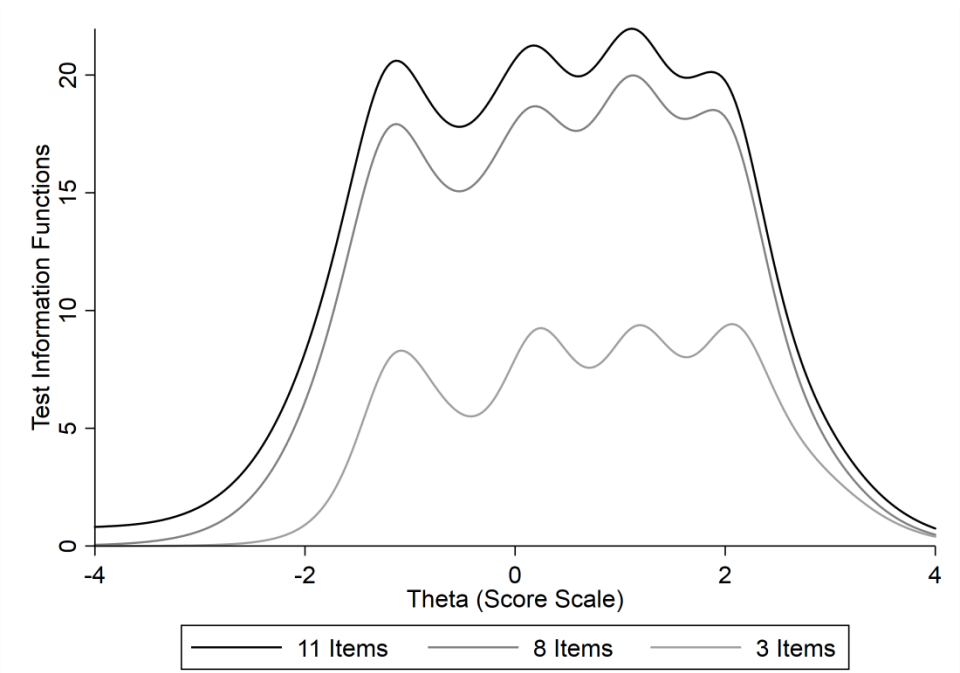
HARVARD
GRADUATE SCHOOL OF EDUCATION

Statistical and psychometric methods for measurement: Scale development and validation



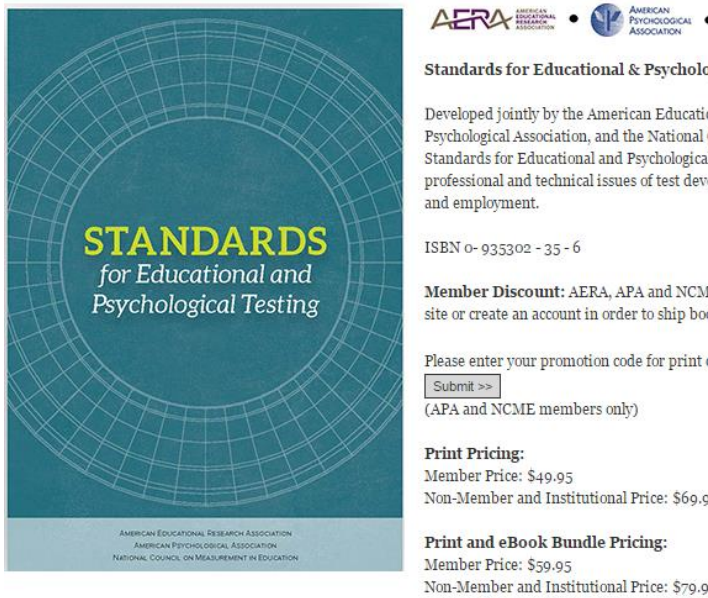
Andrew Ho, *Harvard Graduate School of Education*

The World Bank, Psychometrics Mini Course

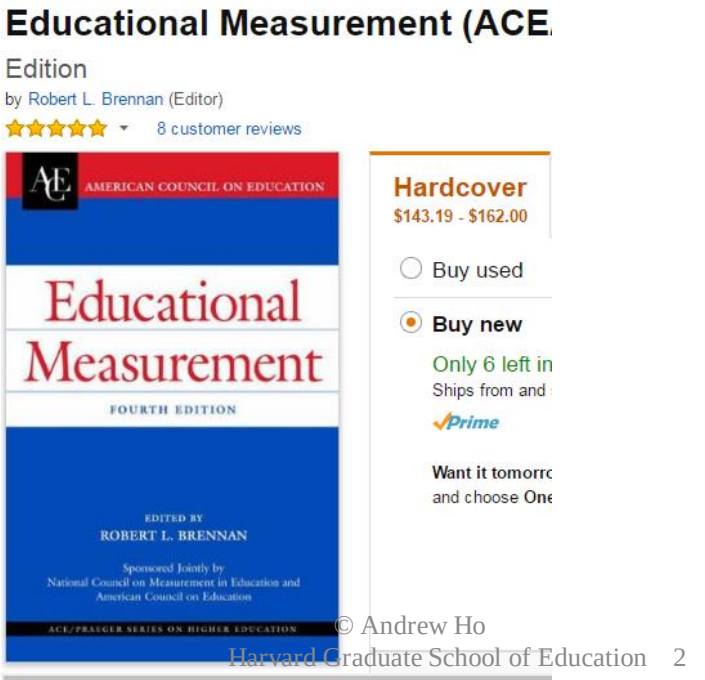
Washington, DC. June 11, 2018

Essential References

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.



2. Brennan, R. L. (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education, Praeger Publishers.



3. Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata, Volumes I and II* (3rd ed.). College Station, TX: Stata Press

Learning Objectives

- How do we “develop and validate” a scale?
  - What is validation?
  - What is reliability?
  - What is factor analysis?
  - What is Item Response Theory?
  - How do we do all this in Stata, and interpret the output accurately?

Some motivating examples

Development and Validation of the Short Grit Scale (Grit–S)

ANGELA LEE DUCKWORTH AND PATRICK D. QUINN

Department of Psychology, University of Pennsylvania

Journal of Personality Assessment, *91*(2), 166–174, 2009

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223890802634290

In this article, we introduce brief self-report and informant-report versions of the Grit Scale, which measures trait-level perseverance and passion for long-term goals. The Short Grit Scale (Grit–S) retains the 2-factor structure of the original Grit Scale (Duckworth, Peterson, Matthews, & Kelly, 2007) with 4 fewer items and improved psychometric properties. We present evidence for the Grit–S's internal consistency, test–retest stability, consensual validity with informant-report versions, and predictive validity. Among adults, the Grit–S was associated with educational attainment and fewer career changes. Among adolescents, the Grit–S longitudinally predicted GPA and, inversely, hours watching television. Among cadets at the United States Military Academy, West Point, the Grit–S predicted retention. Among Scripps National Spelling Bee competitors, the Grit–S predicted final round attained, a relationship mediated by lifetime spelling practice.

Improving Measurement Efficiency of the Inner EAR Scale with Item Response Theory

Otolaryngology–
Head and Neck Surgery
2018, Vol. 158(6) 1093–1100

DOI: 10.1177/0194599818760528
http://otojournal.org
SAGE

Annika Jessen, EdM[1], Andrew D. Ho, PhD[1], C. Eduardo Corrales, MD, PhD[2], Bevan Yueh, MD, MPH[3], and Jennifer J. Shin, MD, SM[2]

How do we "develop and validate" a measure?

We say, "*$\boldsymbol{X}$ is important. No one thinks of $\boldsymbol{X}$. Existing measures of $\boldsymbol{X}$ are off the mark. $\boldsymbol{X}$ matters more than everything else. If only we paid attention to $\boldsymbol{X}$.*" If that is your argument, I suggest this research agenda:

1. Establish the **theoretical construct**
   - This measure should exist.
2. Establish the **latent structure**
   - This components of the measure relate as expected.
3. Establish **reliability**
   - The score you estimate should be precise.
4. Establish **predictions** and **intercorrelations**
   - These scores predict outcomes it should. They also predict outcomes things better than, over and above, other scores.
5. Establish **usefulness**
   - Using these scores achieves the intended purposes.

What is validation? My "5 Cs"

1. **Content**
   - Evidence based on tested content, the measured construct.
     - e.g., Alignment studies, theoretical development
2. **Cognition**
   - Evidence based on response processes
     - e.g., Think-aloud protocols
3. **Coherence**
   - Evidence based on internal structure
     - e.g., Reliability analyses
4. **Correlation**
   - Evidence based on relations to other variables
     - e.g., Convergent evidence
5. **Consequence**
   - Evidence based on consequences of testing
     - e.g., Long-term evaluations

Common uses of "validity", graded:

- "The measure is valid." Grade: C-
- "…is a valid and reliable measure" C-
- "…is a validated measure" C
- "X has validated this measure." C+
- "Has a high validity coefficient" D
- "Validate the score" B
- "Validate the interpretation of the score as…" B+
- "I provide validity evidence for the interpretation of the score as..." A-
- "I provide validity evidence for the use of the score as…" A

Validation (Kane, 2006; 2013)

Validating the Interpretations and Uses of Test Scores

Michael T. Kane
Educational Testing Service

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. In outlining the argument-based approach to validation, this paper makes eight general points. First, it is the proposed score interpretations and uses that are validated and not the test or the test scores. Second, the validity of a proposed interpretation or use depends on how well the evidence supports the claims being made. Third, more-ambitious claims require more support than less-ambitious claims. Fourth, more-ambitious claims (e.g., construct interpretations) tend to be more useful than less-ambitious claims, but they are also harder to validate. Fifth, interpretations and uses can change over time in response to new needs and new understandings leading to changes in the evidence needed for validation. Sixth, the evaluation of score uses requires an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable. Seventh, the rejection of a score use does not necessarily invalidate a prior, underlying score interpretation. Eighth, the validation of the score interpretation on which a score use is based does not validate the score use.

An 8-Step Plan

Step 1: Content and Cognition

Step 2: Scoring and Scaling

Step 3: Correlation and Reliability

Step 4: Classical Item Diagnostics

Step 5: Latent Structure Analysis

Step 6: Item Response Theory (IRT)

Step 7: IRT for Efficient Measurement

Step 8: Correlation and Prediction

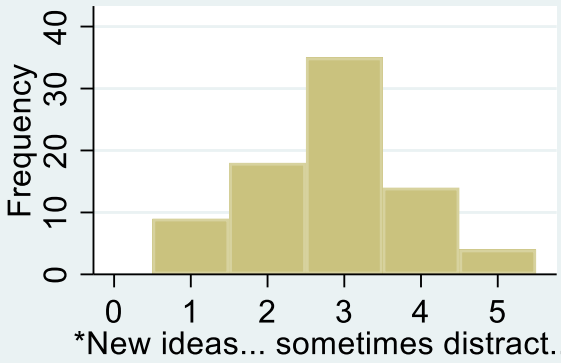
Step 1: Content and Cognition (RTFQ)

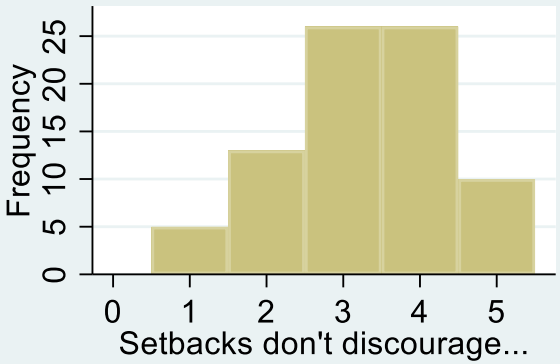
	Very much like me	Mostly like me	Somewhat like me	Not much like me	Not like me at all
1) New ideas and projects sometimes distract me from previous ones.	○				
2) Setbacks don't discourage me.	○				
3) I have been obsessed with a certain idea or project for a short time but later lost interest.	○				
4) I am a hard worker.	○				
5) I often set a goal but later choose to pursue a different one.			○	○	○
6) I have difficulty maintaining my focus on projects that take more than a few months to complete.			○	○	○
7) I finish whatever I begin.			○	○	○
8) I am diligent.			○	○	○

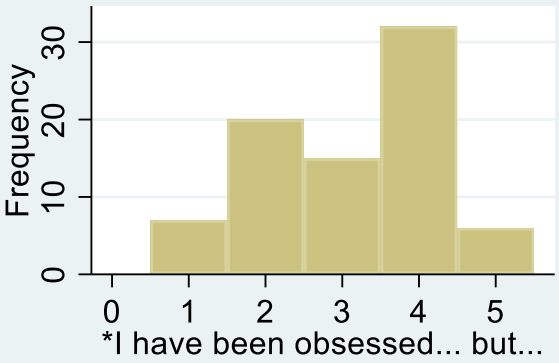
Step 2: Establishing Scoring and Scaling Rules

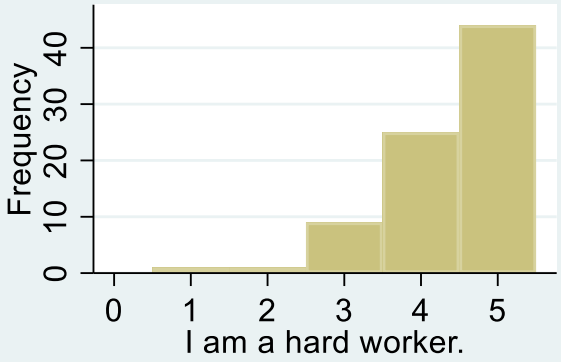
```
*-----------------------------------------------------------------------
* Encode data
*-----------------------------------------------------------------------

forvalues varnum = 1/8 {
    gen score`varnum' = 0
    replace score`varnum' = 1 if x`varnum' == "Not like me at all"
    replace score`varnum' = 2 if x`varnum' == "Not much like me"
    replace score`varnum' = 3 if x`varnum' == "Somewhat like me"
    replace score`varnum' = 4 if x`varnum' == "Mostly like me"
    replace score`varnum' = 5 if x`varnum' == "Very much like me"
}

* Reverse polarity for items 1, 3, 5, and 6
foreach var of varlist score1 score3 score5 score6 {
    replace `var' = 6-`var'
}

label variable score1 "New ideas... sometimes distract..."
label variable score2 "Setbacks don't discourage..."
label variable score3 "I have been obsessed... but..."
label variable score4 "I am a hard worker."
label variable score5 "I often set a goal but later..."
label variable score6 "I have difficulty maintaining..."
label variable score7 "I finish whatever I begin."
label variable score8 "I am diligent."
```

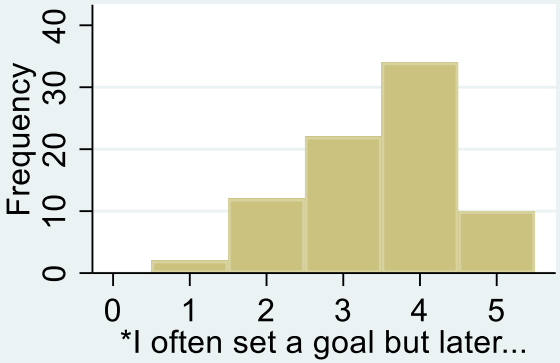
What are the distributions of item scores for grit?

