

# Assessment of Model Fit

Lecture #6

ICPSR Item Response Theory Workshop

# Lecture Overview

- An overview of model fit for IRT Models using Mplus
- Model fit is used to help:
  - Determine if a model fits the data well enough in an absolute sense to use the examinee estimates
  - Select best model among competing models
- Fraction subtraction data will be used to illustrate model fit in practice

# ASSESSMENT OF MODEL FIT

# Assessing Model Fit

- There is no one best way to assess fit in IRT Models
- Techniques typically used can be put into several general categories:
  - Absolute fit
    - ♦ Model based hypothesis tests (if available)
  - Relative fit
    - ♦ Information criteria
  - Item fit
    - ♦ Univariate
    - ♦ Bivariate
- Topics discussed here will mainly focus on fit statistics available in Mplus

# Overall Model Fit: Chi-Squared Test

- For small numbers of items (10-15), the traditional Chi-Squared test of model fit can be used
  - Test is invalid for too many items – sparse data
- Mplus gives this automatically
  - Omits when data are sparse
  - Can omit extreme cells from an analysis
    - ♦ Misleading

# Relative Model Fit: Information Criteria

- Used when comparing between two models
  - 1PL v. 2PL
- Mplus reports:
  - AIC and BIC
  - Sample size adjusted BIC
- All can be used
  - Smallest value is best
- Here, 2PL Model is Preferred using AIC/BIC

- Fraction Subtraction

## 1PL:

MODEL FIT INFORMATION	
Number of Free Parameters	21
Loglikelihood	
H0 value	-4797.178
Information Criteria	
Akaike (AIC)	9636.355
Bayesian (BIC)	9726.322
Sample-Size Adjusted BIC ( $n^* = (n + 2) / 24$ )	9659.661

## 2PL:

MODEL FIT INFORMATION	
Number of Free Parameters	40
Loglikelihood	
H0 value	-4640.159
Information Criteria	
Akaike (AIC)	9360.319
Bayesian (BIC)	9531.684
Sample-Size Adjusted BIC ( $n^* = (n + 2) / 24$ )	9404.711

# Chi-Squared (Deviance) Test

- The 1PL and 2PL are nested models
  - Can use deviance test to statistically test for fit

- Change in  $-2 \times \log\text{likelihood}$ :  
 $-2 \times (-4797.178 - -4640.159) = 314.038$

- Change in DF:  
 $40 - 21 = 19$

- Chi-Square p-value  $< 0.0001$
- Conclusion: 2PL is preferred statistically

- Fraction Subtraction

## 1PL:

MODEL FIT INFORMATION	
Number of Free Parameters	21
Loglikelihood	
H0 Value	-4797.178
Information Criteria	
Akaike (AIC)	9636.355
Bayesian (BIC)	9726.322
Sample-Size Adjusted BIC ( $n^* = (n + 2) / 24$ )	9659.661

## 2PL:

MODEL FIT INFORMATION	
Number of Free Parameters	40
Loglikelihood	
H0 Value	-4640.159
Information Criteria	
Akaike (AIC)	9360.319
Bayesian (BIC)	9531.684
Sample-Size Adjusted BIC ( $n^* = (n + 2) / 24$ )	9404.711

# Item Fit Statistics

- The TECH10 option reports a degree of misfit for each
  - Item individually (Univariate)
  - Pair of two items (Bivariate)
- Uses Chi-Squared test for misfit
  - Values for each item are distributed as Chi-square with 1 df (for binary items)
- Misfitting items can be investigated
  - Items can be removed
  - Multidimensional model may be used



# Item Fit Statistics: Univariate Fit

- Univariate fit attempts to determine if the model fits each item marginally
  - Limited information statistic
- Not \*that\* useful in IRT
  - Scale of fit is usually small
  - Most items “fit”
- H1: Observed Probability
- H0: IRT Model Prediction
- Chi-Square critical values
  - (0.05) = 3.84
  - (0.01) = 6.63

## UNIVARIATE MODEL FIT INFORMATION

variable	Estimated Probabilities		Standardized Residual (z-score)
	H1	H0	
x1			
Category 1	0.468	0.479	-0.511
Category 2	0.532	0.521	0.511
Univariate Pearson Chi-Square			0.261
Univariate Log-Likelihood Chi-Square			0.262
x2			
Category 1	0.424	0.434	-0.487
Category 2	0.576	0.566	0.487
Univariate Pearson Chi-Square			0.237
Univariate Log-Likelihood Chi-Square			0.238
x3			
Category 1	0.485	0.498	-0.611
Category 2	0.515	0.502	0.611
Univariate Pearson Chi-Square			0.373
Univariate Log-Likelihood Chi-Square			0.373

