 2pl        botherB bluesB keepmindB depressB tiredB dislikedB ///
               sadB feltgoodBR happyBR enjoylifeBR

irt_me,        help

Marginal Effects of + 1 Increase in Latent Variable (theta) N=5105
```

	PrStart	PrEnd	ME	SE(ME)	P>|z|
botherB	0.239	0.574	0.335	0.013	0.000
bluesB	0.016	0.285	0.269	0.013	0.000
keepmindB	0.523	0.754	0.231	0.010	0.000
depressB	0.021	0.412	0.391	0.017	0.000
tiredB	0.586	0.747	0.162	0.009	0.000
dislikedB	0.136	0.298	0.162	0.007	0.000
sadB	0.193	0.768	0.575	0.020	0.000
feltgoodBR	0.025	0.054	0.029	0.002	0.000
happyBR	0.002	0.012	0.010	0.002	0.000
enjoylifeBR	0.001	0.005	0.005	0.001	0.000

```
PrStart : Pr(y=1) at theta = -0.5
PrEnd   : Pr(y=1) at theta = 0.5
ME      : PrEnd - PrStart
```

Option **help** adds footnotes to the table to explain the output

`irt_me` command after `gsem` commands

1. If using **`gsem`**, have to specify the name of the latent variable in the **`latent( )`** option