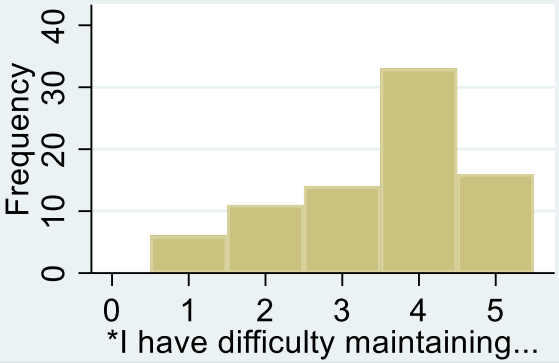


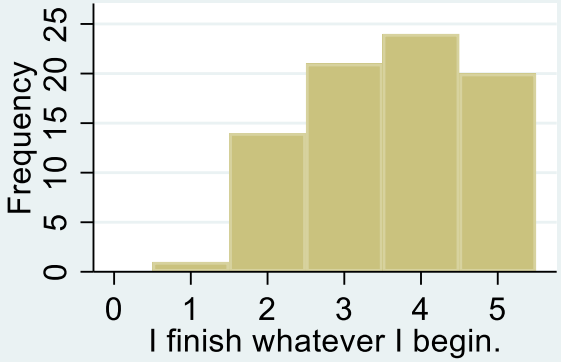


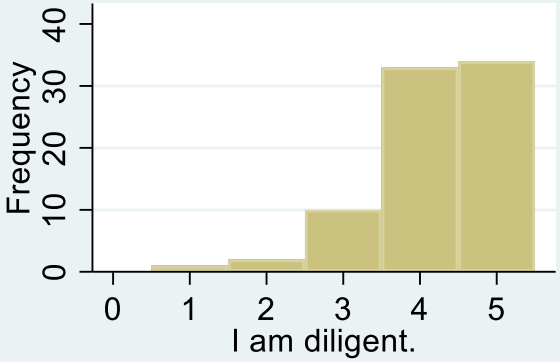


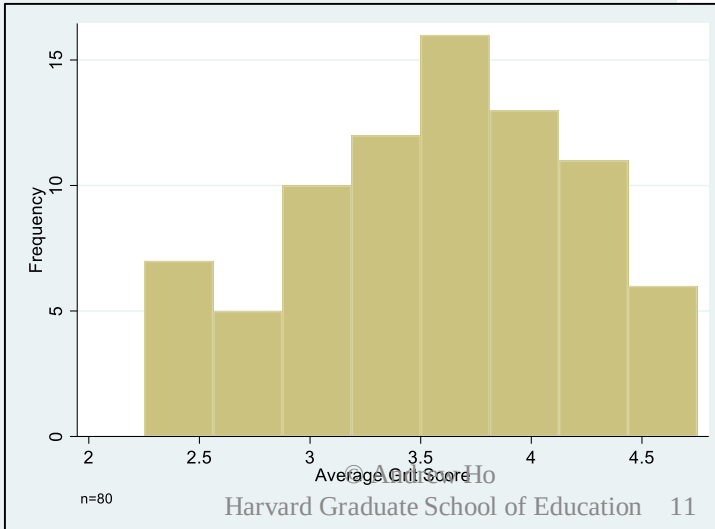












n=80

Step 1: Content and Cognition (RTFQ)

Inner EAR

Instructions: The purpose of this questionnaire is to identify the problems your hearing loss may be causing you. **Please answer all the questions based on your experiences over the past 2 weeks.** If you use a hearing aid, please answer the way you hear based on normal use of your hearing aid. You may leave any question unanswered.

1. Over the past 2 weeks, how would you rate your ability to hear?

Circle one number 0-10

0	1	2	3	4	5	6	7	8	9	10
Hate it					It's OK					Love it

2. Over the past 2 weeks, how would you rate your ability to understand:	Poor	Fair	Good	Very Good	Excellent
what family and close friends are saying?	1	2	3	4	5
speech in a quiet room?	1	2	3	4	5
what is said by those at your table in a crowded restaurant?	1	2	3	4	5
what you want to hear, and filter out unwanted noises?	1	2	3	4	5
telephone conversations?	1	2	3	4	5

3. Over the past 2 weeks, how would you rate your ability to hear:					
in different listening situations?	1	2	3	4	5
soft household sounds (car turn signal, clock ticking)?	1	2	3	4	5

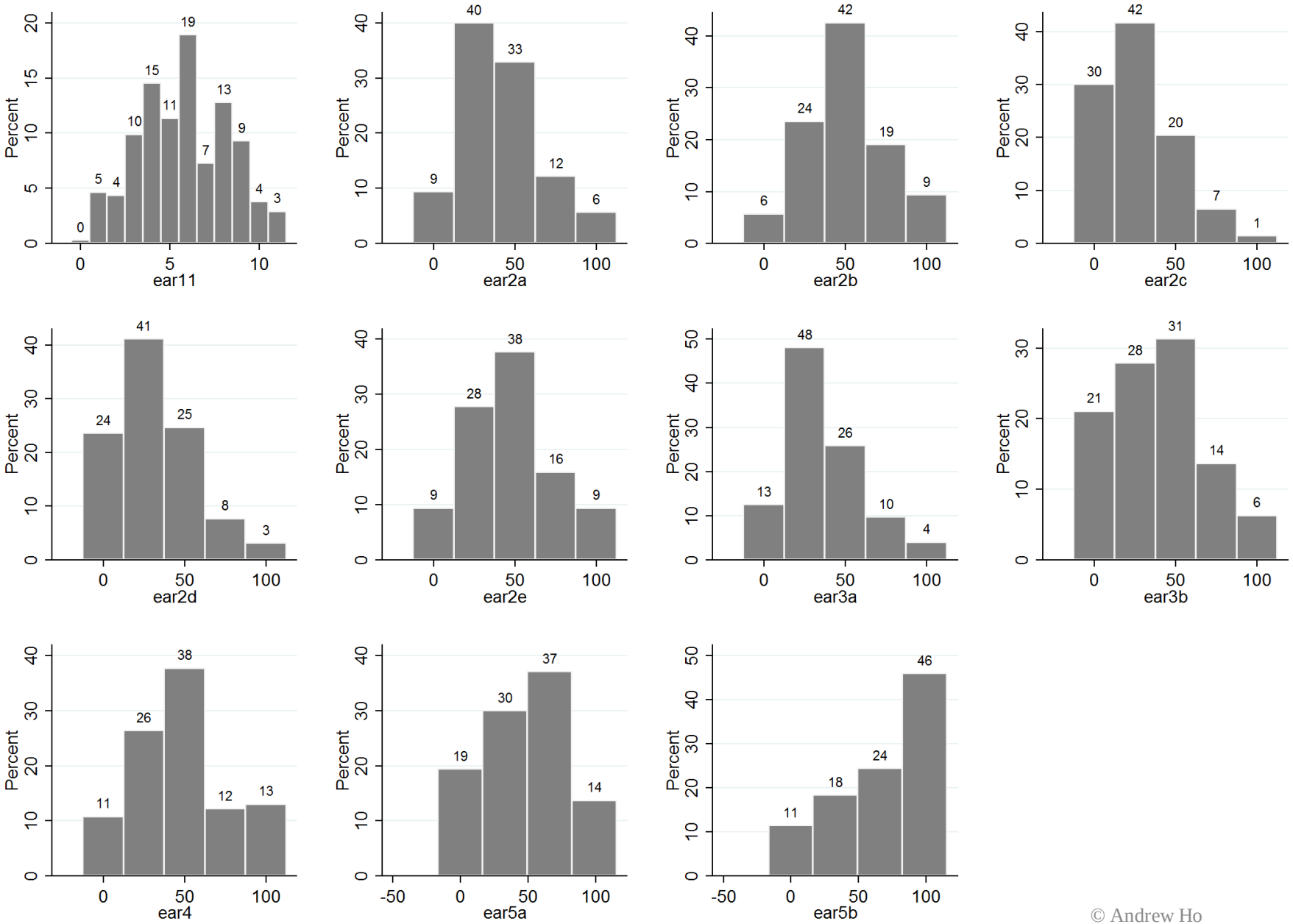
4. Over the past 2 weeks, how would you rate your mood based on your ability to hear?	1	2	3	4	5

5. Over the past 2 weeks, how bothered are you:	Very Bothered	Bothered	A little Bothered	Not Bothered
by having to ask people to repeat things?	1	2	3	4

Step 2: Establishing Scoring and Scaling Rules

```
*-----------------------------------------------------------------------------
* Loading dataset, Rescale variables to the same scale in new variables
*-----------------------------------------------------------------------------

   import excel "Raw Data\Export for AH AJ 11232016.xls", sheet("Sheet1") firstrow clear

   rename innerearl1 earl1

   recode innerearrarray234innerear2a (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2a)
   recode innerearrarray234innerear2b (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2b)
   recode innerearrarray234innerear2c (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2c)
   recode innerearrarray234innerear2d (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2d)
   recode innerearrarray234innerear2e (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2e)
   recode innerearrarray234innerear3a (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear3a)
   recode innerearrarray234innerear3b (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear3b)
   recode innerearrarray234innerear4 (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear4)
   recode innerearrarray5innerear5a (1=0) (2=33) (3=67) (4=100), gen(ear5a)
   recode innerearrarray5innerear5b (1=0) (2=33) (3=67) (4=100), gen(ear5b)
 *-----------------------------------------------------------------------------
 * Rescale variables to the same scale (just earl1)
 *-----------------------------------------------------------------------------

     gen earl1a = .
     replace earl1a = 0 if earl1 == 0
     replace earl1a = 10 if earl1 == 1
     replace earl1a = 20 if earl1 == 2
     replace earl1a = 30 if earl1 == 3
     replace earl1a = 40 if earl1 == 4
     replace earl1a = 50 if earl1 == 5
     replace earl1a = 60 if earl1 == 6
     replace earl1a = 70 if earl1 == 7
     replace earl1a = 80 if earl1 == 8
     replace earl1a = 90 if earl1 == 9
     replace earl1a = 100 if earl1 == 10
```

What are the distributions of item scores for Inner Ear?

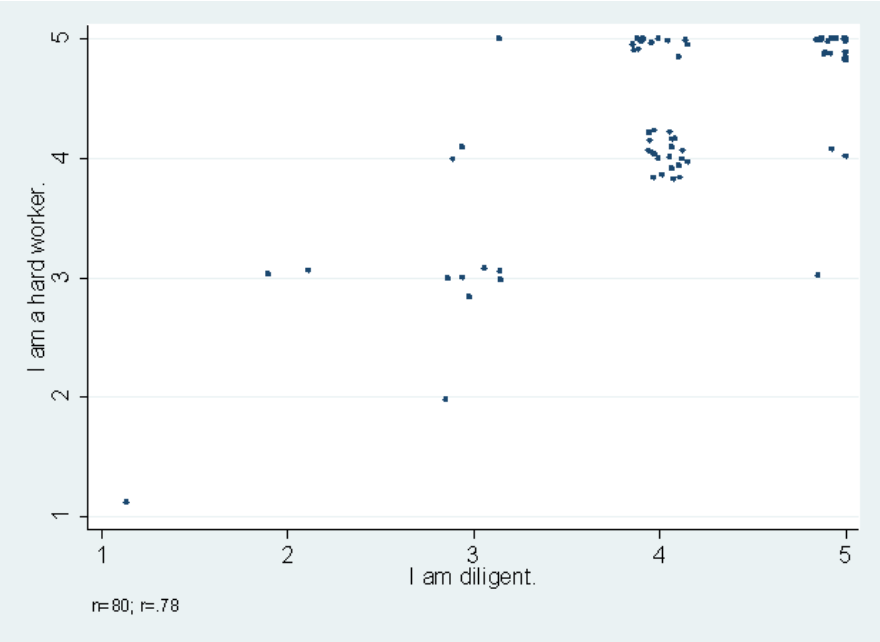


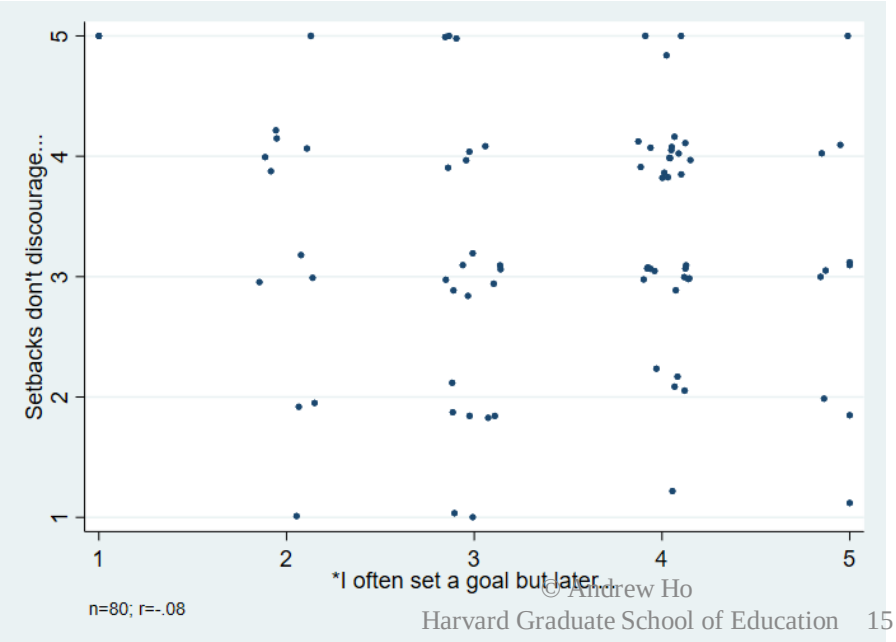
Step 3: Correlations and Cronbach's alpha

```
pwcorr score1-score8, star(.05)
```

	score1	score2	score3	score4	score5	score6	score7
score1	1.0000						
score2	-0.0804	1.0000					
score3	0.3359*	0.1244	1.0000				
score4	0.0037	0.2302*	0.1233	1.0000			
score5	0.4149*	-0.0827	0.3990*	0.1977	1.0000		
score6	0.4684*	-0.0404	0.4775*	0.2997*	0.4930*	1.0000	
score7	0.3028*	0.2392*	0.3270*	0.5040*	0.4303*	0.2845*	1.0000
score8	0.1604	0.2625*	0.1677	0.7783*	0.3472*	0.3283*	0.5716*

```
score1 "New ideas... sometimes distract..."
score2 "Setbacks don't discourage..."
score3 "I have been obsessed... but..."
score4 "I am a hard worker."
score5 "I often set a goal but later..."
score6 "I have difficulty maintaining..."
score7 "I finish whatever I begin."
score8 "I am diligent."
```





Step 3: Correlations and Cronbach's alpha

Table 1. Item Means (SD) and Interitem Correlations (n = 334-353).

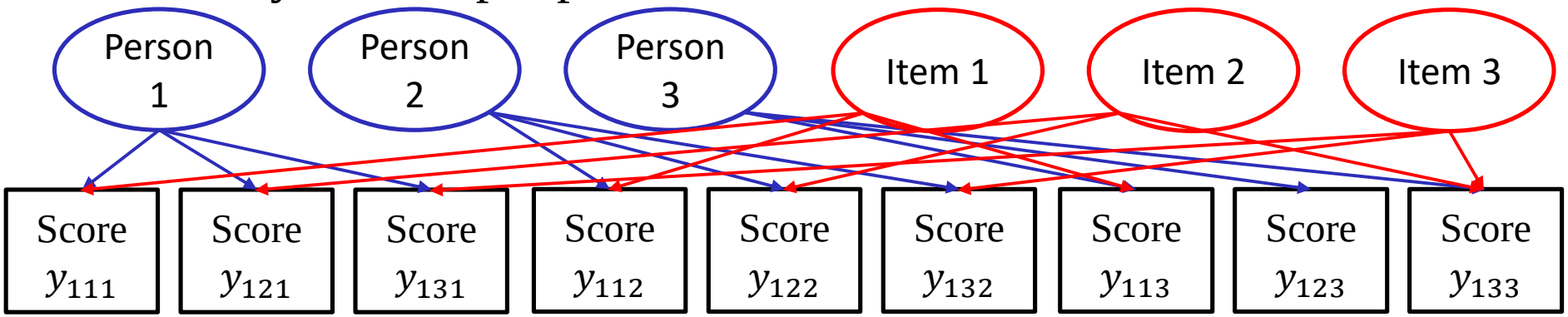
	Item										
	ear11	ear2a	ear2b	ear2c	ear2d	ear2e	ear3a	ear3b	ear4	ear5a	ear5b
Mean (0-100)	56.23	41.22	50.71	26.91	31.44	47.03	36.15	39.06	47.59	48.36	68.21
SD (0-100)	23.700	25.020	25.340	23.570	25.050	27.080	24.130	28.700	28.850	31.940	35.040
ear11: Ability to hear	1.00										
ear2: Ability to understand . . .											
Family and close friends (ear2a)	0.62	1.00									
Speech in a quiet room (ear2b)	0.49	0.74	1.00								
Conversation in a crowded restaurant (ear2c)	0.49	0.69	0.59	1.00							
Filter out unwanted noises (ear2d)	0.48	0.72	0.59	0.80	1.00						
Telephone conversations (ear2e)	0.41	0.63	0.61	0.51	0.60	1.00					
ear3: Ability to hear . . .											
In different listening situations (ear3a)	0.54	0.76	0.70	0.72	0.74	0.68	1.00				
Soft household sounds (ear3b)	0.47	0.65	0.68	0.56	0.61	0.72	0.70	1.00			
ear4: Mood based on your ability to hear	0.45	0.59	0.53	0.52	0.59	0.62	0.67	0.64	1.00		
ear5: Bother from . . .											
Need for repetition (ear5a)	0.33	0.52	0.39	0.52	0.52	0.42	0.54	0.48	0.58	1.00	
Restricted activities because of hearing (ear5b)	0.42	0.47	0.42	0.44	0.44	0.39	0.50	0.46	0.59	0.65	1.00

Reliability: Measurement as a random crossed effects model

- A response i to item j by person k

$$y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk};$$
$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$

- Note: Only 1 score per person/item combination

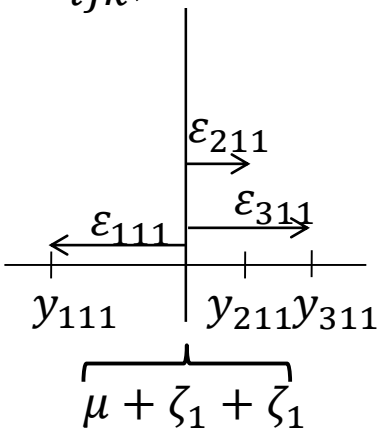


- μ – Overall average score
- ζ_j – Item location (easiness), ψ_1 –variance of item effects
- ζ_k –Person location (proficiency), ψ_2 –variance of person effects
- ε_{ijk} –Person-item interactions and other effects, θ –error variance

Reliability: What are two relevant intraclass correlations?