# Item Fit Statistics: Bivariate Fit

- Bivariate fit is an index of fit for a pair of items
- Compares observed data with frequency expected under IRT model
  - Produces a 1-df Chi-Squared test for binary items
- Can help identify items that do not fit model
  - Rough approximation

BIVARIATE MODEL FIT INFORMATION

variable	variable	H1	Estimated Probabilities		Standardized Residual (z-score)
			H0		
x1	x2				
Category 1	Category 1	0.390	0.338		2.542
Category 1	Category 2	0.078	0.141		-4.185
Category 2	Category 1	0.034	0.096		-4.903
Category 2	Category 2	0.498	0.425		3.438
Bivariate Pearson Chi-Square					47.850
Bivariate Log-Likelihood Chi-Square					57.544
x1	x3				
Category 1	Category 1	0.410	0.363		2.304
Category 1	Category 2	0.058	0.117		-4.245
Category 2	Category 1	0.075	0.136		-4.126
Category 2	Category 2	0.457	0.385		3.429
Bivariate Pearson Chi-Square					41.247
Bivariate Log-Likelihood Chi-Square					47.257
x1	x4				
Category 1	Category 1	0.315	0.325		-0.481
Category 1	Category 2	0.153	0.154		-0.083
Category 2	Category 1	0.157	0.152		0.288
Category 2	Category 2	0.375	0.368		0.315
Bivariate Pearson Chi-Square					0.295
Bivariate Log-Likelihood Chi-Square					0.296

# Fraction Subtraction Results: Model Fit

- Univariate model fit
  - Compares model predicted and observed frequencies of responses for all items marginally
  - Of 20 items, none had p-values less than 0.01
- Bivariate model fit
  - Compares model predicted and observed frequencies of responses for all pairs of items
  - Of 190 item pairs 35 had p-values less than 0.01
  - Items most indicated
    - ♦ Item 4 (8 pairs) :  $3\frac{1}{2} - 2\frac{3}{2}$
    - ♦ Item 6 (7 pairs):  $\frac{6}{7} - \frac{4}{7}$
    - ♦ Item 20 (7 pairs):  $4\frac{1}{3} - 1\frac{5}{3}$
- Indicates some items are not fit well by model
  - We will ignore this and continue with analysis as example