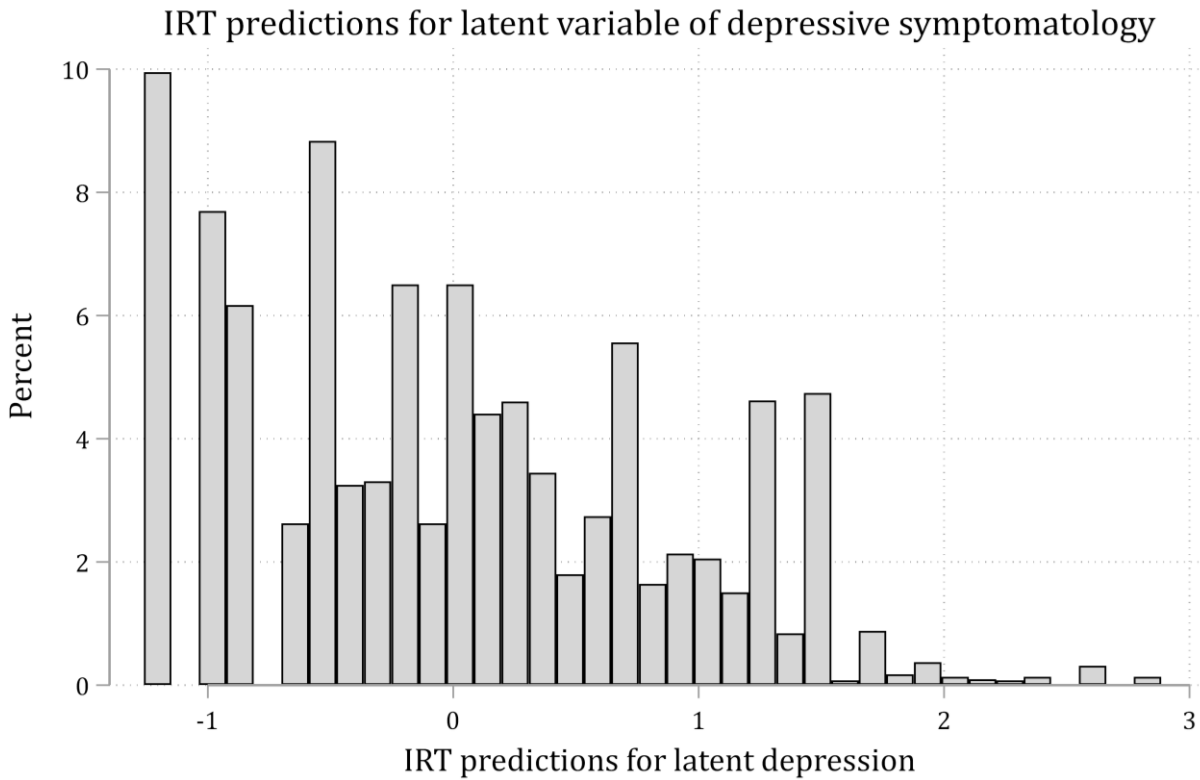
```
gsem            (Depress -> botherB bluesB keepmindB depressB tiredB dislikedB ///
                            sadB feltgoodBR happyBR enjoylifeBR) ///
                , logit var(Depress@1)

irt_me,         latent(Depress)
```

```
Marginal Effects of + 1 Increase in Latent Variable (theta) N=5105
```

	PrStart	PrEnd	ME	SE(ME)	P>|z|
botherB	0.239	0.574	0.335	0.013	0.000
bluesB	0.016	0.285	0.269	0.013	0.000
keepmindB	0.523	0.754	0.231	0.010	0.000
depressB	0.021	0.412	0.391	0.017	0.000
tiredB	0.586	0.747	0.162	0.009	0.000
dislikedB	0.136	0.298	0.162	0.007	0.000
sadB	0.193	0.768	0.575	0.020	0.000
feltgoodBR	0.025	0.054	0.029	0.002	0.000
happyBR	0.002	0.012	0.010	0.002	0.000
enjoylifeBR	0.001	0.005	0.005	0.001	0.000

Calculating MEs across larger ranges of θ

1. Consider the distribution of the predictions for θ in the sample