- A response i to item j by person k : $y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk}$;

$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$



- μ – Overall average score
- ζ_j – Item location (easiness). Variance: ψ_1
- ζ_k –Person location (proficiency). Variance: ψ_2
- ε_{ijk} –Person-item interactions and other effects. Variance: θ
- Intraclass correlation: $\rho = \frac{\psi_2}{\psi_2 + \theta}$. The correlation between two <u>item responses</u> within persons. The proportion of *relative* <u>response</u> variation due to persons.
- Intraclass correlation: $\rho_\alpha = \frac{\psi_2}{\psi_2 + \frac{\theta}{n_j}}$. Cronbach's alpha: The correlation between two average (or sum) <u>scores</u> within persons. The proportion of *relative* <u>score</u> variance due to persons.

Estimation in Stata

- A response i to item j by person k : $y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk}$;

$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$

```
. mixed response || _all: R.person || _all: R.item, variance reml
```

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
_all: Identity				
var(R.person)	416.449	33.72802	355.3231	488.0904
_all: Identity				
var(R.item)	138.3177	62.28706	57.22236	334.3411
var(Residual)	334.8282	8.012387	319.4868	350.9063

LR test vs. linear model: chi2(2) = 2687.38 Prob > chi2 = 0.0000

- Relevant intraclass correlation:

$$\hat{\rho}_\alpha = \frac{\hat{\psi}_{person}}{\hat{\psi}_{person} + \frac{\hat{\theta}}{11}} = \frac{416.45}{416.45 + \frac{334.83}{11}} = .93$$

Cronbach's alpha directly, in Stata

- A response i to item j by person k : $y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk}$;

$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$

- Classical computational formula for Cronbach's alpha:

$$\hat{\rho}_\alpha = \frac{n_j}{n_j - 1}\left(1 - \frac{\sum_j \sigma^2_{X_j}}{\sigma^2_X}\right),$$

- where $\sigma^2_{X_j}$ is the variance of each item score X_j, and σ^2_X is the variance of a total (summed) score, X.
- In Stata:

```
. alpha ear11a-ear5b, asis

Test scale = mean(unstandardized items)

Average interitem covariance:      410.7885
Number of items in the scale:            11
Scale reliability coefficient:       0.9304
```

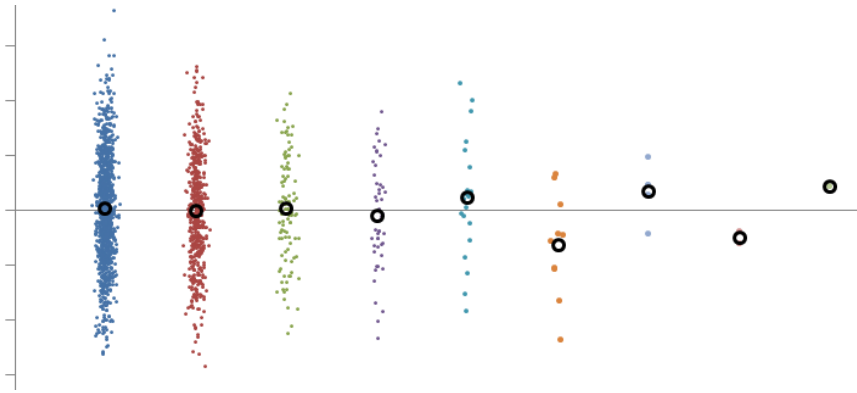
How should I think about reliability?

Three Necessary Intuitions

1. Any observed score is ***one of many possible replications***.

2. Any observed score is the ***sum of a "true score"*** (average of all theoretical replications) ***and an error term***.

3. Averaging over replications gives us ***better estimates*** of "true" scores by ***averaging over error terms***.



What is the reliability of grit scores?

```
.   alpha score1-score8, asis

Test scale = mean(unstandardized items)

Average interitem covariance:     .2955753
Number of items in the scale:            8
Scale reliability coefficient:      0.7565
```

Three Interpretations of Reliability

1. Reliability is the ***correlation between two sets of observed scores*** from a ***replication of a measurement procedure***.

2. Reliability is the ***proportion of "observed score variance"*** that is ***accounted for by "true score variance."***

3. Reliability that starts with ***an average of pairwise part covariances***, then increases this average as a function of the number of replications? Why? Because ***averaging over replications decreases error variance***.

$$\rho = \mathbf{E}\big(Corr(\bar{y}, \bar{y}')\big)$$

Reliability (ρ) is the expected value (long run average, **E**) of the correlations between average scores $\bar{y}$ and average scores of a replication $\bar{y}$.

$$\rho = \frac{\psi_2}{\psi_2 + \theta/n_j}$$

Reliability (ρ), "true" between-person variance (ψ) vs. observed score variance, $\sigma^2_{\bar{X}_i}$

$$\rho = \frac{n_j \overline{\rho_{jj'}}}{1 + (n_j - 1)\overline{\rho_{jj'}}}$$

Given some average pairwise part covariance, $\overline{\rho_{jj'}}$, the reliability of $\bar{X}_i$ is ρ.

Cronbach's α is a particular type of reliability, one of the most limited, but easy to estimate. Cronbach's α only considers correlations of scores (or variance) across replications of items.

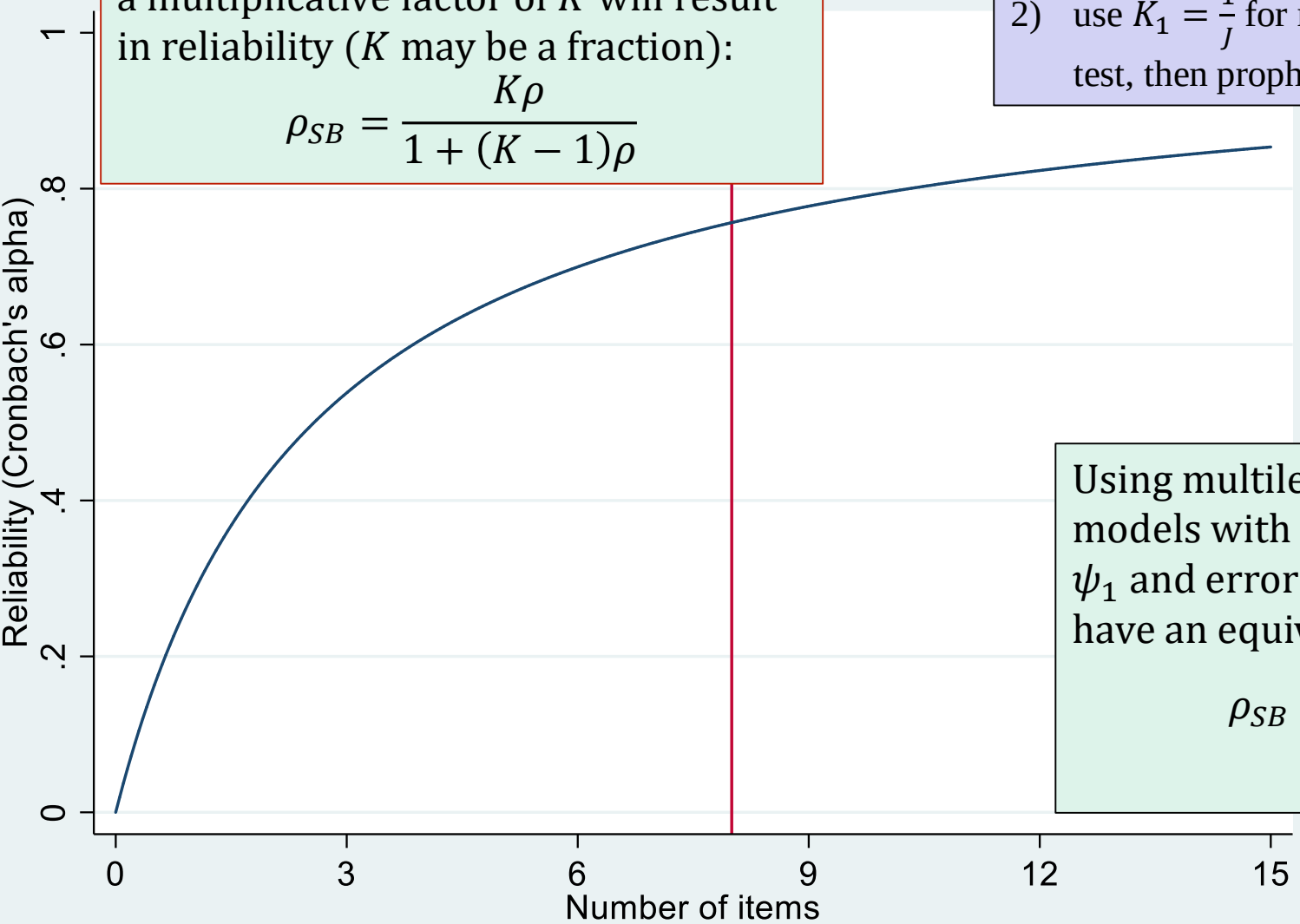
Spearman-Brown "Prophecy": How many items do I need for precision?

From some baseline reliability, ρ, Spearman-Brown "prophesizes" that increasing the replications (items?) by a multiplicative factor of K will result in reliability (K may be a fraction):

$$\rho_{SB} = \frac{K\rho}{1 + (K-1)\rho}$$

Note: Given ρ_α for a J-item test, and prophecy for a J'-item test, you can

1) calculate $K = \frac{J'}{J}$, or
2) use $K_1 = \frac{1}{J}$ for reliability of a 1-item test, then prophesize using $K_2 = J'$.



Using multilevel crossed effects models with person variance ψ_1 and error variance θ, we have an equivalent formula:

$$\rho_{SB} = \frac{\psi_2}{\psi_2 + \frac{\theta}{n_j}}$$

Step 4: Classical Item Diagnostics

```
. alpha ear11a-ear5b, asis item

Test scale = mean(unstandardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
ear11a	334	+	0.6614	0.5945	435.7858	0.9295
ear2a	353	+	0.8513	0.8168	406.6422	0.9198
ear2b	353	+	0.7775	0.7277	415.1858	0.9234
ear2c	353	+	0.7872	0.7430	419.3079	0.9232
ear2d	353	+	0.8164	0.7748	411.07	0.9216
ear2e	353	+	0.7683	0.7130	411.6634	0.9240
ear3a	352	+	0.8720	0.8430	407.1448	0.9191
ear3b	352	+	0.8139	0.7646	400.3318	0.9215
ear4	353	+	0.8009	0.7485	401.6557	0.9222
ear5a	351	+	0.7171	0.6388	407.004	0.9283
ear5b	349	+	0.7004	0.6099	402.6394	0.9310
Test scale					410.7885	0.9304

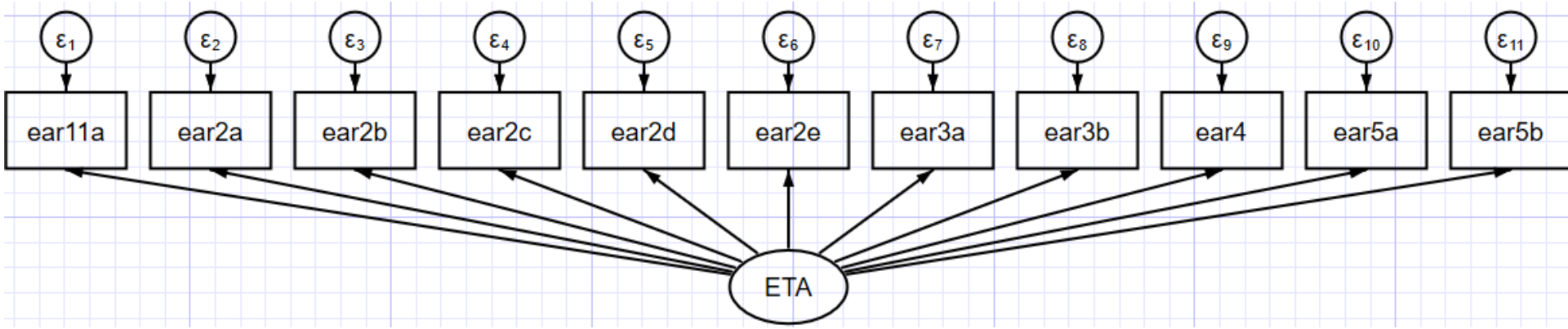
Step 4: Classical Item Diagnostics

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
ear11a	334	+	0.6614	0.5945	435.7858	0.9295
ear2a	353	+	0.8513	0.8168	406.6422	0.9198
ear2b	353	+	0.7775	0.7277	415.1858	0.9234
ear2c	353	+	0.7872	0.7430	419.3079	0.9232
ear2d	353	+	0.8164	0.7748	411.07	0.9216
ear2e	353	+	0.7683	0.7130	411.6634	0.9240
ear3a	352	+	0.8720	0.8430	407.1448	0.9191

These are ***diagnostics that explain item functioning*** and sometimes, with additional analysis, warrant item adaptation or exclusion. However, ***no item should be altered or excluded on the basis of these statistics alone.***

- ***Item-Test Correlation*** is a simple correlation between each item response and total test scores (the higher the better). This correlation is sometimes called "classical item discrimination." Think of it as "item information."
- ***Item-Rest Correlation*** is the similar, but the total test score excludes the target item (the higher the better). This avoids a part-whole confounding in correlation.
- ***Interitem Correlation*** shows the would-be interitem covariance if the item were excluded (the lower the better).
- ***Alpha (excluded-item alpha)*** shows the would-be ρ_α estimate if the item were excluded (the lower the better).

Step 5: Latent Structure Analysis – Principal Factor Analysis



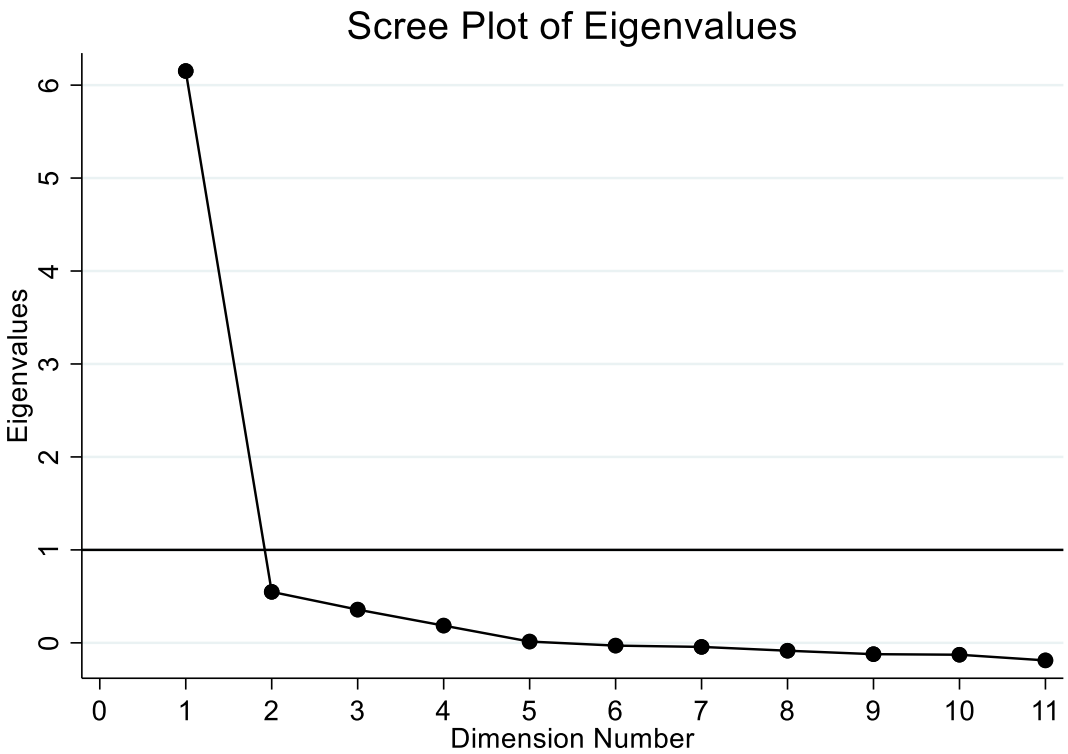
ear11: Ability to hear	1.00										
ear2: Ability to understand . . .											
Family and close friends (ear2a)	0.62	1.00									
Speech in a quiet room (ear2b)	0.49	0.74	1.00								
Conversation in a crowded restaurant (ear2c)	0.49	0.69	0.59	1.00							
Filter out unwanted noises (ear2d)	0.48	0.72	0.59	0.80	1.00						
Telephone conversations (ear2e)	0.41	0.63	0.61	0.51	0.60	1.00					
ear3: Ability to hear . . .											
In different listening situations (ear3a)	0.54	0.76	0.70	0.72	0.74	0.68	1.00				
Soft household sounds (ear3b)	0.47	0.65	0.68	0.56	0.61	0.72	0.70	1.00			
ear4: Mood based on your ability to hear	0.45	0.59	0.53	0.52	0.59	0.62	0.67	0.64	1.00		
ear5: Bother from . . .											
Need for repetition (ear5a)	0.33	0.52	0.39	0.52	0.52	0.42	0.54	0.48	0.58	1.00	
Restricted activities because of hearing (ear5b)	0.42	0.47	0.42	0.44	0.44	0.39	0.50	0.46	0.59	0.65	1.00

1) Replace diagonals with an estimate of reliability
2) Conduct a principal components analysis.

Step 5: Latent Structure Analysis – Principal Factor Analysis

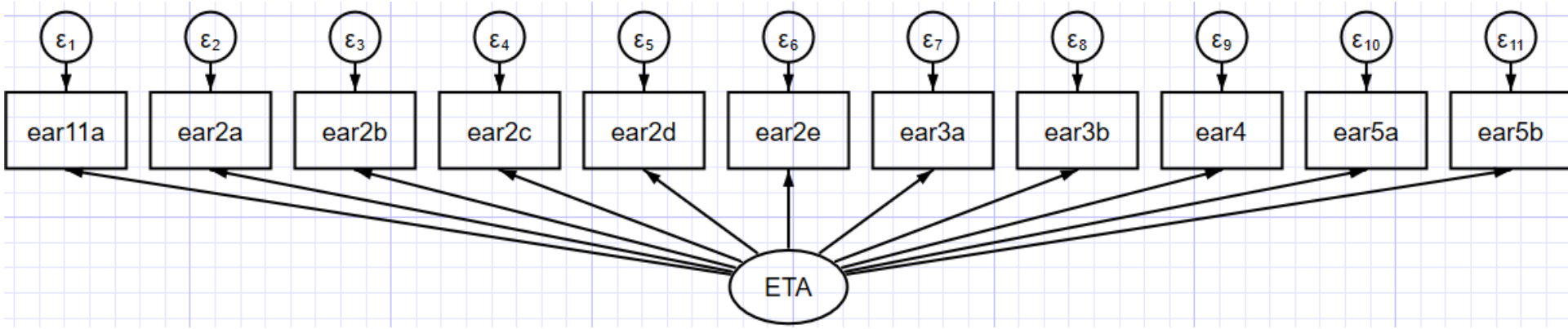
```
. factor ear11a-ear5b, factors(1)
(obs=329)