# **“TRADITIONAL” IRT FIT METHODS**

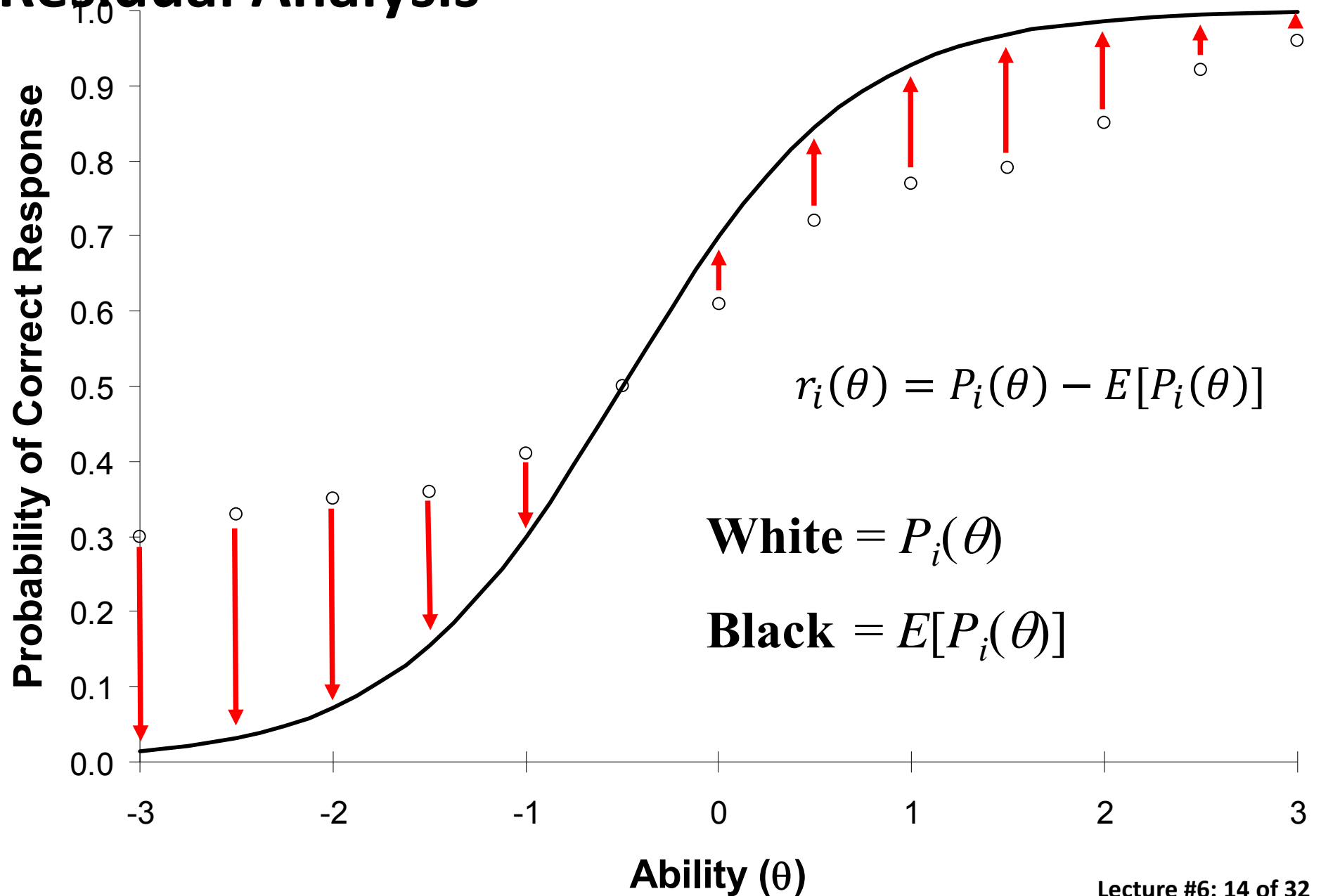
# Residual Analysis

- Assess the accuracy of model predictions versus actual data
  - Residual = difference between observed proportion and predicted probability:

$$r_i(\theta) = P_i(\theta) - E[P_i(\theta)]$$

- $P_i(\theta)$  = observed proportion correct for a given **theta** level