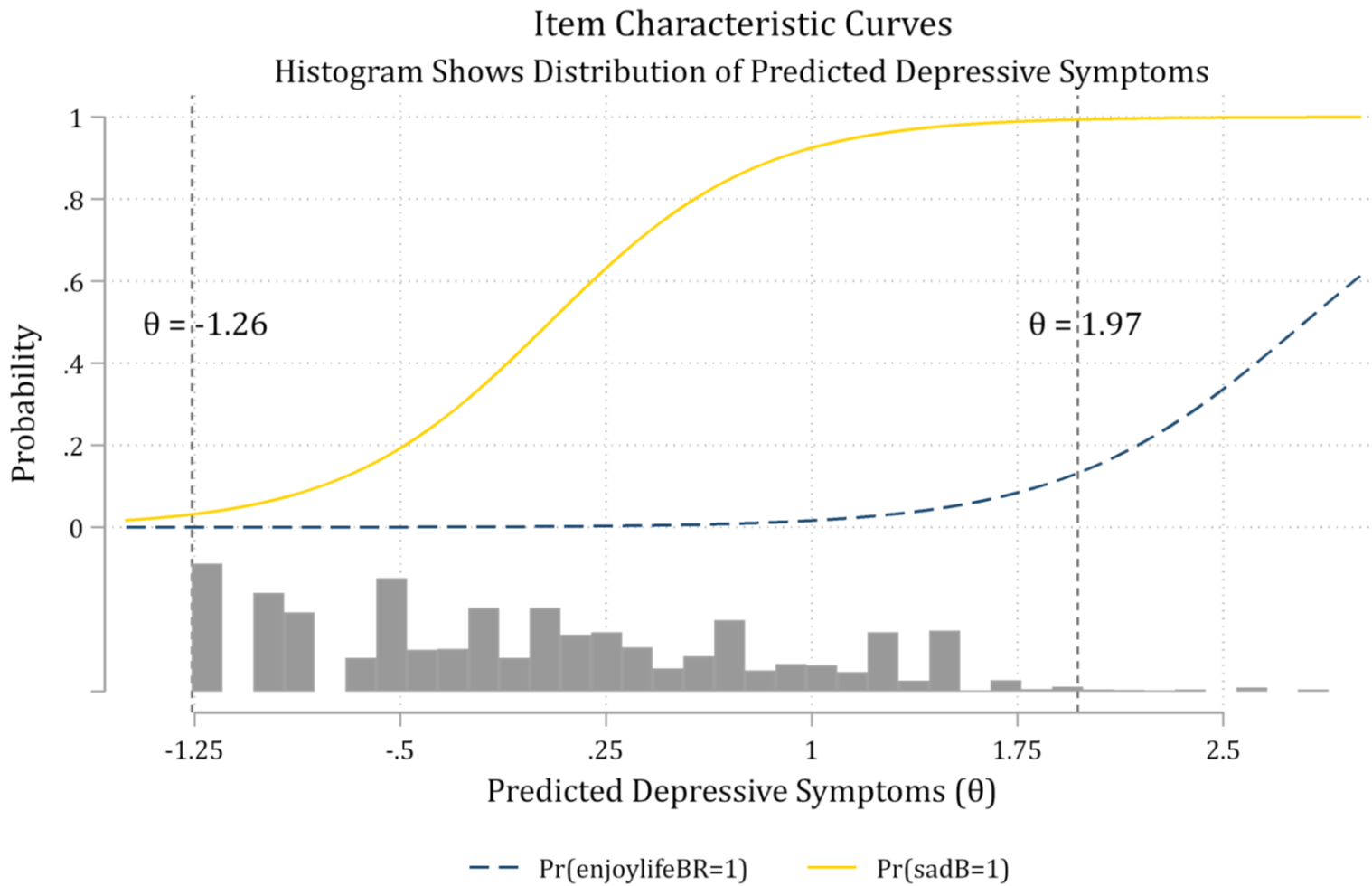


Using the `range` option

1. To get a sense of effects across most of the distribution, we can calculate an ME for a change in θ from the 1st to the 99th percentile

$$ME = \eta(\theta = P_{99}) - \eta(\theta = P_1) = \eta(\theta = 1.97) - \eta(\theta = -1.26)$$

2. Use the **`range`** option for MEs across the trimmed range (1st to 99th percentile)


Item Characteristic Curves
Histogram Shows Distribution of Predicted Depressive Symptoms
Probability
1
.8
.6
.4
.2
0
θ = -1.26
θ = 1.97
-1.25
-.5
.25
1
1.75
2.5
Predicted Depressive Symptoms (θ)
Pr(enjoylifeBR=1)
Pr(sadB=1)

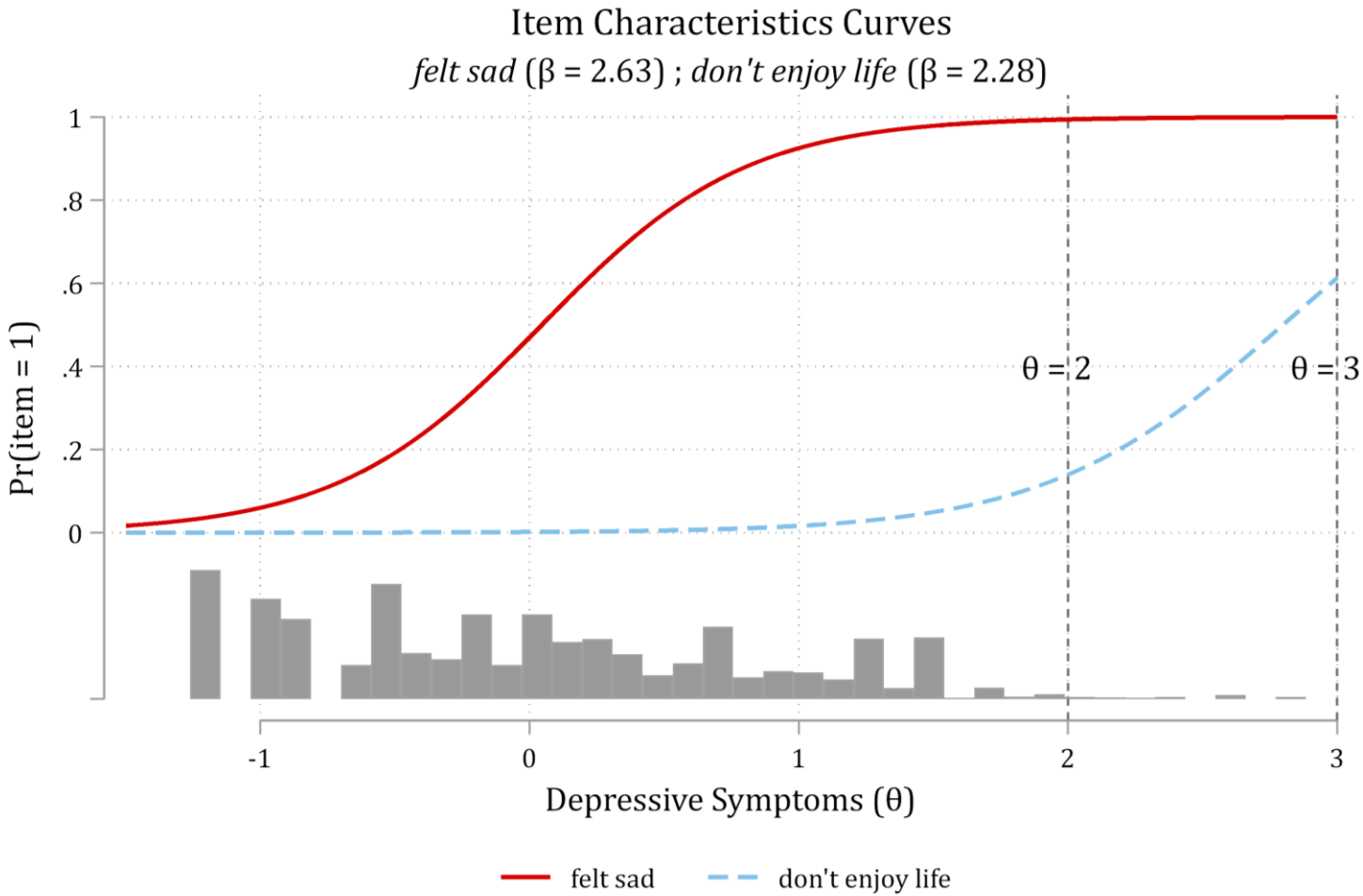
Using the `range` option

```
irt_me,   range help

Marginal Effects of + 3.227 Increase in Latent Variable (theta) N=5105

             |    PrStart       PrEnd     ME Est.   Std. Err.       P>|z|
-------------+------------------------------------------------------------