Factor analysis/correlation                      Number of obs    =        329
    Method: principal factors                    Retained factors =          1
    Rotation: (unrotated)                        Number of params =         11
```

Factor	Eigenvalue
Factor1	6.15209
Factor2	0.54844
Factor3	0.35753
Factor4	0.18585
Factor5	0.01411
Factor6	-0.02968
Factor7	-0.04374
Factor8	-0.08464
Factor9	-0.12194
Factor10	-0.12772
Factor11	-0.18856



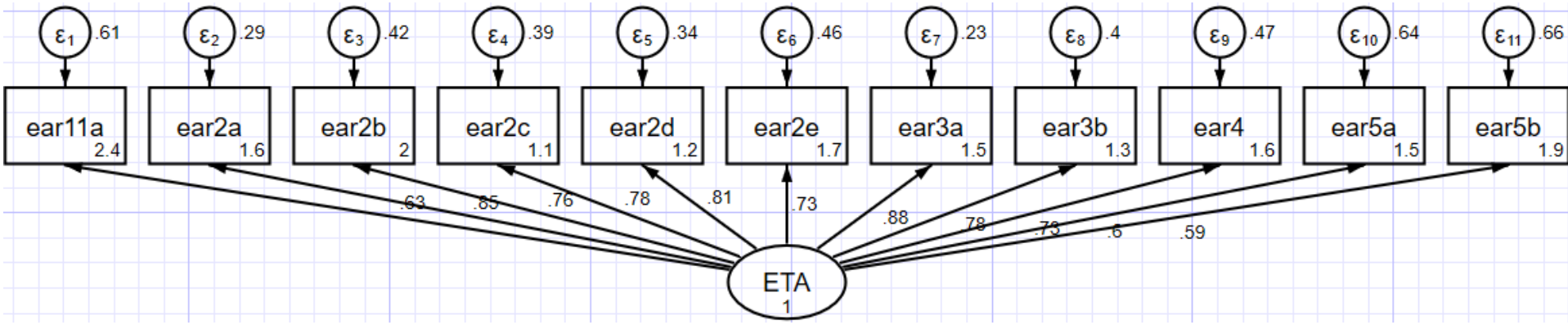

```
LR test: independent vs. saturated:  chi2(55) = 2464.79 Prob>chi2 = 0.0000
```

Step 5: Latent Structure Analysis – Principal Factor Analysis



Variable	Factor1	Uniqueness
ear11a	0.6229	0.6120
ear2a	0.8410	0.2927
ear2b	0.7546	0.4306
ear2c	0.7760	0.3978
ear2d	0.8081	0.3470
ear2e	0.7298	0.4674
ear3a	0.8641	0.2533
ear3b	0.7817	0.3889
ear4	0.7460	0.4435
ear5a	0.6314	0.6013
ear5b	0.6218	0.6134

Step 5: Latent Structure Analysis – Structural Equation Modeling




```
. sem (ear11a-ear5b <- ETA), var(ETA@1) standardized
```

Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Measurement						
ear11a						
ETA	.6265437	.0353141	17.74			
_cons	2.362817	.1073507	22.01			
ear2a						
ETA	.8454712	.0179031	47.22			
_cons	1.642187	.0844864	19.44			
ear2b						
ETA	.7639354	.0249605	30.61			
_cons	1.991468	.0952197	20.91			

var(e.ear11a)	.607443	.0442517
var(e.ear2a)	.2851784	.0302732
var(e.ear2b)	.4164027	.0381364
var(e.ear2c)	.3924561	.0370933
var(e.ear2d)	.3394463	.0339581
var(e.ear2e)	.4649219	.0403897
var(e.ear3a)	.2323609	.0264024
var(e.ear3b)	.3965921	.0371671
var(e.ear4)	.4696493	.0406111
var(e.ear5a)	.6364757	.0445871
var(e.ear5b)	.6569394	.0446471
var(ETA)	1	(constrained)

LR test of model vs. saturated: chi2(44) =

SEM Goodness of Fit (briefly): Baseline Comparison

```
. estat gof, stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(44)	335.693	model vs. saturated
p > chi2	0.000	
chi2_bs(55)	2498.970	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA		
90% CI, lower bound		
upper bound		
pclose		
Information criteria		
AIC		
BIC		
Baseline comparison		
CFI	0.881	Comparative fit index
TLI	0.851	Tucker-Lewis index
Size of residuals		
SRMR	0.062	Standardized root mean
CD	0.942	Coefficient of determin

The comparative fit index (Bentler 1990) is computed as

$$\mathrm{CFI} = 1 - \left[\frac{(\chi^2_{ms} - \mathrm{df}_{ms})}{\max\{(\chi^2_{bs} - \mathrm{df}_{bs}), (\chi^2_{ms} - \mathrm{df}_{ms})\}}\right]$$

The Tucker–Lewis index (Bentler 1990) is computed as

$$\mathrm{TLI} = \frac{(\chi^2_{bs}/\mathrm{df}_{bs}) - (\chi^2_{ms}/\mathrm{df}_{ms})}{(\chi^2_{bs}/\mathrm{df}_{bs}) - 1}$$

What percent of worst possible (baseline vs. saturated) fit does my model account for?
Here, .881 is 88.1% of the bad fit.
Around .9 is generally "okay."

SEM Goodness of Fit (briefly): Population Error

```
Population error
                RMSEA        0.142   Root mean squared error of approximation
 90% CI, lower bound         0.128
        upper bound          0.156
              pclose         0.000   Probability RMSEA <= 0.05
```

$$RMSEA = \sqrt{\frac{\chi^2 - df}{N * df}}$$

- Favors simpler models and larger sample sizes. Lower the better.
- Can we be somewhat sure that the standardized distance (badness of fit) is low? (lower bound of 90% CI less than .05)
- And can we be somewhat sure that the standardized distance (badness of fit) is not high? (upper bound of 90% CI less than .10).

An 8-Step Plan

Step 1: Content and Cognition

Step 2: Scoring and Scaling

Step 3: Correlation and Reliability

Step 4: Classical Item Diagnostics

Step 5: Latent Structure Analysis

Step 6: Item Response Theory (IRT)

Step 7: IRT for Efficient Measurement

Step 8: Correlation and Prediction

Step 6: Item Response Theory. Why IRT?

- Item response theory (IRT) supports the vast majority of large-scale educational assessments.
  - State testing programs
  - National and international assessments (NAEP, TIMSS, PIRLS, PISA).
  - Selection testing (SAT, ACT)
- Many presentations of IRT use unfamiliar jargon and specialized software.
  - We will try to connect IRT to other more flexible statistical modeling frameworks.
  - We will use Stata.

Classical Test Theory vs. Item Response Theory

- CTT: A response i to item j by person k :

$$y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk};$$
$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$

- IRT: A response i to item j by person k :

$$\log\left(\frac{P(y_{ijk} = 1)}{1 - P(y_{ijk} = 1)}\right) = \alpha_j + \zeta_k;$$
$$\zeta_k \sim N(0,1).$$

  - A logistic model vs. a linear model.
  - Fixed item effects (α_j) vs. random item effects (ζ_j).
  - Both models have random effects for persons.
  - IRT extends to a fixed slope coefficient for items, β_j, on the random slope:

$$\log\left(\frac{P(y_{ijk} = 1)}{1 - P(y_{ijk} = 1)}\right) = \alpha_j + \beta_j \zeta_k; \quad \zeta_k \sim N(0,1)$$

Slope-Intercept vs. Discrimination-Difficulty Parameterizations

Slope-Intercept Parameterization

We're familiar with logistic regression models of the form:

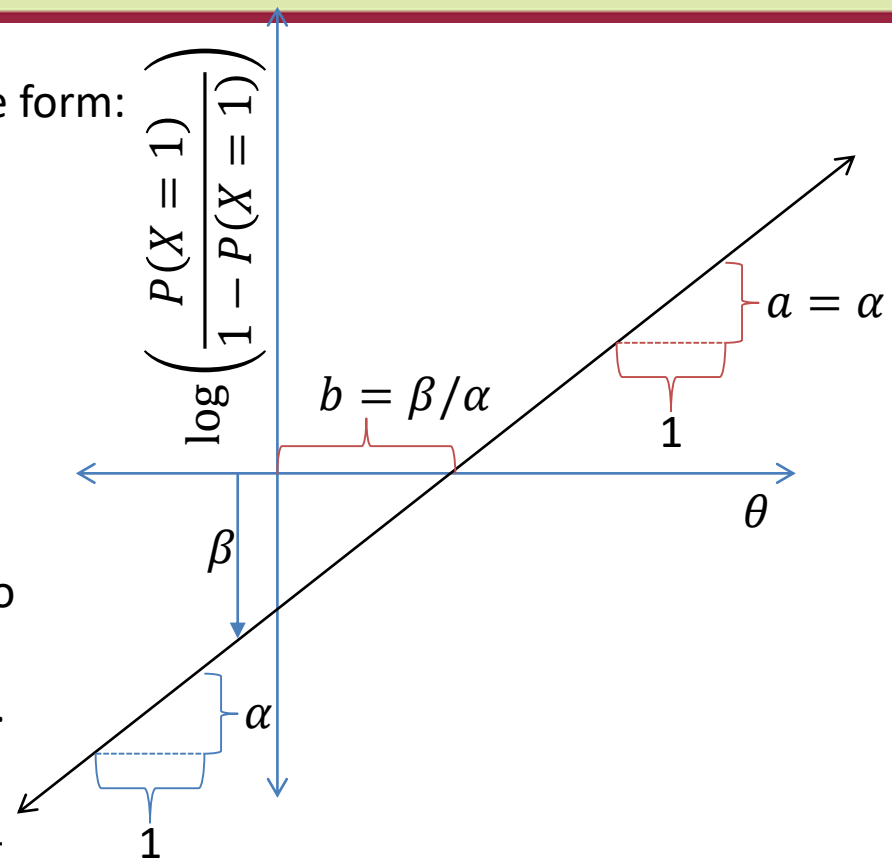
$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

β_0 is the y-intercept, and β_1 is the slope.

IRT models can have a similar parameterization:

$$\log\left(\frac{P(X=1)}{1-P(X=1)}\right) = \alpha\theta - \beta$$

Note β is the *negative* y-intercept, corresponding to difficulty, and α is the slope. See the figure at right. Notice that β is on the logit scale, the $-y$-intercept.



Discrimination-Difficulty Parameterization

In contrast, in IRT, we prefer to think of difficulty on the same scale as θ, as the x-intercept.