- $E[P_i(\theta)]$  = expected proportion correct (i.e., probability from the IRT model)

# Residual Analysis



# Standardized Residuals

- Raw residuals do not take into account the error associated with the expected proportion correct, so we standardize each by dividing by its standard error:

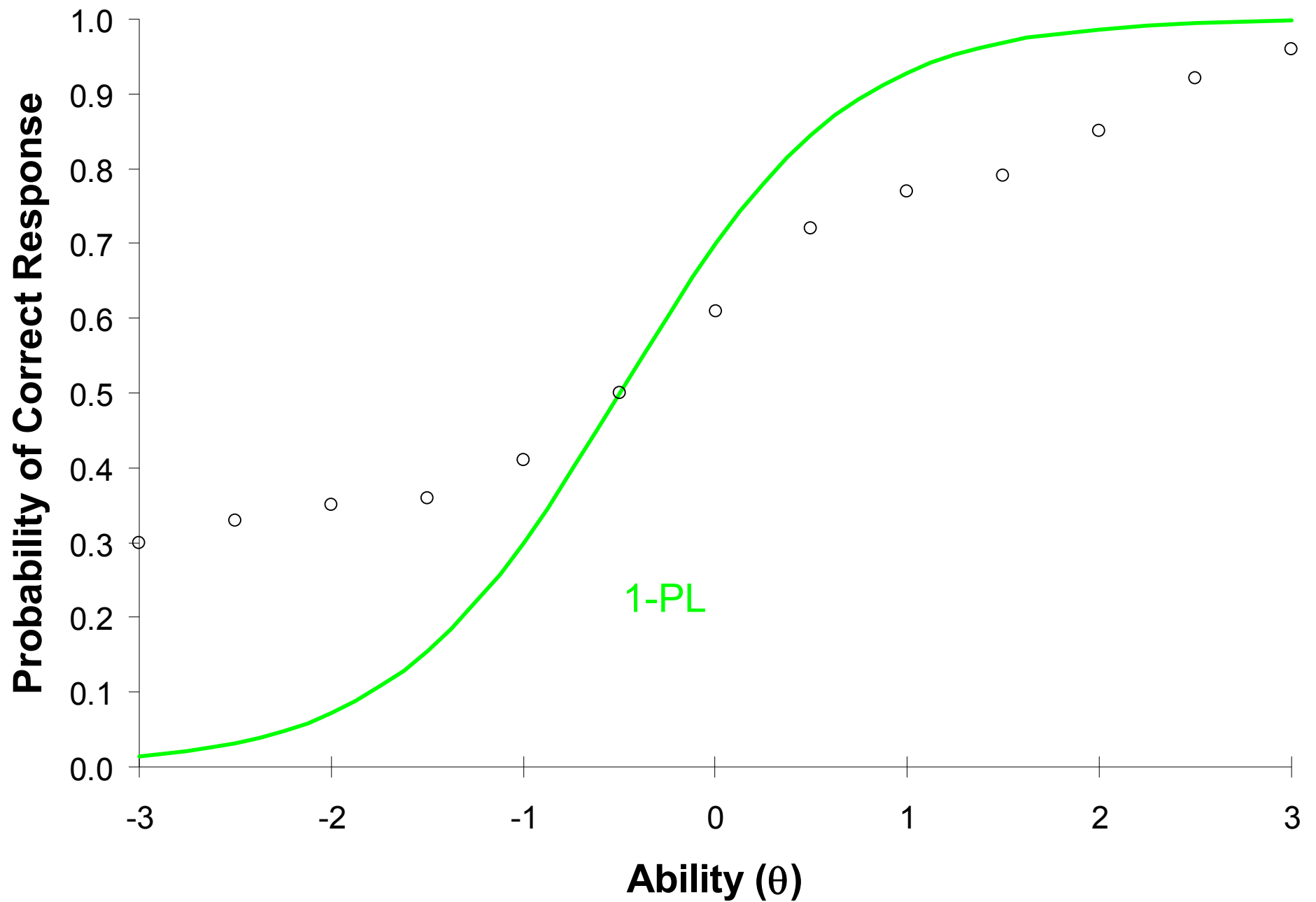
$$SE(E[P_i(\theta)]) = \sqrt{\frac{E[P_i(\theta)]E[1 - P_i(\theta)]}{N(\theta)}}$$