     botherB |      0.094       0.919       0.825       0.015       0.000
      bluesB |      0.001       0.978       0.976       0.006       0.000
   keepmindB |      0.334       0.933       0.598       0.019       0.000
    depressB |      0.001       0.992       0.990       0.003       0.000
      tiredB |      0.446       0.897       0.451       0.020       0.000
   dislikedB |      0.069       0.645       0.576       0.024       0.000
        sadB |      0.031       0.994       0.962       0.006       0.000
  feltgoodBR |      0.013       0.158       0.144       0.018       0.000
     happyBR |      0.000       0.163       0.162       0.020       0.000
 enjoylifeBR |      0.000       0.131       0.131       0.018       0.000


PrStart : Pr(y=1) at theta = -1.261
PrEnd   : Pr(y=1) at theta = 1.967
ME      : PrEnd - PrStart
```

Specify custom values/ideal types

1. Options **`start(#)`** and **`end(#)`** allow you to specify custom values of θ to make the predictions
    a. E.g., Ideal types of "very depressed" from 2 to 3 (SDs above the mean)

```
irt_me,   start(2) end(3) help
```



Ideal Types

For someone with high depressive symptomology, an increase from high (2) to extremely high (3) depressive symptoms:

$ME_{sad} = 0.01$

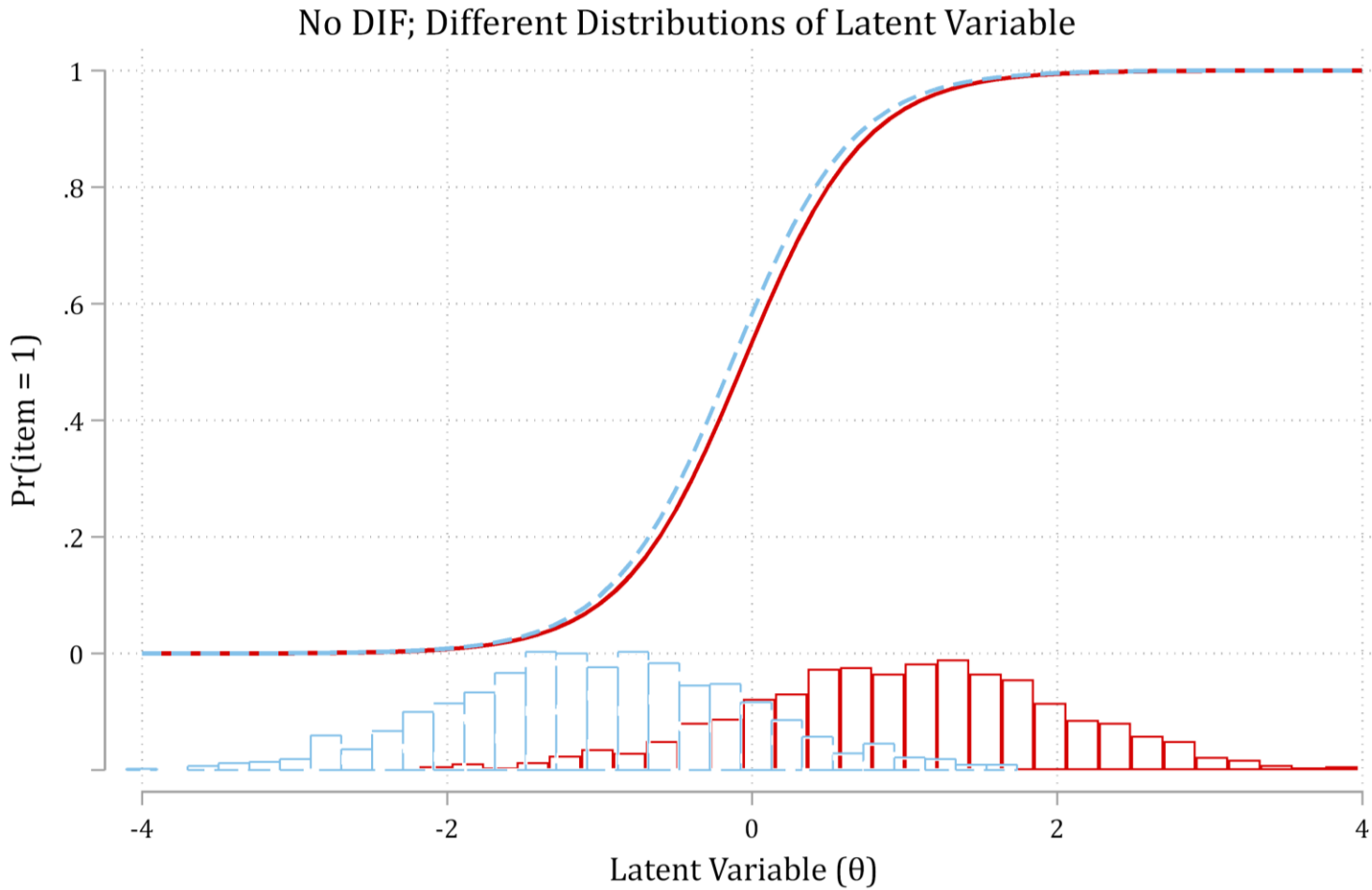
$ME_{don't\ enjoy\ life} = 0.47$

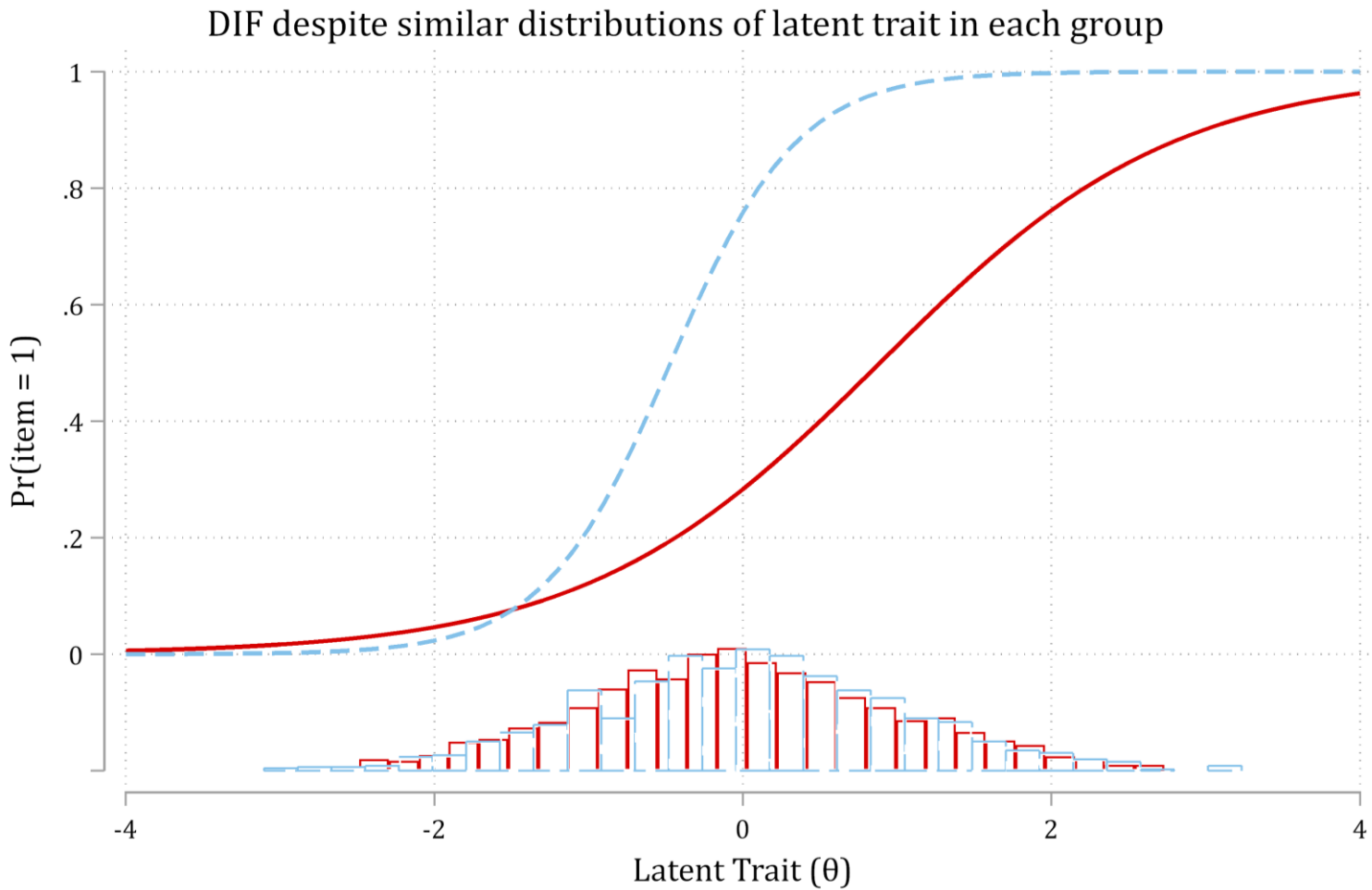
Differential Item Functioning (DIF)

1. The utility of using MEs in IRT is highlighted by tests of item bias

2. Tests and scales are made up of questions/items
    a. A **biased test** is one that includes **biased items**

3. Differential item functioning (DIF) is a test of item bias
    a. E.g., Is the question *"did you feel sad last week?"* (item) equally good at picking up on *depressive symptoms* (latent variable) for *men* and for *women* (groups)?