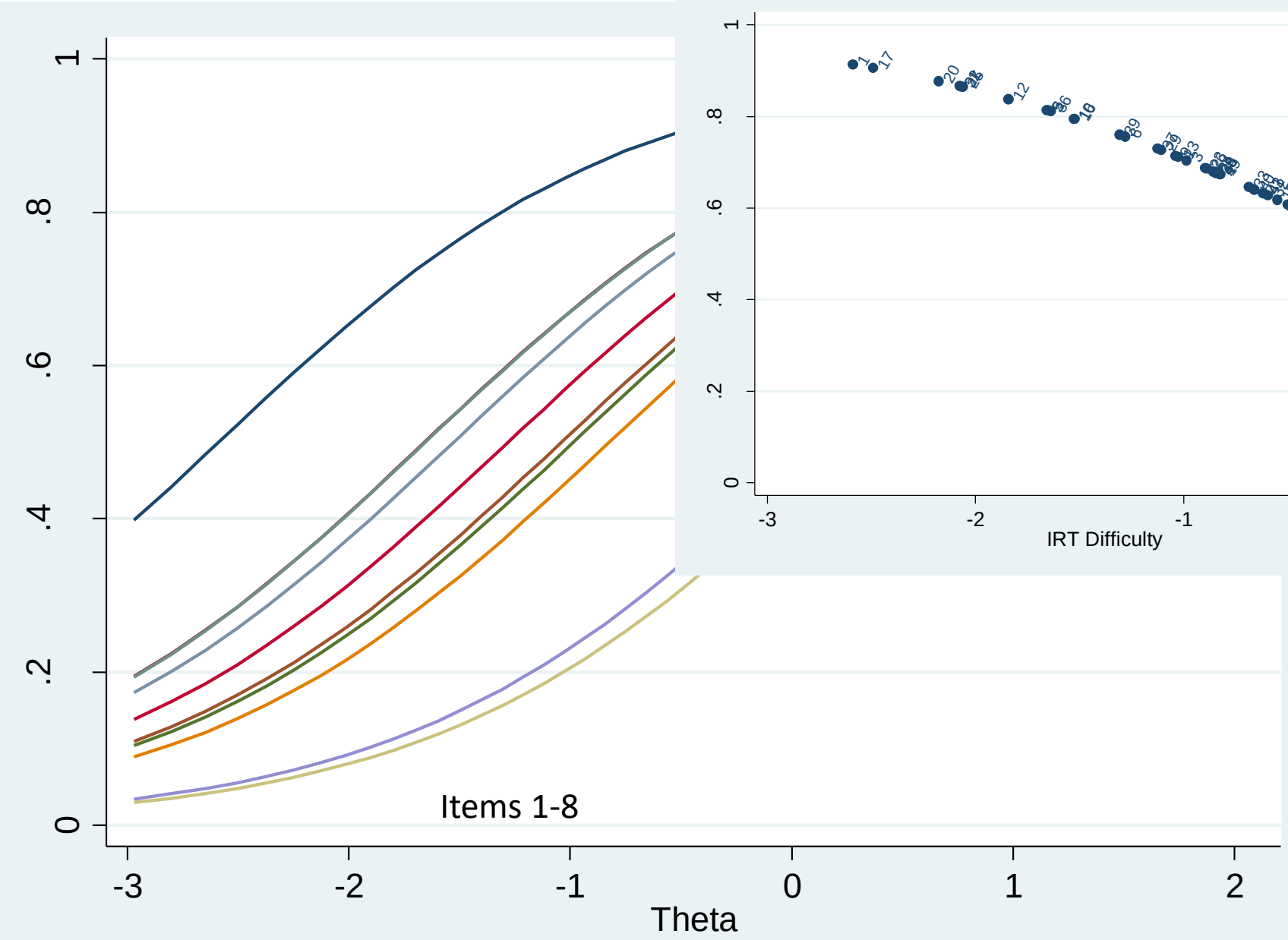
So, we use the parameterization:

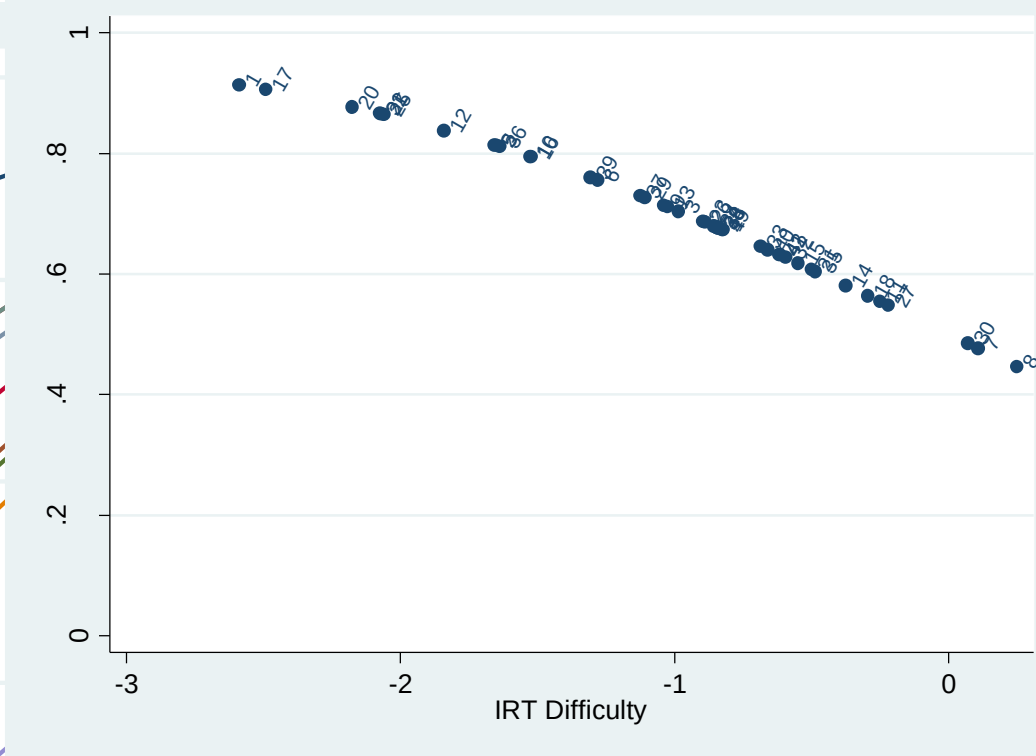
$$\log\left(\frac{P(X=1)}{1-P(X=1)}\right) = a(\theta - b)$$

The slope here, a, is equal to α in the slope-intercept parameterization, but $b = \beta/\alpha$ and $\beta = ab$.

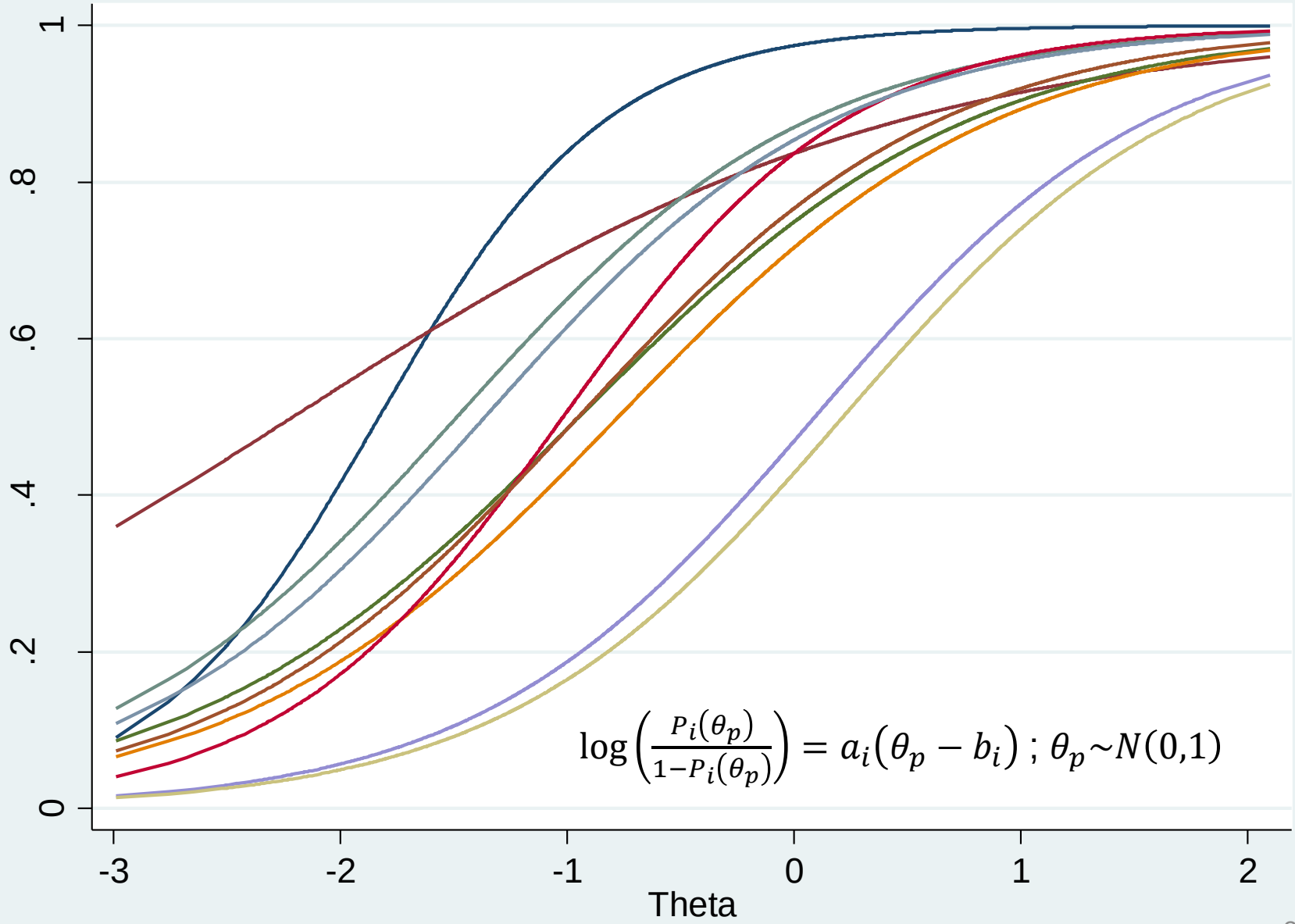
- In slope-intercept parameterization, β is the log-odds of getting an item wrong when $\theta = 0$.
- In discrimination-difficulty parameterization, b is the θ you need for even odds (50%) of a correct answer.

1 Parameter Logistic (1PL) Item Characteristic Curves (ICCs)





2 Parameter Logistic (1PL) Item Characteristic Curves (ICCs)



$$\log\left(\frac{P_i(\theta_p)}{1-P_i(\theta_p)}\right) = a_i(\theta_p - b_i)\ ;\ \theta_p \sim N(0,1)$$

Item Characteristic Curve (ICC) Slider Questions

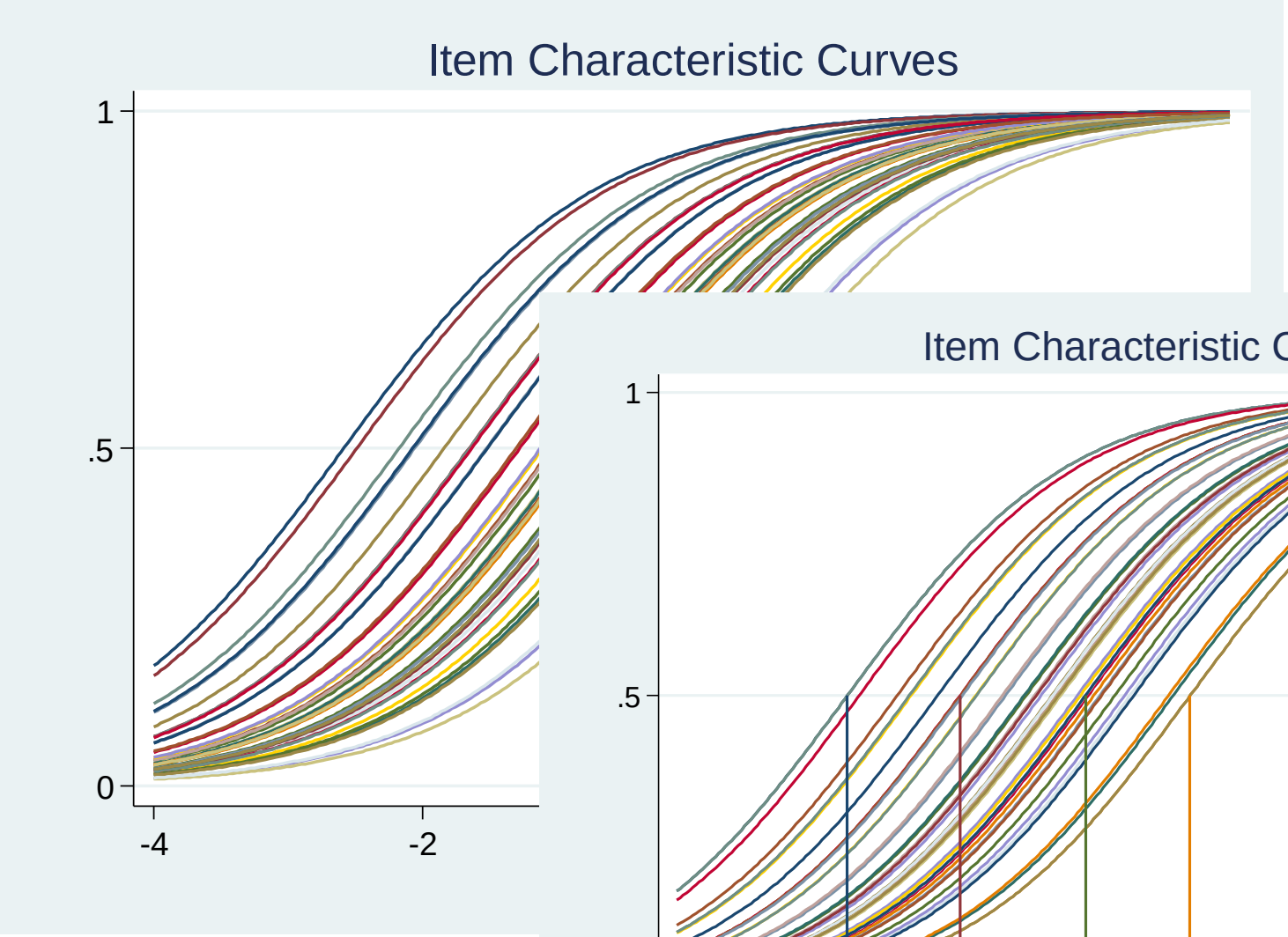
- What happens when we increase a for the blue item? Which item is more discriminating?
- What happens when we increase b for the blue item? Which item is more difficult?
- What happens when we increase c for the blue item? Which item is more discriminating?
- Try setting blue to .84, 0, .05 and red to .95,.3, .26. Why might the c parameter be the most difficult to estimate in practice?
- Given this overlap, comparisons of items in terms of item parameters instead of full curves will be shortsighted.
  - Difficulty for which θ? Discrimination for which θ?
- For reference, the probability of a correct response when $\theta_p = b_i$ is $\frac{1+c_i}{2}$. The slope at this inflection point is $\frac{a_i(1-c_i)}{4}$.

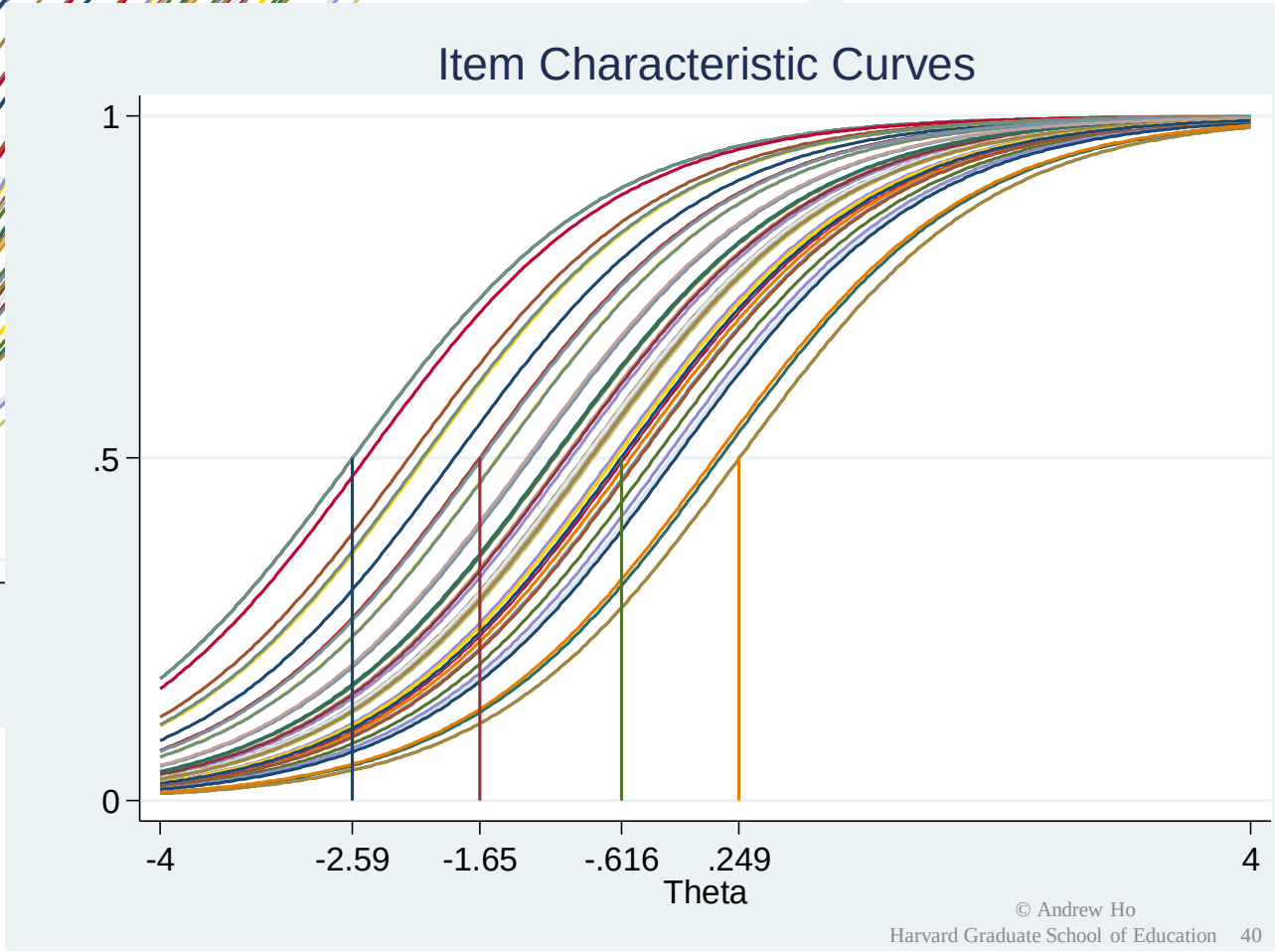
IRT in Stata: 1PL (The Rasch Model)

```
*-----------------------------------------------------------------------------
*
* 1 Parameter Logistic (1PL) Model: The Rasch Model
*
*-----------------------------------------------------------------------------

    * Fit the item response model
    irt 1pl item1-item42
    estimates store mcas1pl

    * List items by difficulty
    estat report, sort(b) byparm

    * Plotting Item Characteristic Curves
    irtgraph icc, legend(off)
    irtgraph ic· estat report, sort(b) byparm

    * Estimatin One-parameter logistic model                    Number of obs     =      5,000
    predict the Log likelihood = -108475.31
```