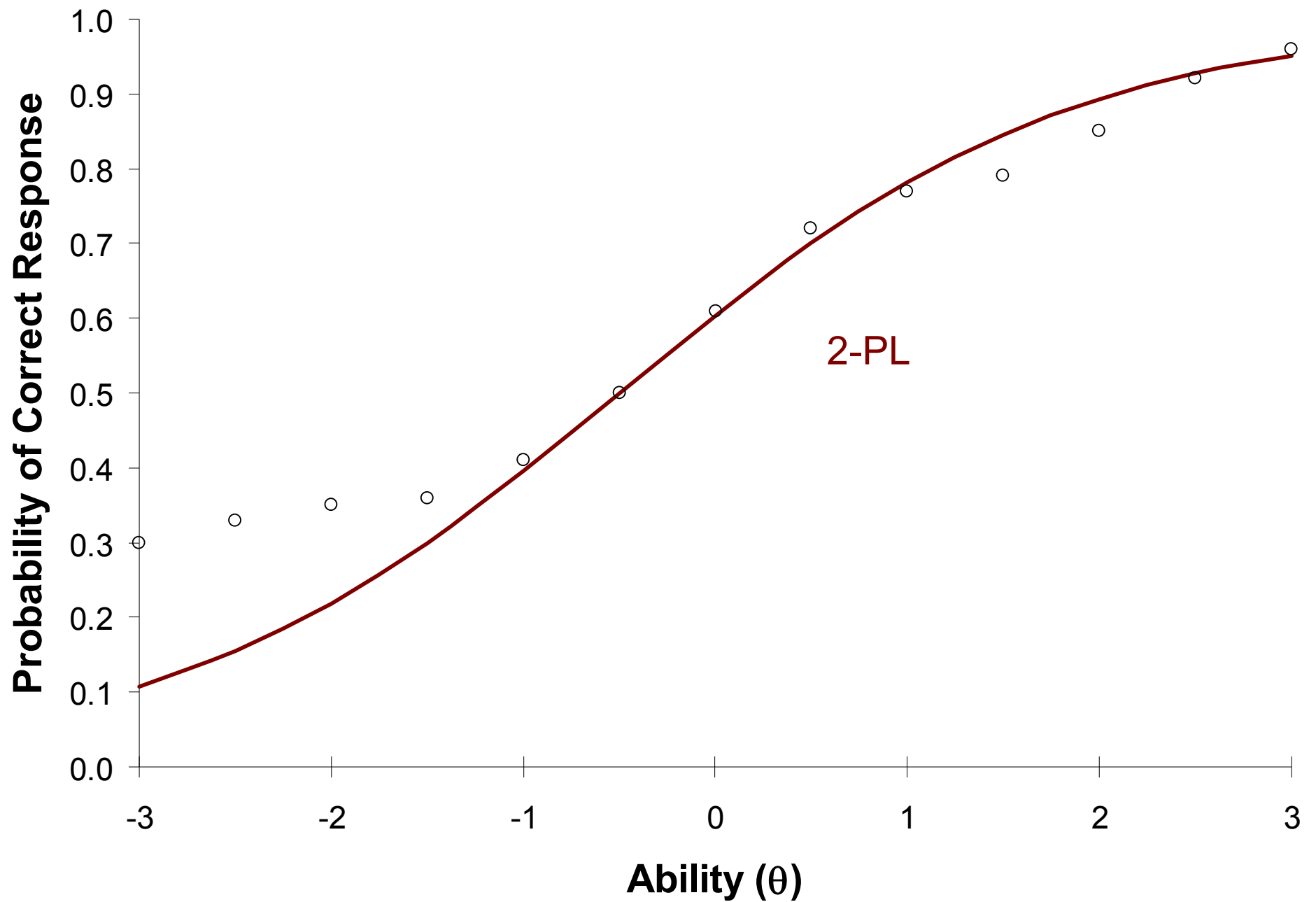
# Standardized Residuals

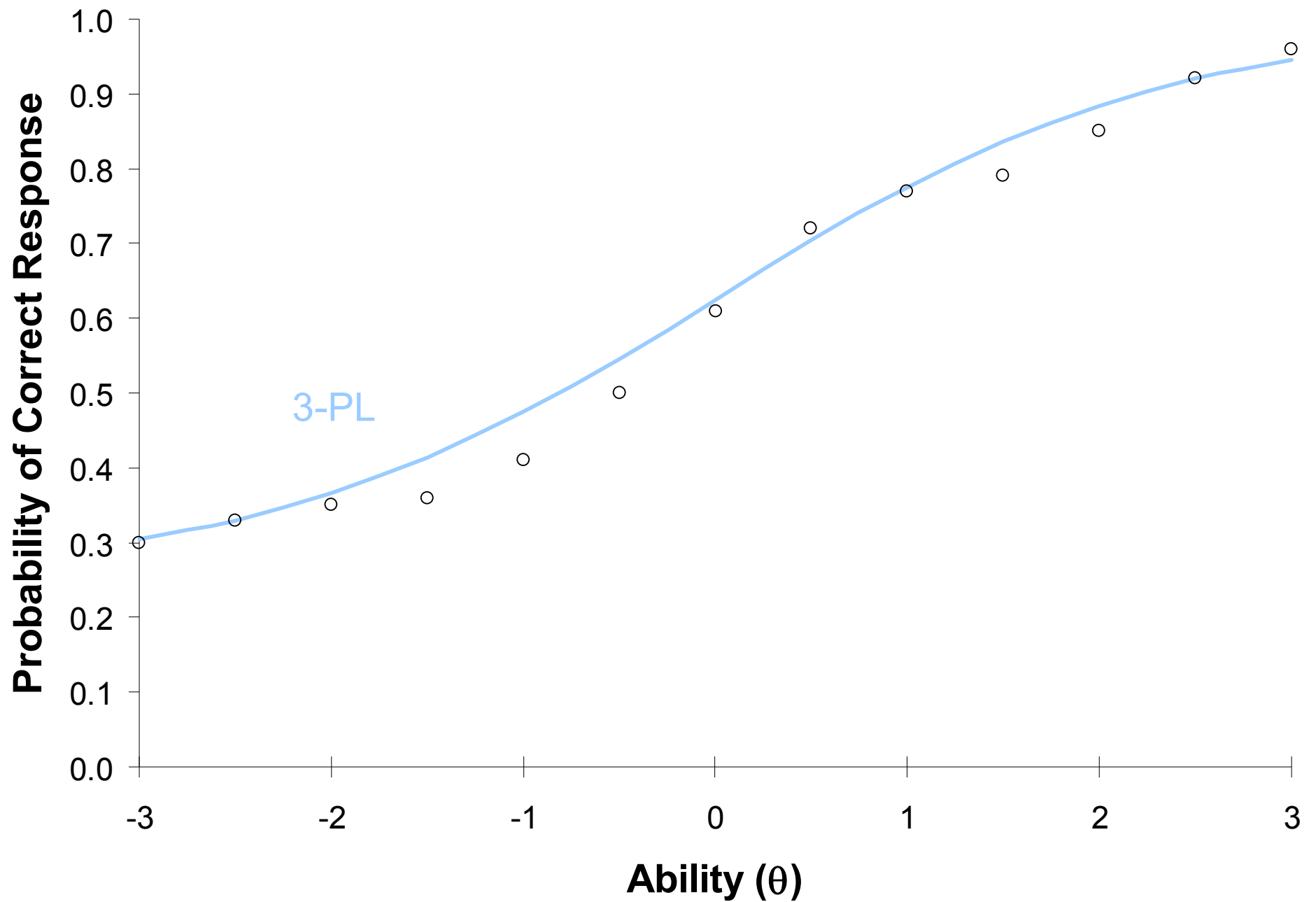
$$SR_i(\theta) = \frac{P_i(\theta) - E[P_i(\theta)]}{\sqrt{\frac{E[P_i(\theta)]E[1 - P_i(\theta)]}{N}}}$$

SR values should be homoscedastic for each item and follow an approximately standard normal distribution across all items of the test.