No DIF; Different Distributions of Latent Variable
Pr(item = 1)
1
.8
.6
.4
.2
0
-4
-2
0
2
4
Latent Variable (θ)


DIF despite similar distributions of latent trait in each group
Pr(item = 1)
1
.8
.6
.4
.2
0
-4
-2
0
2
4
Latent Trait (θ)

Tests of differential item functioning in IRT

1. Either use a naïve estimate (t_i) of the latent variable = sum of all of the items; or the IRT parameter (θ)
   a. A binary grouping variable of interest (e.g., gender) = g_i

2. For a test of DIF: estimate binary logit model (or IRT) for each item (k):

$$logit\{\Pr(item_{k,i})\} = \alpha_k + \beta_k t_i + \beta_g g_i + \beta_{tXg}(t_i * g_i)$$

$$logit\{\Pr(item_{k,i})\} = \alpha_k + \beta_k \theta_i + \beta_g g_i + \beta_{\theta Xg}(\theta_i * g_i)$$

   b. The significance test on the interaction (product term) coefficient β_{tXg} (or $\beta_{\theta Xg}$) is interpreted as a test of (non-uniform) DIF
      i. As implemented in **`diflogistic`**

Issues with tests of differential item functioning (DIF)

1. There is no issue with the logic of tests of DIF
    a. However, it is now well-established that tests of interaction in the coefficient metric ≠ a test of interaction in the prediction metric (Mize 2019)
    b. Solution: Use tests of predictions and marginal effects in prediction metric as recommended in categorical data analysis literature

2. Example
    a. Performance of depression items across men and women

DIF tests using 2nd differences of MEs

1. Fit the IRT model using **gsem** with the **group( )** option

```
gsem          (Depress -> botherB bluesB keepmindB depressB tiredB dislikedB ///
                          sadB feltgoodBR happyBR enjoylifeBR, logit) ///
              , mean(Depress@0) var(Depress@1) group(woman) ginvariant(none)
```

2. Use **nlcom** to calculate equality of MEs across groups (i.e., 2nd differences)

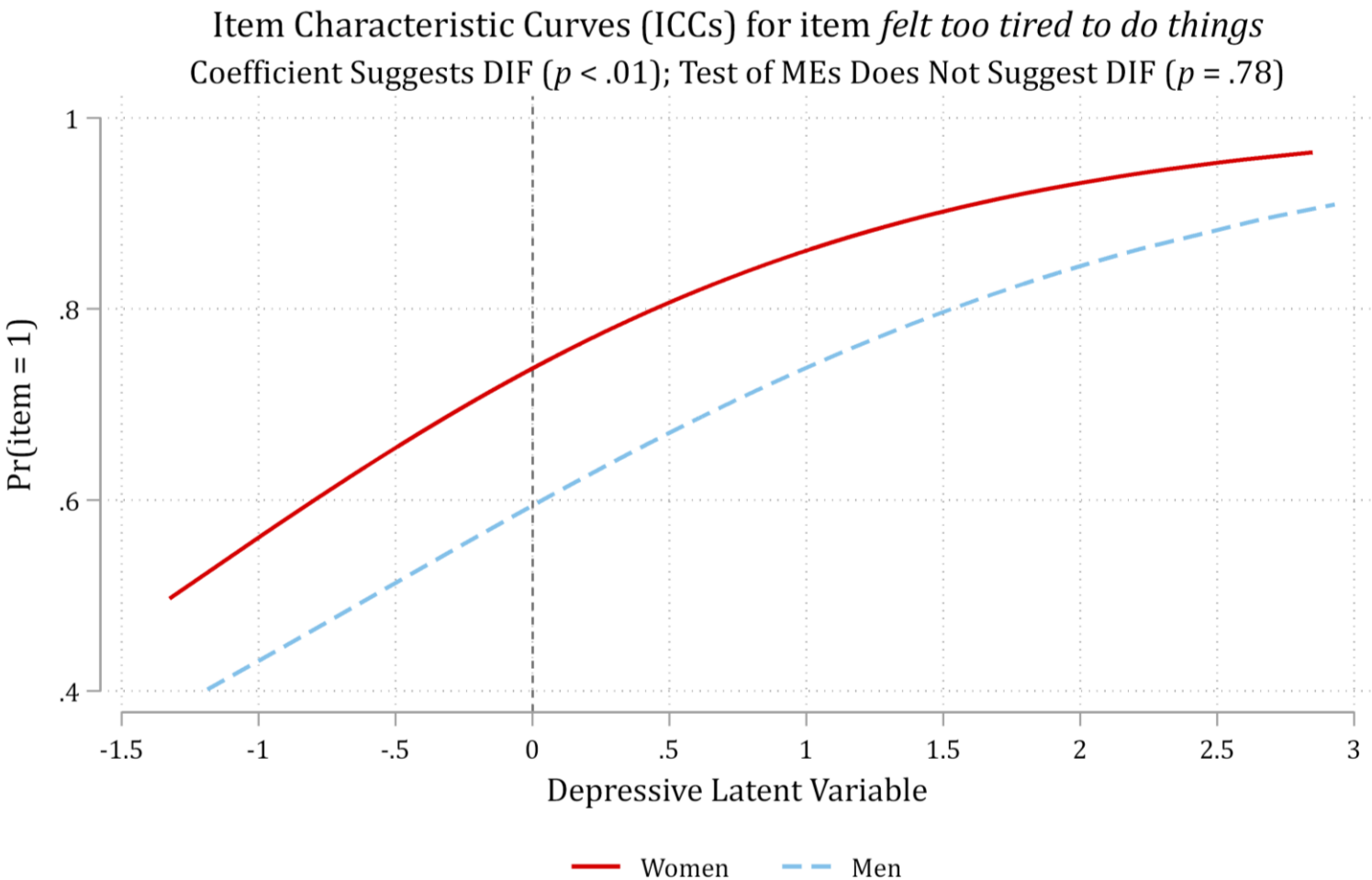
```
nlcom         [[exp(_b[bluesB:0.woman] + _b[bluesB:0.woman#c.Depress]*0.5)     / ///
              (1 + exp(_b[bluesB:0.woman] + _b[bluesB:0.woman#c.Depress]*0.5))] ///
              - ///
              [exp(_b[bluesB:0.woman] + _b[bluesB:0.woman#c.Depress]*-0.5)     / ///
              (1 + exp(_b[bluesB:0.woman] + _b[bluesB:0.woman#c.Depress]*-0.5))]] ///
              - ///
              [[exp(_b[bluesB:1.woman] + _b[bluesB:1.woman#c.Depress]*0.5)     / ///
              (1 + exp(_b[bluesB:1.woman] + _b[bluesB:1.woman#c.Depress]*0.5))] ///
              - ///
              [exp(_b[bluesB:1.woman] + _b[bluesB:1.woman#c.Depress]*-0.5)     / ///
              (1 + exp(_b[bluesB:1.woman] + _b[bluesB:1.woman#c.Depress]*-0.5))]]
```

Comparison of tests of DIF (*p*-values shown)

	diflogistic Uniform	diflogistic Non-Uniform	ME 2nd Diff
botherB	0.023	0.792	0.556
bluesB	0.652	0.718	0.000
keepmindB	0.165	0.118	0.070
depressB	0.395	0.004	0.000
tiredB	0.000	0.008	0.782
dislikedB	0.000	0.219	0.755
sadB	0.057	0.528	0.181
feltgoodBR	0.072	0.065	0.298
happyBR	0.577	0.893	0.127
enjoylifeBR	0.789	0.469	0.318

Note: "uniform" DIF is for differences in intercepts across groups; "non-uniform" for differences in coefficients across groups.