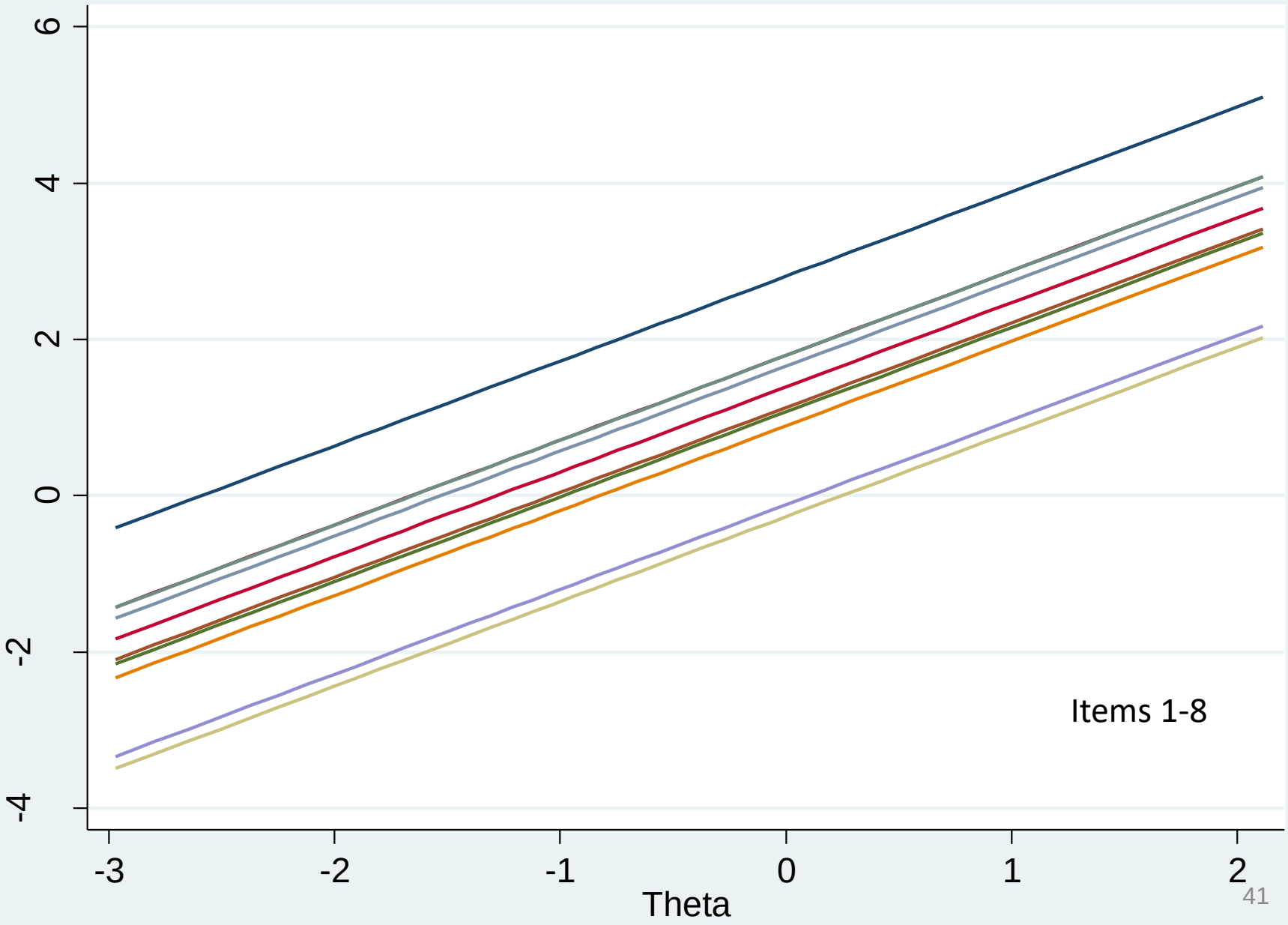
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim	1.084066	.0131721	82.30	0.000	1.058249	1.109882
Diff						
item1	-2.58876	.0586296	-44.15	0.000	-2.703671	-2.473848
item17	-2.491368	.0566288	-43.99	0.000	-2.602359	-2.380378
item20	-2.177048	.0508108	-42.85	0.000	-2.276635	-2.077461
item31	-2.075762	.0491291	-42.25	0.000	-2.172053	-1.979471
item22	-2.073947	.0490997	-42.24	0.000	-2.170181	-1.977713

1 Parameter Logistic (1PL) Item Characteristic Curves (ICCs)

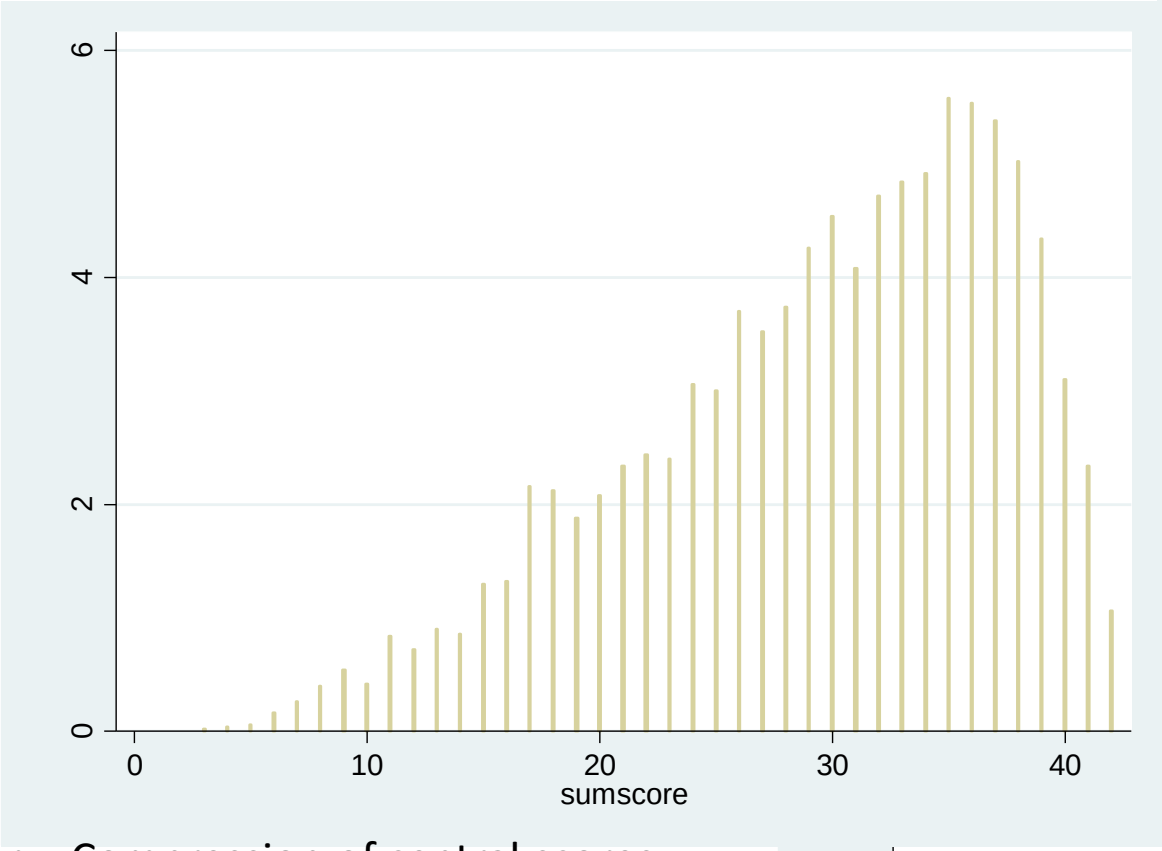


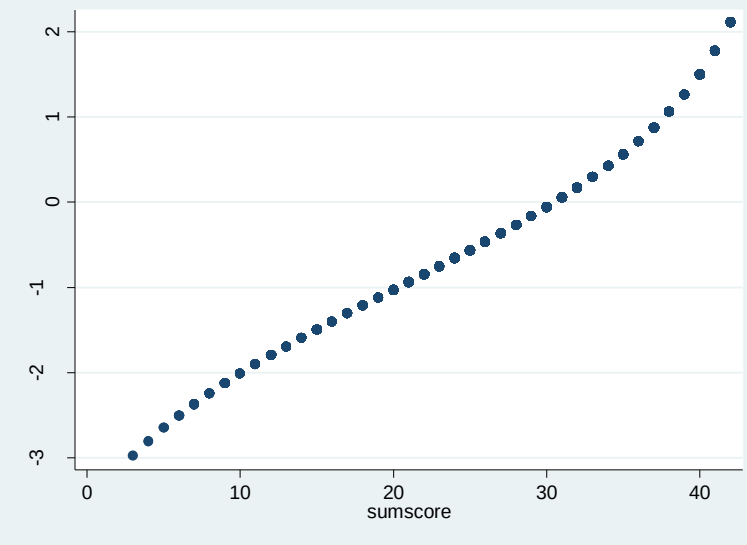


1 Parameter Logistic (1PL) ICCs in Logit Space (Linear)

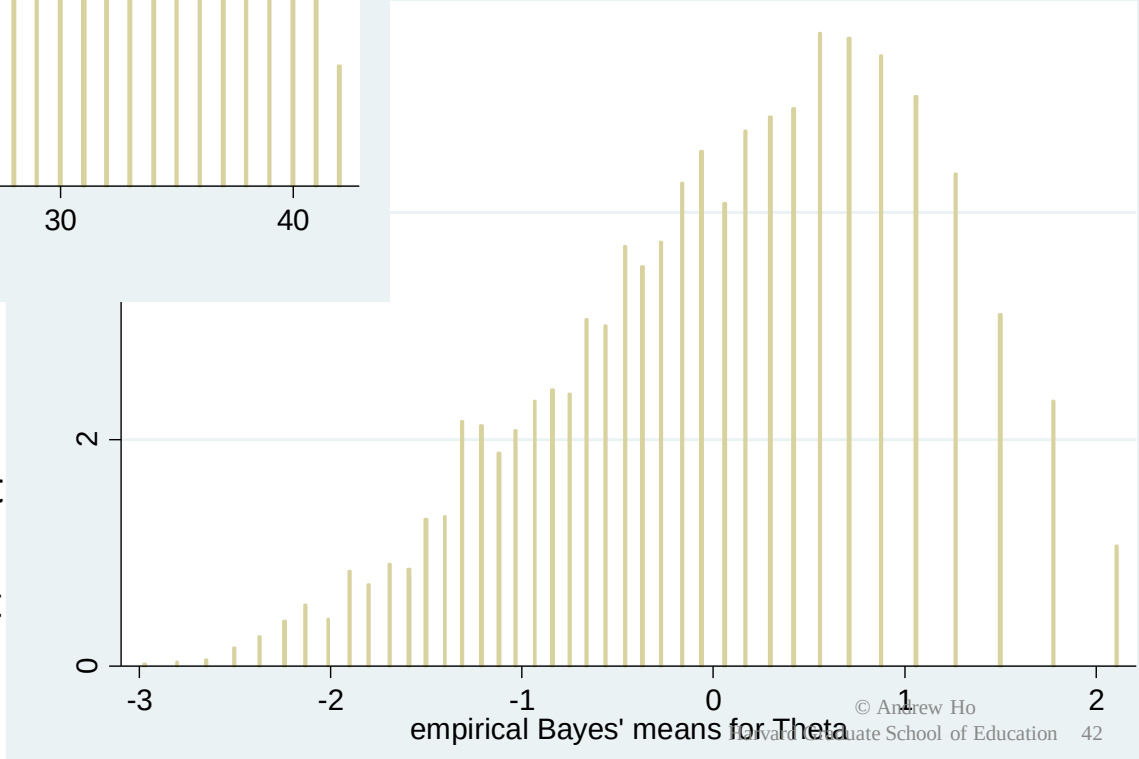


The Rasch (1PL) Scale Transformation





- Compression of central scores, stretching of extremes.
- Relative error initially greater in central scores, afterwards greater at extremes.
- Information initially concentrated at extreme score points, afterwards concentrated centrally.