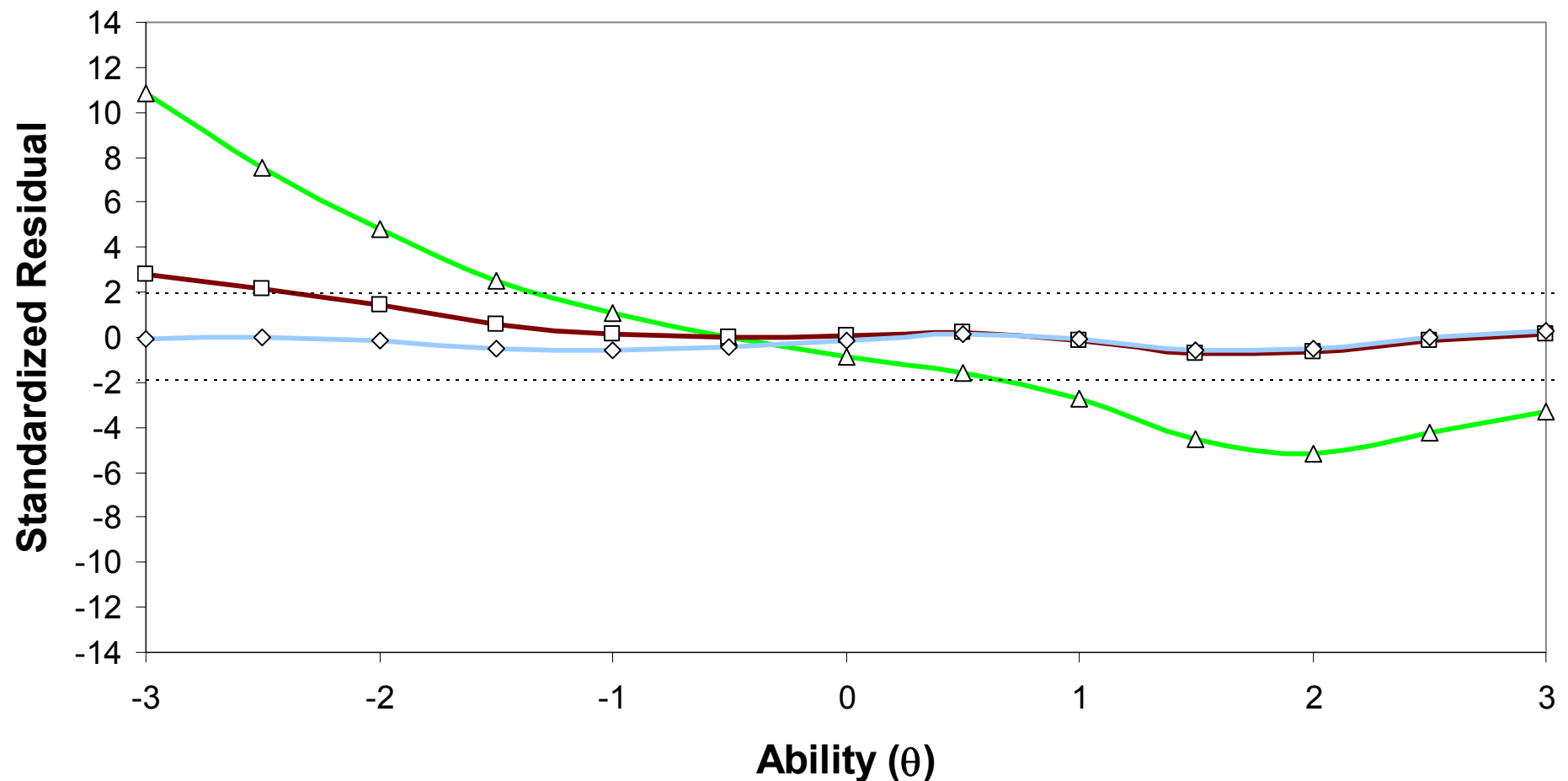








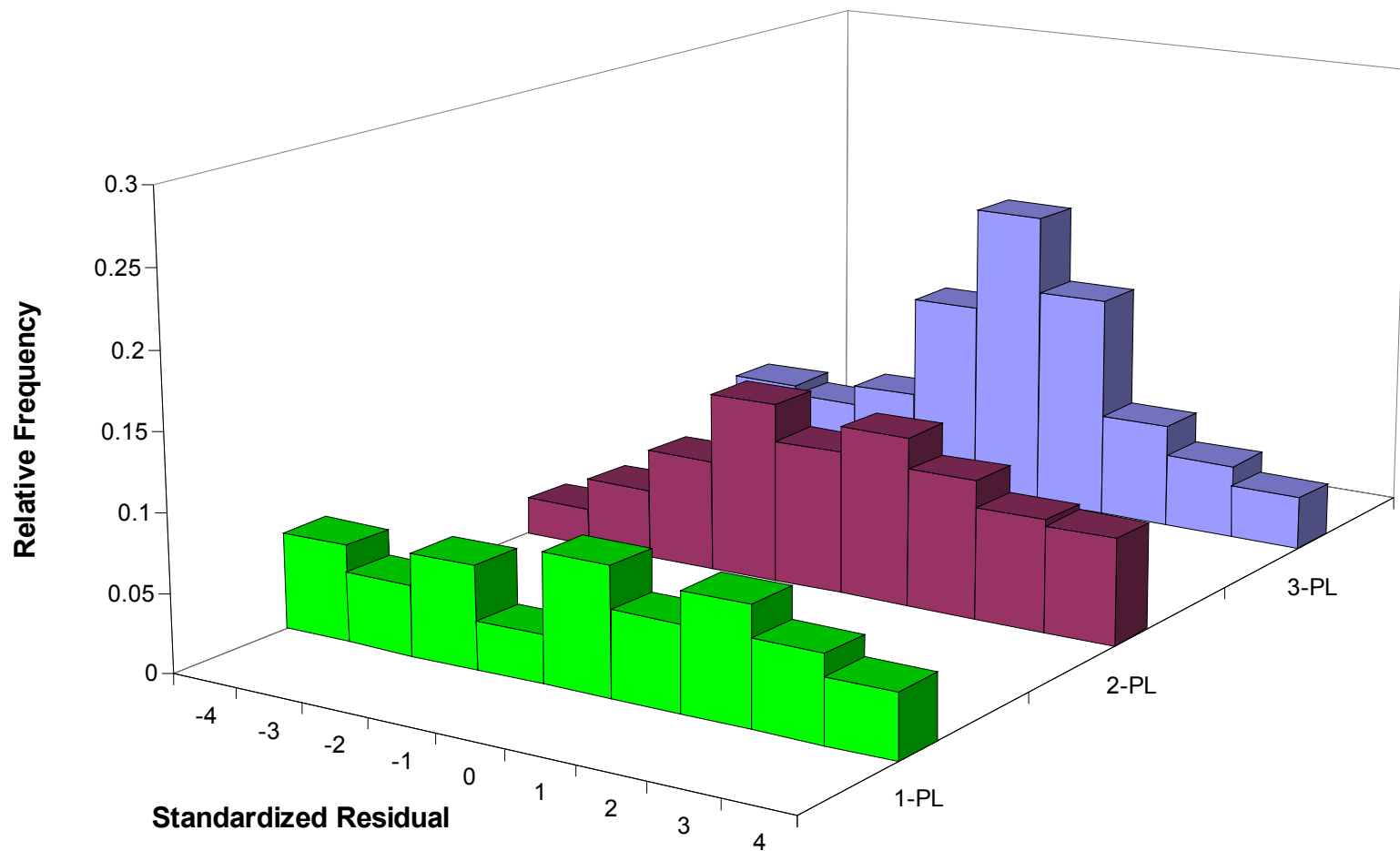
SR values are homoscedastic for this item when fit by a 3-PL model, but systematic errors are present for the 1- and 2-PL models



# Test-level Fit

- Similar to the comparison done for individual items, but instead we compare Expected Proportion Correct (TCC) to observed proportion correct (raw score/ $N$ )
- SRs across the test should be homoscedastic and follow an approximate normal distribution

Across all items, SR values are approximately normally distributed when fit by a 3-PL model, but more uniform for the 1- and 2-PL models