ME 2nd Diff approach uses both in its calculations



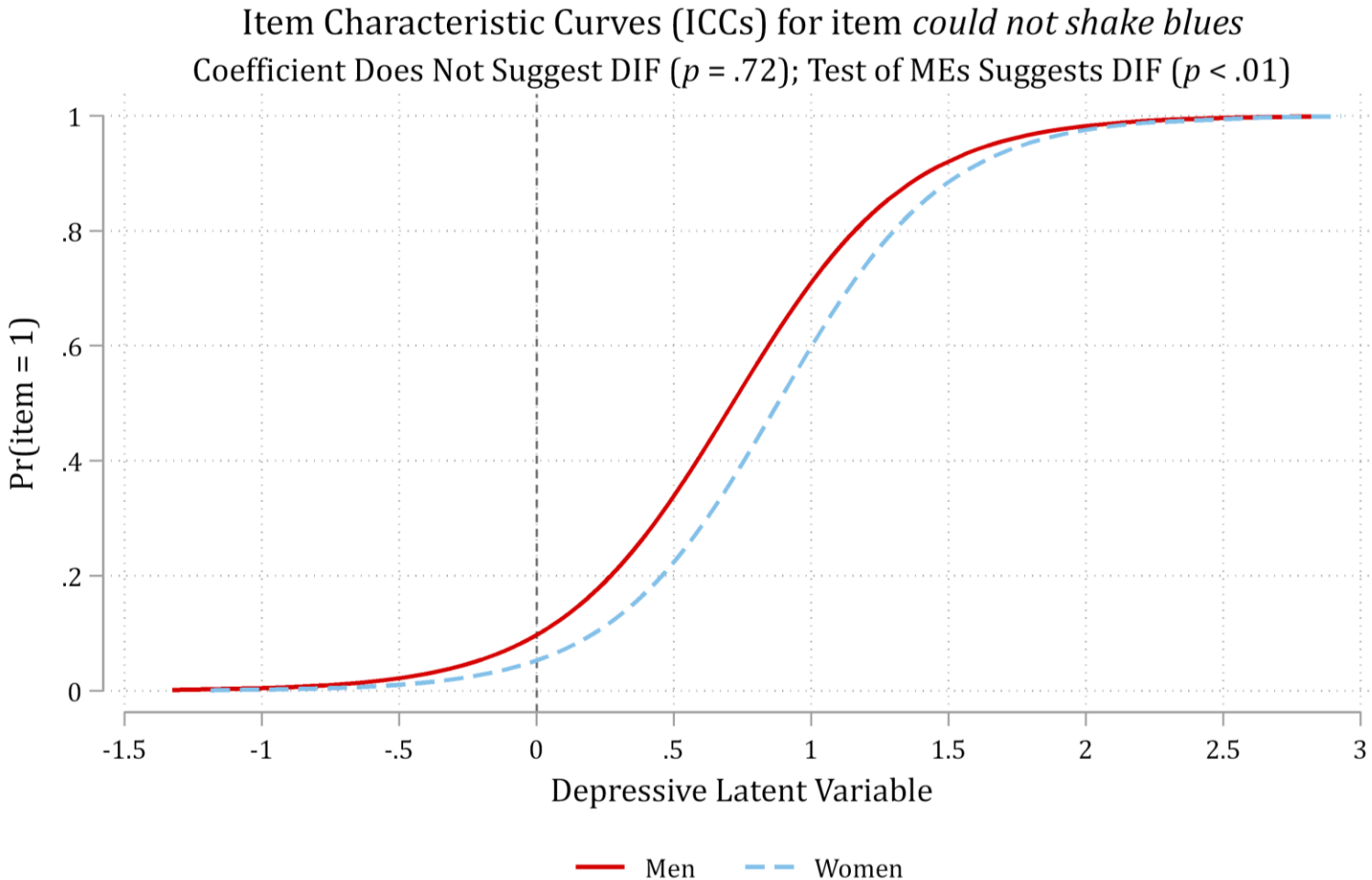
For the average person:

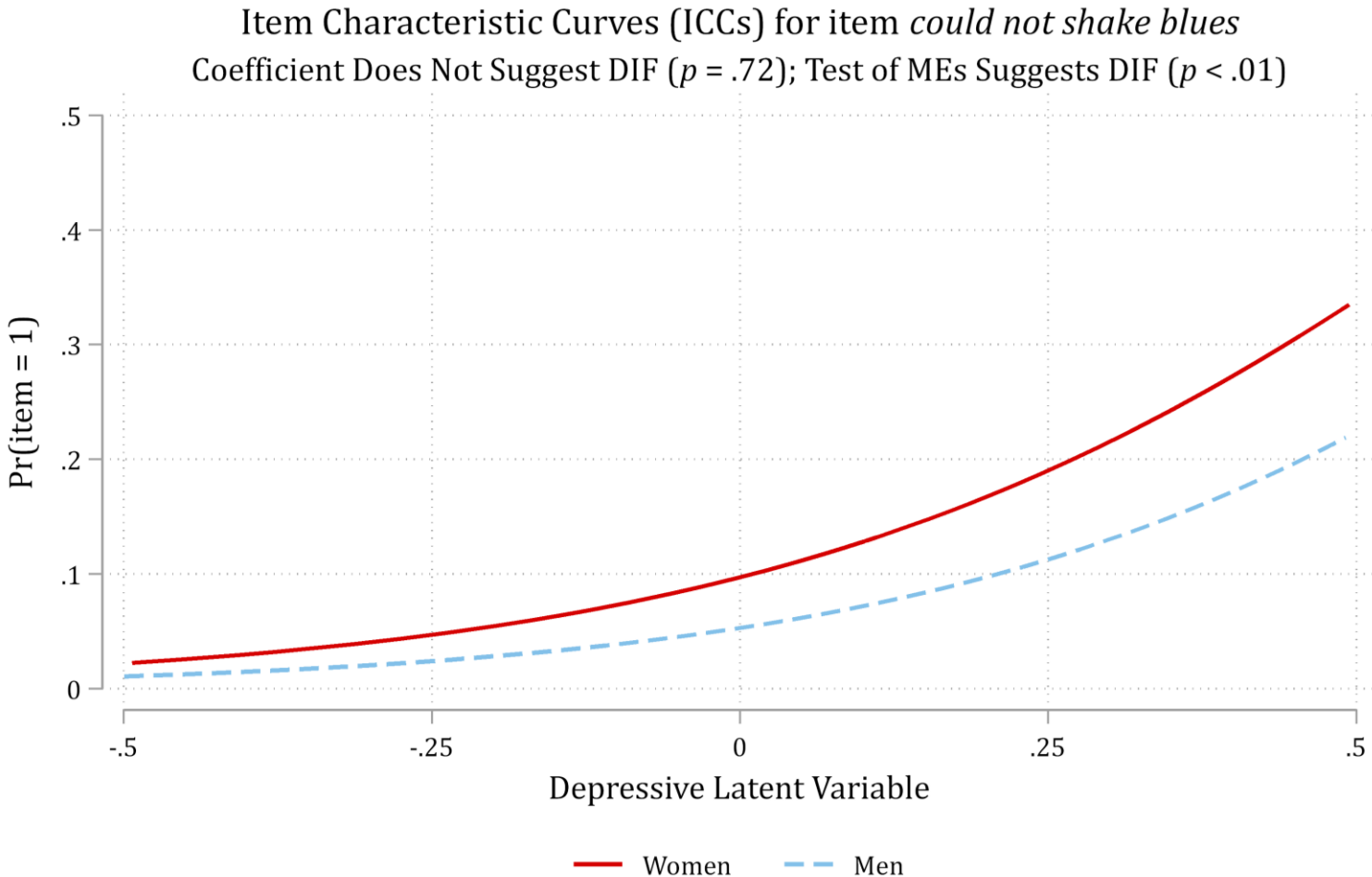
ME_{men} = 0.16

ME_{women} = 0.15

Gender diff. in MEs:

$p = .782$


Item Characteristic Curves (ICCs) for item *could not shake blues*
Coefficient Does Not Suggest DIF (p = .72); Test of MEs Suggests DIF (p < .01)
Pr(item = 1)
1
.8
.6
.4
.2
0
-1.5
-1
-.5
0
.5
1
1.5
2
2.5
3
Depressive Latent Variable
Men
Women



For the average person:

ME_{men} = 0.22

ME_{women} = 0.34

Gender diff. in MEs: $p < .001$

Questions?

1. Slides, example code, and a beta-version of the **`irt_me`** command available at:

   www.trentonmize.com/software/irt_me

2. **`irt_dif`** command to automate tests of differential item functioning in progress

3. Questions on the content and/or requests for additions to the commands can be sent to:

   tmize@purdue.edu