The 2-Parameter Logistic (2PL) IRT Model

$$\log\left(\frac{P_i(\theta_p)}{1-P_i(\theta_p)}\right) = a_i(\theta_p - b_i)$$

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

Two-parameter logistic model                    Number of obs     =      5,000
Log likelihood =  -107593.1
```

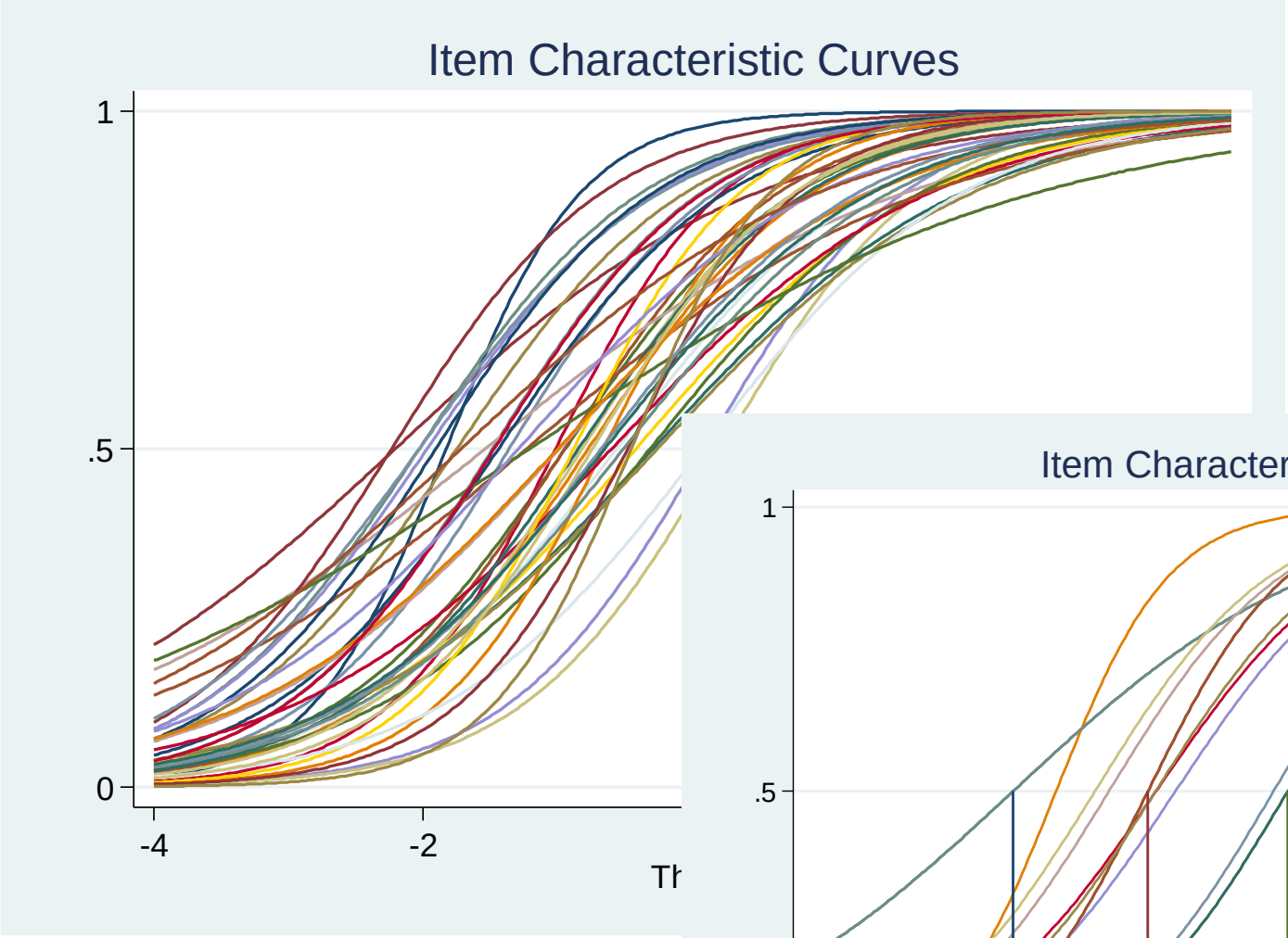
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
item33	.5270587	.0348567	15.12	0.000	.4587408	.5953765
item13	.6359186	.0378883	16.78	0.000	.5616589	.7101783
item24	.6692801	.0374937	17.85	0.000	.5957938	.7427663
item2	.7392442	.0444423	16.63	0.000	.652139	.8263495
item39	.7490874	.0414677	18.06	0.000	.6678122	.8303627
item21	.8375887	.0389584	21.50	0.000	.7612317	.9139458
item27	.8473439	.0387146	21.89	0.000	.7714647	.923223

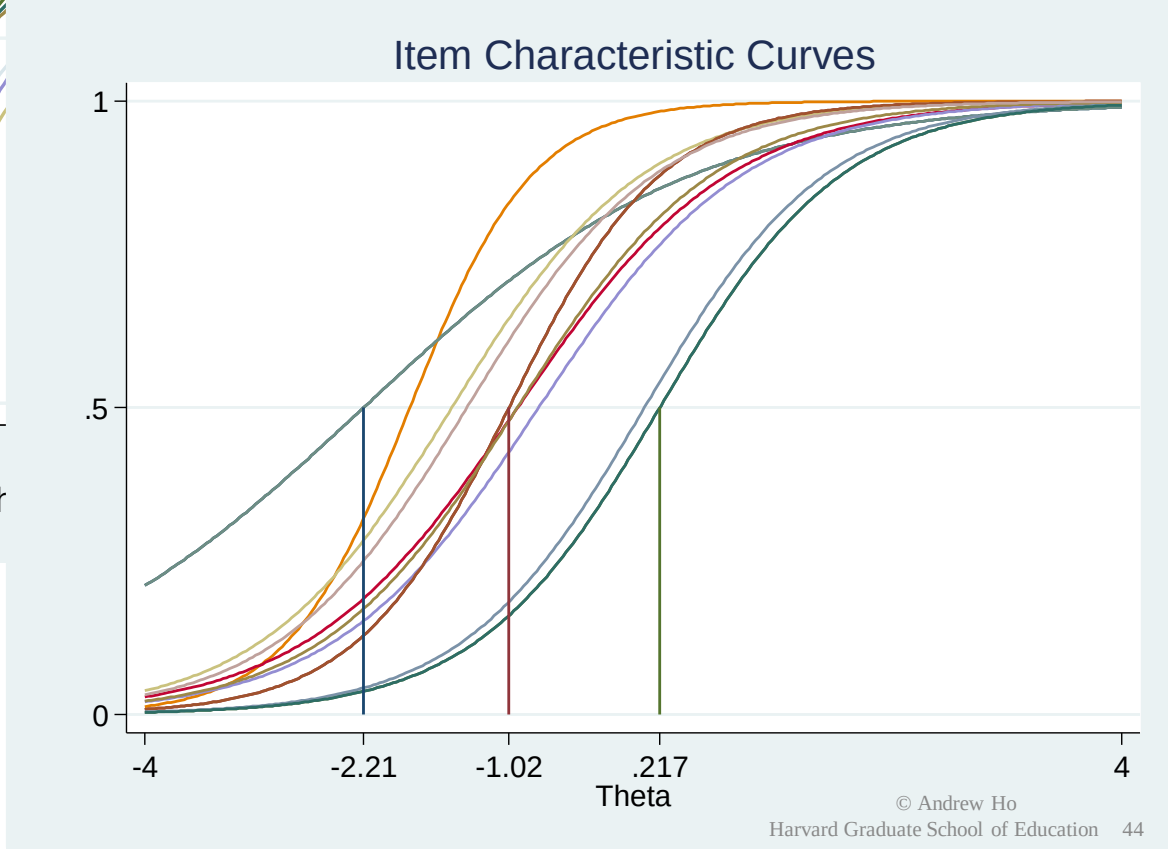
Discrimination is the difference in the log-odds of a correct answer for every SD distance of θ_p from b_i.

```
.          lrtest mcas1pl mcas2pl

Likelihood-ratio test                                 LR chi2(41) =   1764.42
(Assumption: mcas1pl nested in mcas2pl)               Prob > chi2 =    0.0000
```

Likelihood ratio test: reject null hypothesis that discrimination parameters are jointly equal. 2PL fits better than 1PL.

2-Parameter Logistic (2PL) ICCs





The 3-Parameter Logistic (3PL) IRT Model

$$P_i(\theta_p) = c + \frac{(1-c)\left(\exp\left(a_i(\theta_p - b_i)\right)\right)}{1 + \exp\left(a_i(\theta_p - b_i)\right)}$$

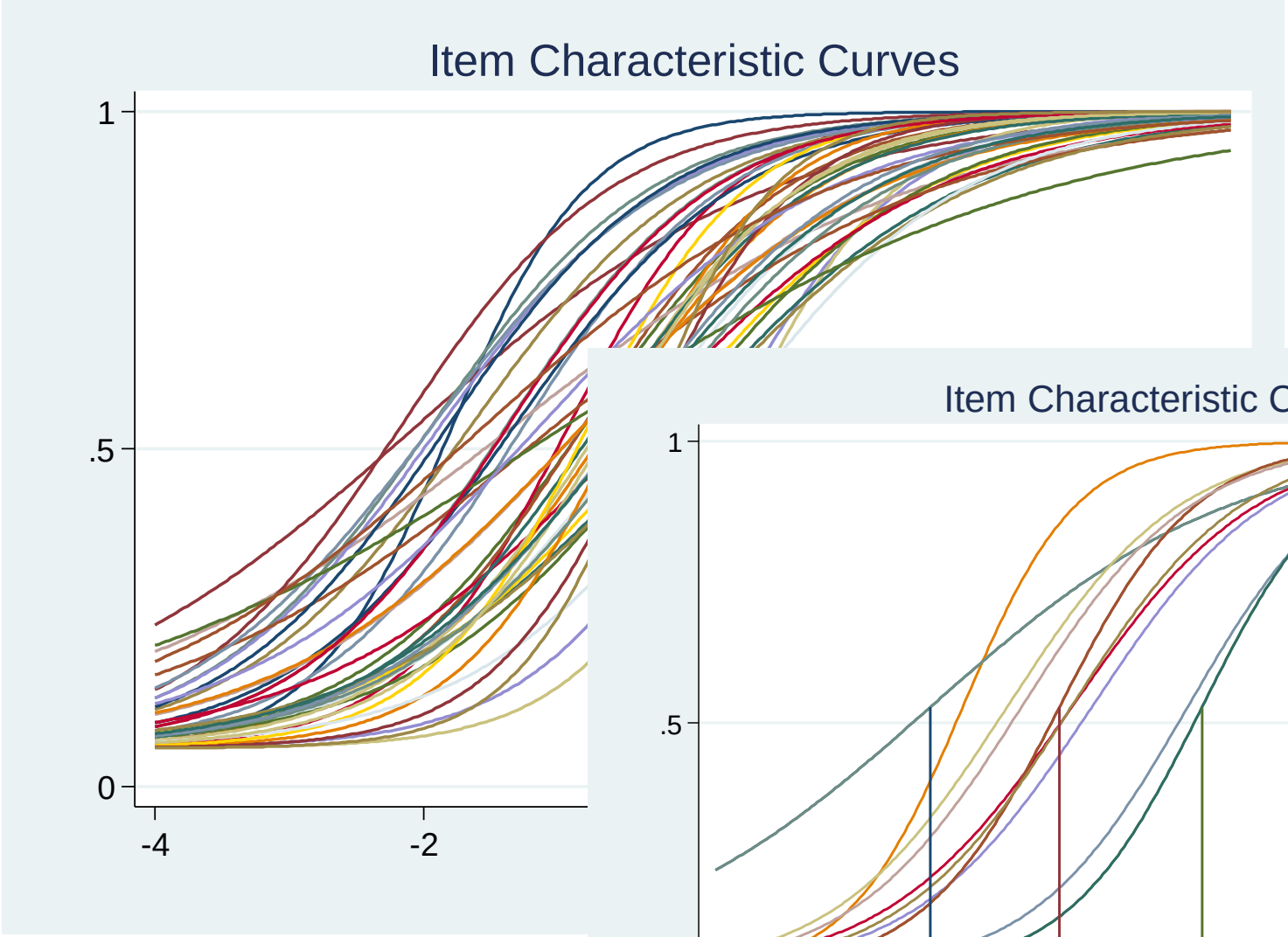
item14	-.2654169	.045632	-5.82	0.000	-.3548539	-.1759799
item18	-.1773411	.0421862	-4.20	0.000	-.2600245	-.0946577
item11	-.1407813	.0463368	-3.04	0.002	-.2315998	-.0499628
item27	-.1040767	.0468187	-2.22	0.026	-.1958397	-.0123137
item30	.2031855	.0406553	5.00	0.000	.1235025	.2828685
item7	.2036091	.0339582	6.00	0.000	.1370522	.270166
item8	.3319137	.0322655	10.29	0.000	.2686745	.3951529
Guess	.056637	.0092217	6.14	0.000	.0385629	.0747112

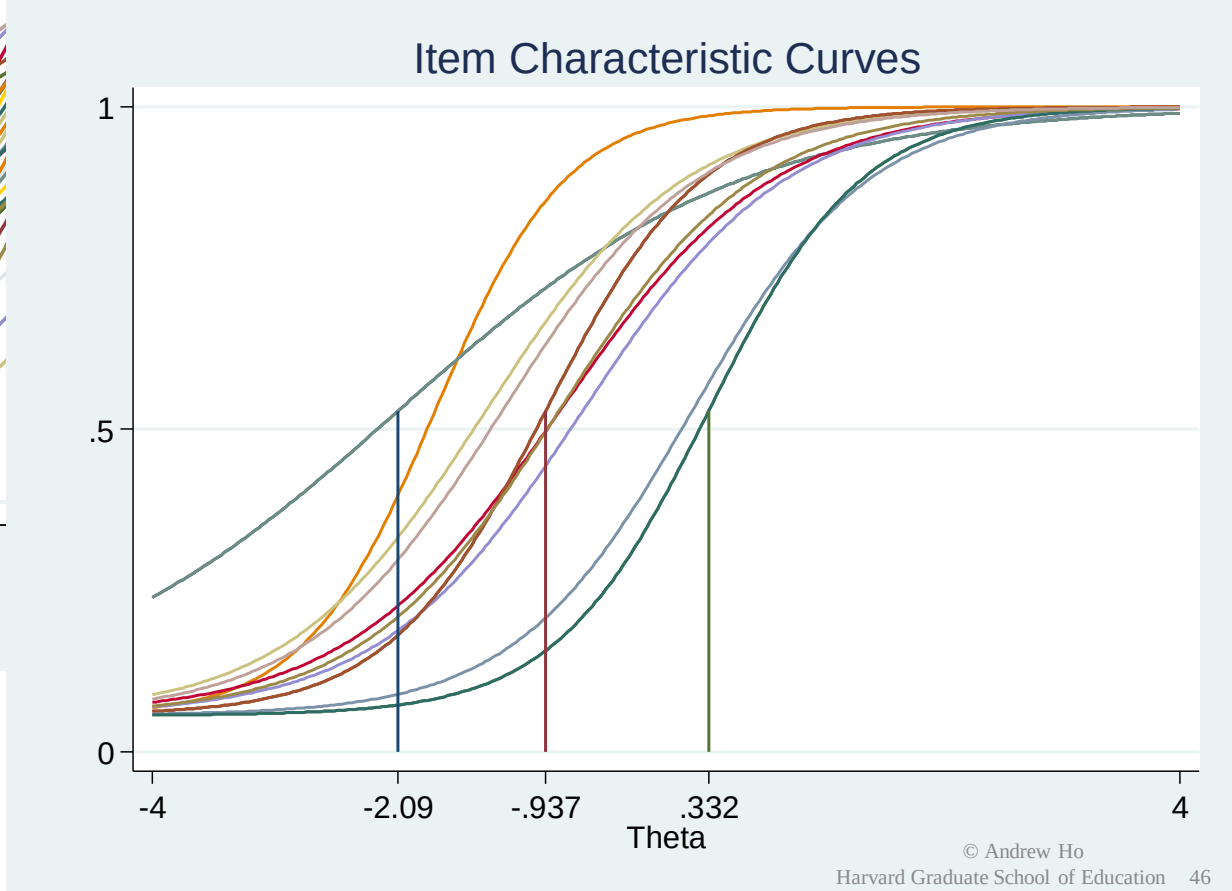
The common c parameter estimate is an estimated lower asymptote, the pseudo-guessing parameter. Estimated in common across items c rather than c_i due to considerable estimation challenges in practice.