# Significance Testing

## Q1 chi-square (Yen, 1981)

$$Q1_j = \sum_{i=1}^m SR_{ij}^2 \quad Q1 \sim \chi^2 \quad df = m - p$$

$m$  = # of quadrature points

$p$  = # of item parameters

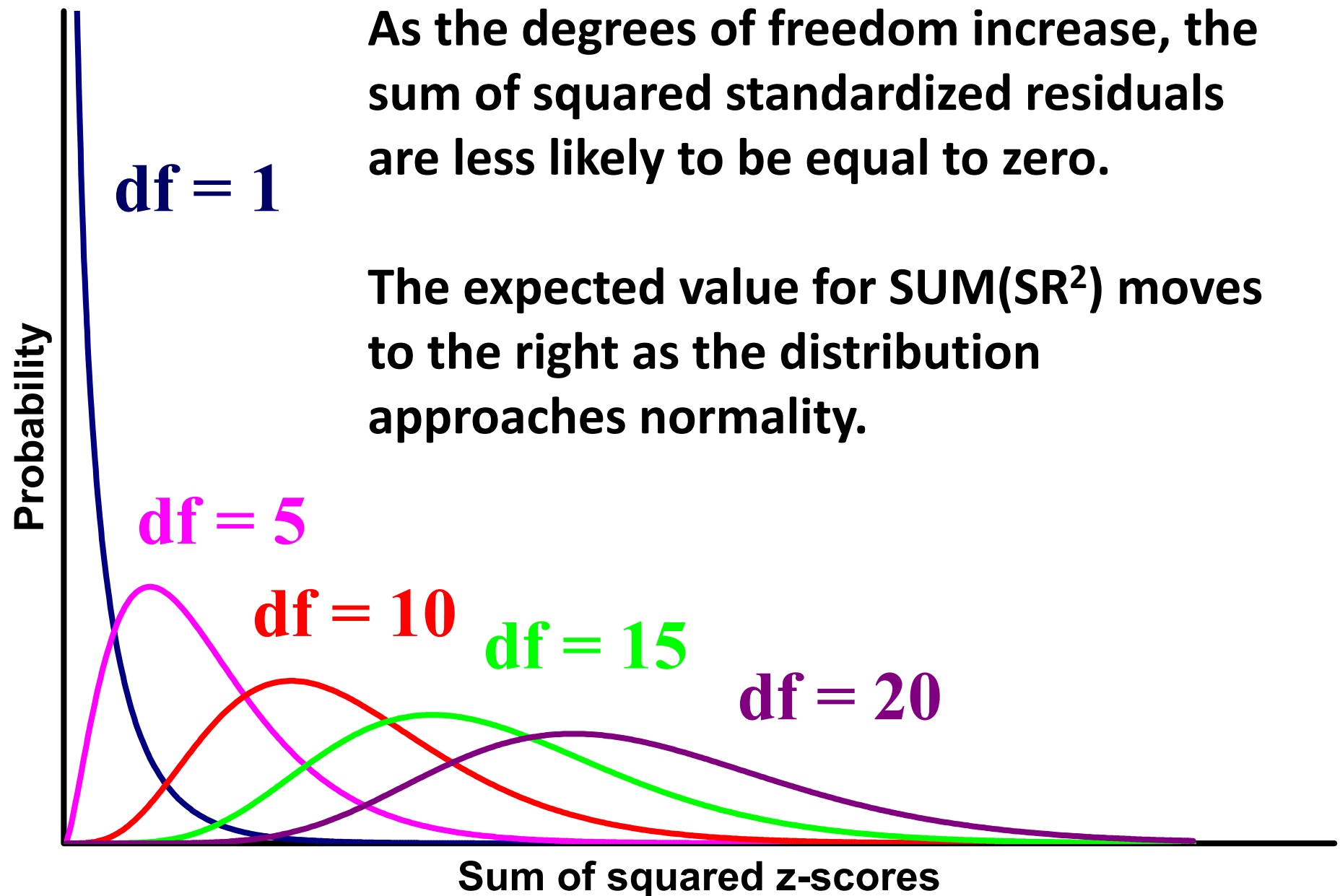
# Chi-square test

- Standardized residuals are essentially prediction errors that have been turned into z-scores
- The sum of squared z-scores follow a Chi-square distribution
  - Much like Sums of Squares and variances follow a Chi-square distribution in ANOVA



# Standard Normal Distribution





# Goodness-of-Fit

- This is what we call “goodness-of-fit”:
- We hope that the chi-square test will NOT be significant
  - This indicates that the differences between observed and expected is small
  - Significant differences would mean that observed proportions are far from what the model predicted

# Significance Testing in BILOG and PARSCALE

- The goodness of fit information contained in BILOG and PARSCALE use the Chi-square test described in the previous slides

➤ These values can be found for all items

SUBTEST PRETEST ; ITEM PARAMETERS AFTER CYCLE 30							
ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
MATH01	1.041 0.107*	0.651 0.082*	-1.599 0.242*	0.545 0.069*	0.186 0.084*	29.0 (0.0007)	9.0
MATH02	2.230 0.165*	0.600 0.114*	-3.717 0.610*	0.514 0.098*	0.199 0.089*	9.5 (0.0920)	5.0
MATH03	0.428 0.106*	0.693 0.084*	-0.618 0.190*	0.569 0.069*	0.159 0.071*	63.2 (0.0000)	7.0
MATH04	-0.601 0.216*	1.391 0.268*	0.432 0.095*	0.812 0.156*	0.217 0.040*	10.7 (0.2204)	8.0

# The problem of sample size

- Statistical tests of model-data fit present an interesting duality:
- Due to sensitivity to sample size, almost any departure of data from the model results in rejecting  $H_0$
- For small samples, model-data misfit can be overlooked, and SEs for item parameters are large

# CONCLUDING REMARKS

# Concluding Remarks: Model Fit

- Assessment of model fit in IRT Models is currently a difficult task
  - Easily accessible options are limited
  - Can quickly find options that take longer to assess fit than to estimate model
  - Mplus options are adequate for initial screening
- IRT models share this problem general categorical data analysis techniques
- Other model fit options are available and forthcoming