```
.          lrtest mcas2pl mcas3pl

Likelihood-ratio test                                 LR chi2(1)  =     31.08
(Assumption: mcas2pl nested in mcas3pl)               Prob > chi2 =    0.0000
```

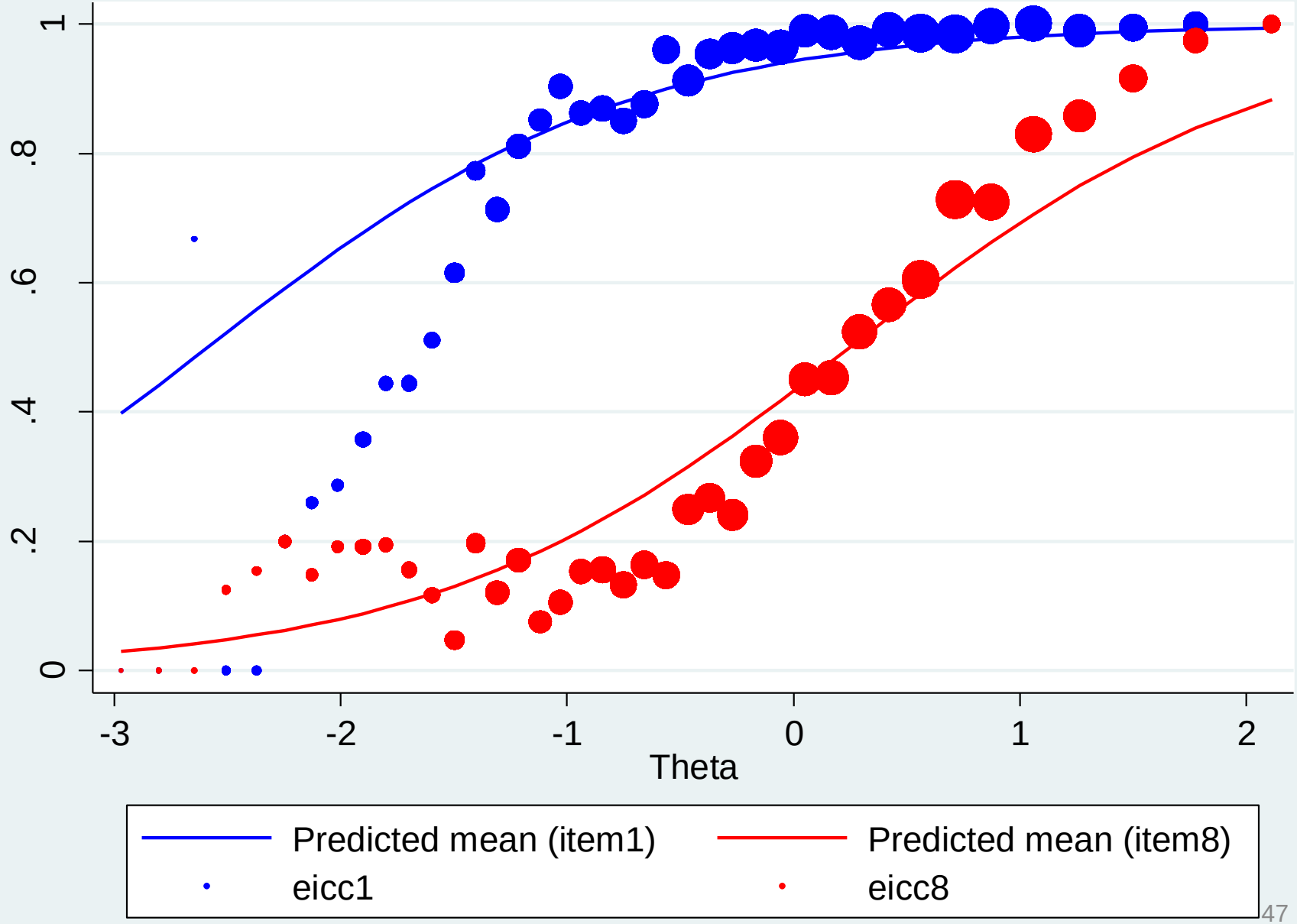
Likelihood ratio test: reject null hypothesis that the common pseudo-guessing parameter is zero.

3-Parameter Logistic (3PL) ICCs





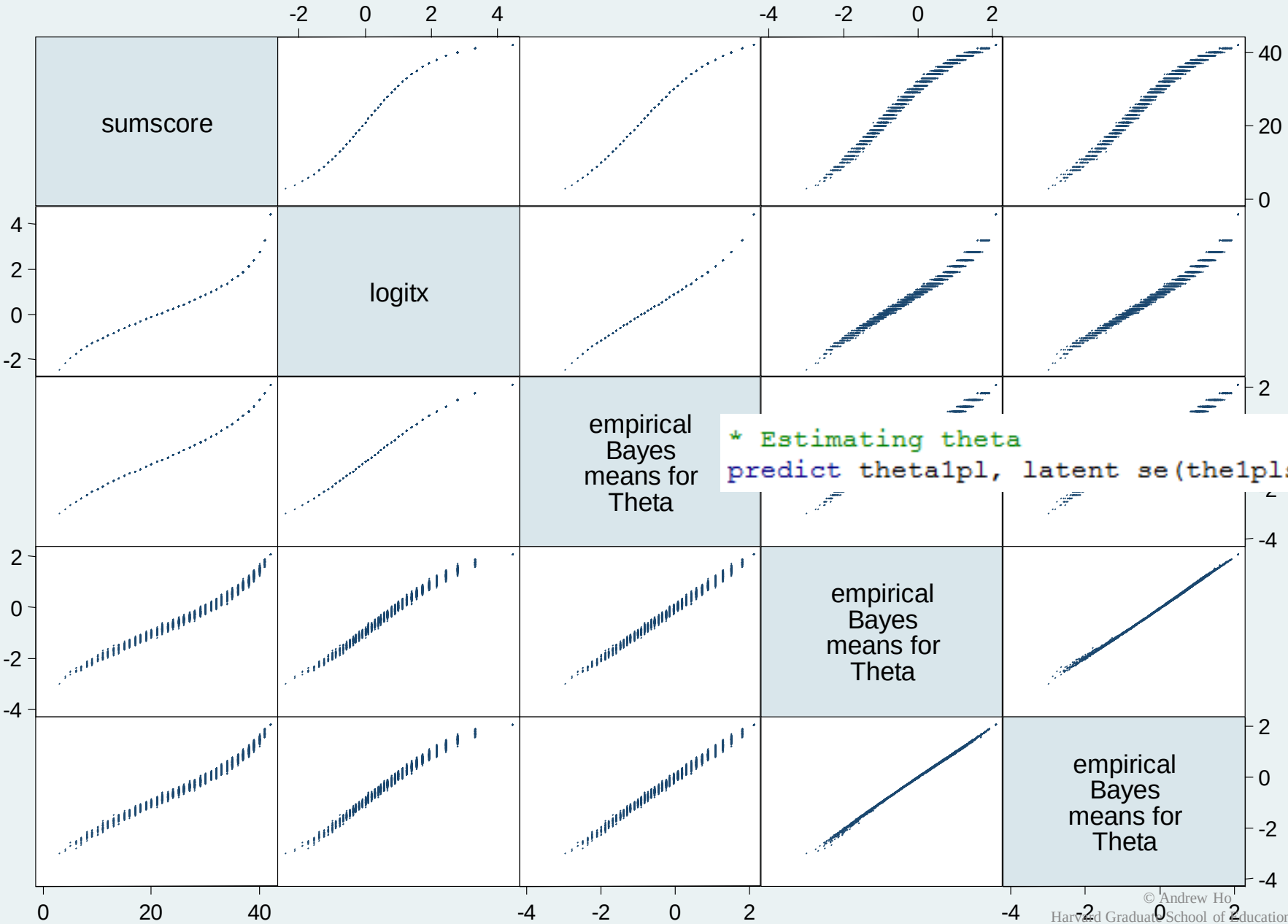
Graphical Goodness of Fit for Items 1 and 8



Loose Sample Size Guidelines (Yen & Fitzpatrick)

- Rasch (1PL): 20 items and 200 examinees.
- Hulin, Lissak, and Drasgow:
  - 2PL: 30 items, 500 examinees.
  - 3PL: 60 items, 1000 examinees.
  - Tradeoffs, maybe 30 items and 2000 examinees.
- Swaminathan and Gifford:
  - 3PL: 20 items, 1000 examinees.
- Low scoring examinees needed for 3PL.
- Large samples (above 3500) needed for polytomous items (scored 0/1/2/…), particularly high or low difficulty items that will have even higher or lower score points.

Estimating θ_p via Empirical Bayes (EAP)




```
* Estimating theta
predict theta1pl, latent se(the1plse)
```

Dichotomous vs. Polytomous IRT

- Define $k = 0,1,\ldots,K$ categories, where $K = 1$ is the dichotomous case (responses scored 0 or 1) and $K \geq 2$ is the polytomous case.
  - Note that K refers to the number of category boundaries or "cut scores."

Dichotomous IRT:

$$P(X_{pi} = 1 | a_i, b_i, \theta_p) = \frac{1}{1 + \exp\left(-a_i(\theta_p - b_i)\right)}$$

$$\log\left(\frac{P_i(\theta_p)}{1 - P_i(\theta_p)}\right) = a_i(\theta_p - b_i)$$

Polytomous (Graded Response Model):

$$P(X_{pi} \geq k | a_i, b_{ik}, \theta_p) = \frac{1}{1 + \exp\left(-a_i(\theta_p - b_{ik})\right)}$$

$$\log\left(\frac{P_{ik}(\theta_p)}{1 - P_{ik}(\theta_p)}\right) = a_i(\theta_p - b_{ik})$$

Graded Response Model (GRM), Slope-Intercept Parameterization:

$$P(X_{pi} \geq k | \alpha_i, \beta_{ik}, \theta_p) = \frac{1}{1 + \exp\left(-(\alpha_i\theta_p - \beta_{ik})\right)}$$

$$\boxed{\log\left(\frac{P_{ik}(\theta_p)}{1 - P_{ik}(\theta_p)}\right) = \alpha_i\theta_p - \beta_{ik}; \qquad a_i = \alpha_i; b_{ik} = \frac{\beta_{ik}}{\alpha_i}}$$

Step 6: Polytomous IRT for the INNER Ear Scale

Table 2. Item Information and Location Parameter Estimates Based on a Graded Response Model (n = 353).[a]

		Location Parameter Estimates				
Item Code	Information Parameter Estimates: a	b1	b2	b3	b4	b5
ear11 (b1-b5)	1.63	–4.24	–2.39	–1.92	–1.30	–0.70
ear11 (b6-b10)		–0.28	0.45	0.79	1.55	2.57
ear2a	3.79	–1.37	–0.10	0.94	1.88	
ear2b	2.72	–1.78	–0.65	0.62	1.62	
ear2c	2.70	–0.65	0.62	1.72	2.90	
ear2d	3.05	–0.84	0.37	1.42	2.32	
ear2e	2.33	–1.55	–0.42	0.79	1.72	
ear3a	4.52	–1.17	0.20	1.16	2.09	
ear3b	2.63	–0.97	–0.09	0.98	1.96	
ear4	2.45	–1.43	–0.42	0.79	1.41	
ear5a	1.69	–1.23	–0.05	1.59		
ear5b	1.86	–1.60	–0.73	0.10		

[a]The ear11 item has 11 score points and thus 10 location parameters that delineate adjacent score points. The ear2, ear3, and ear4 items have 5 score points and thus 4 location parameters. The ear5 items have 4 score points and thus 3 location parameters. See **Table 1** for item prompts, score means, and standard deviations.

Step 7: Efficient Measurement - Item Information

- We can define item information as the ratio of the squared slope of the logistic curve to the conditional variance (think of a Bernoulli trial) as follows:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

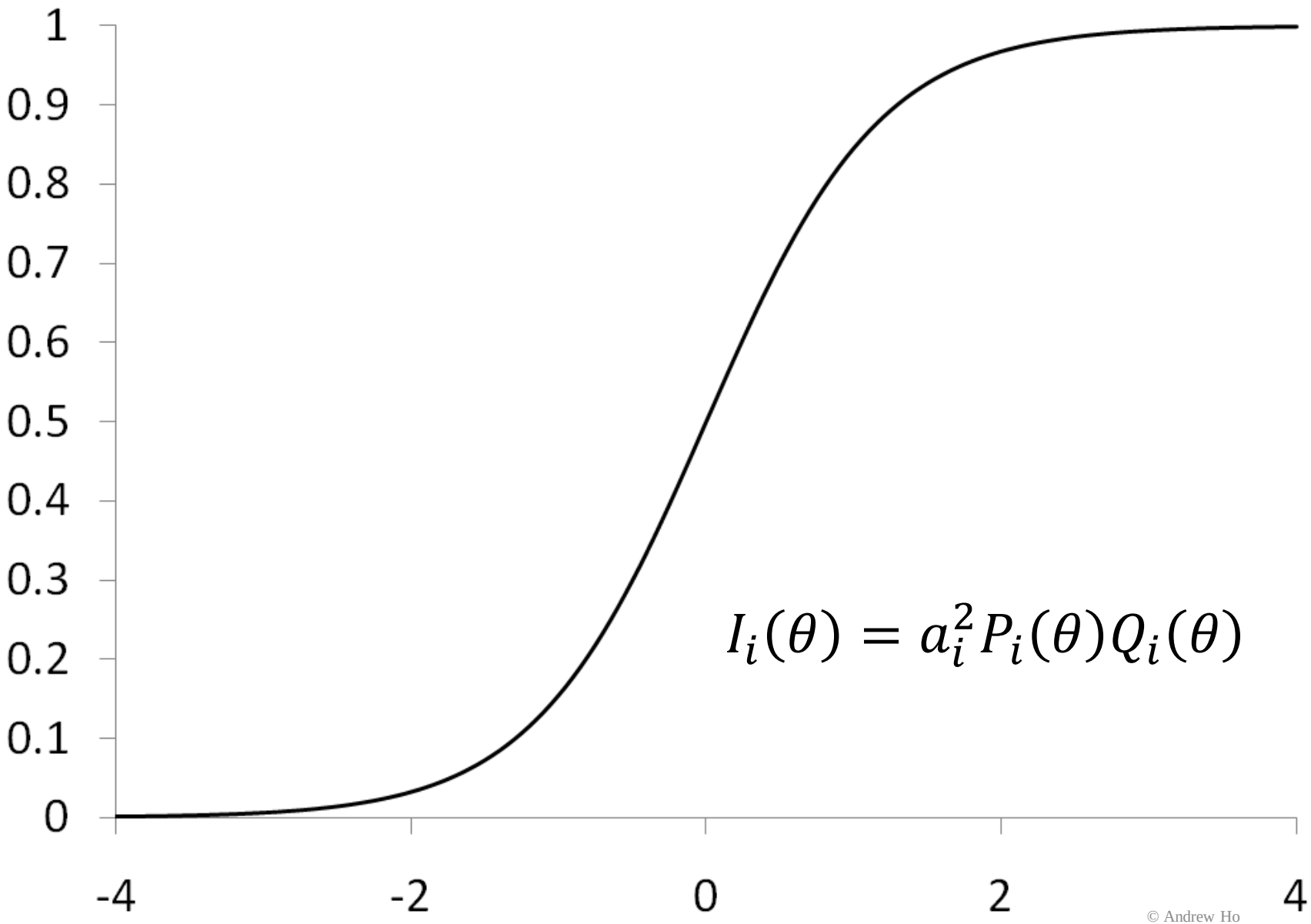
- For 1PL and 2PL:

$$I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta)$$

- Maximized when $P_i = .5$.

- The steeper the slope of the ICC, the greater the information.

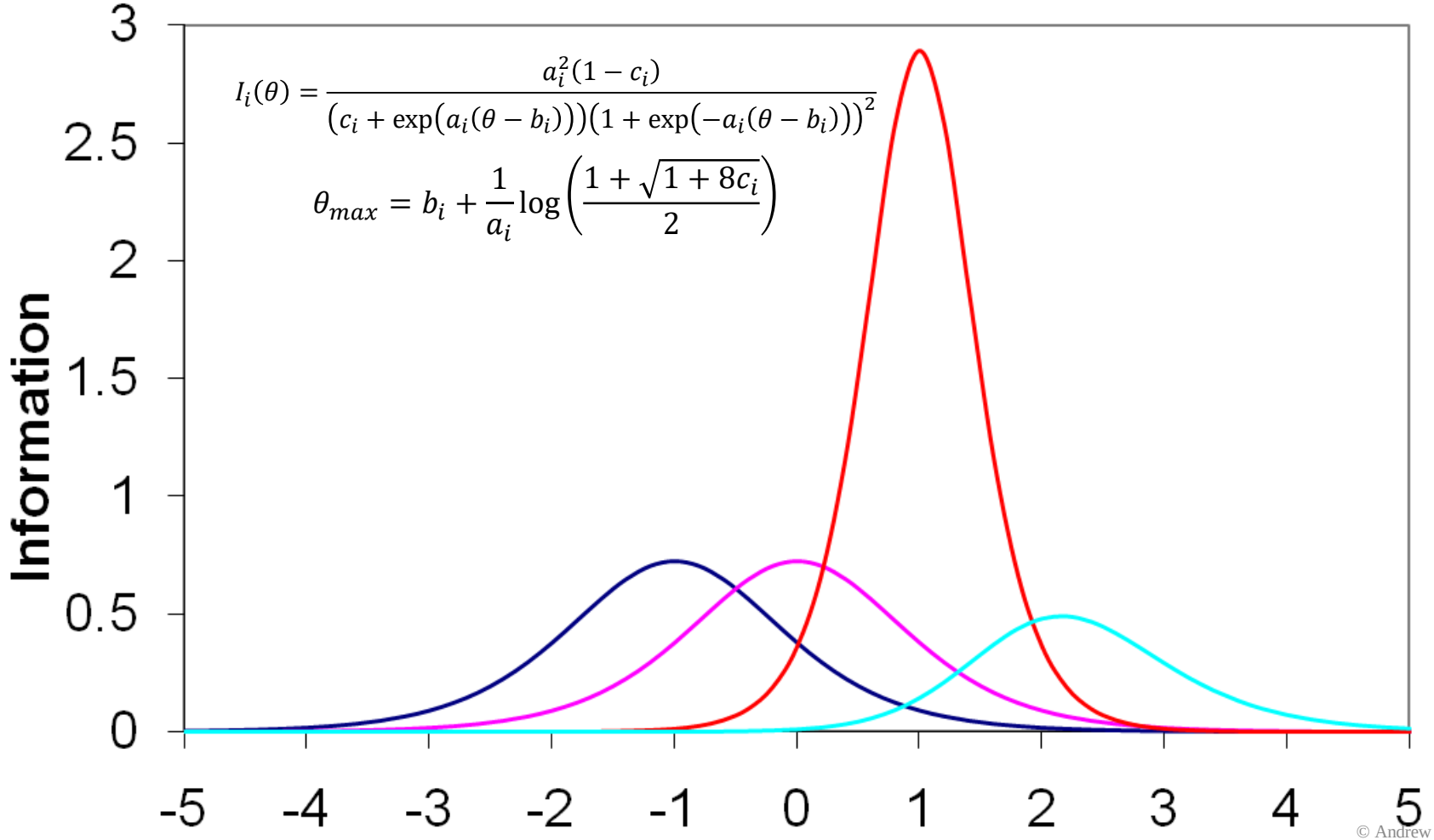
Visualizing Information from an ICC



$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta)$$

Intuition for Item Information

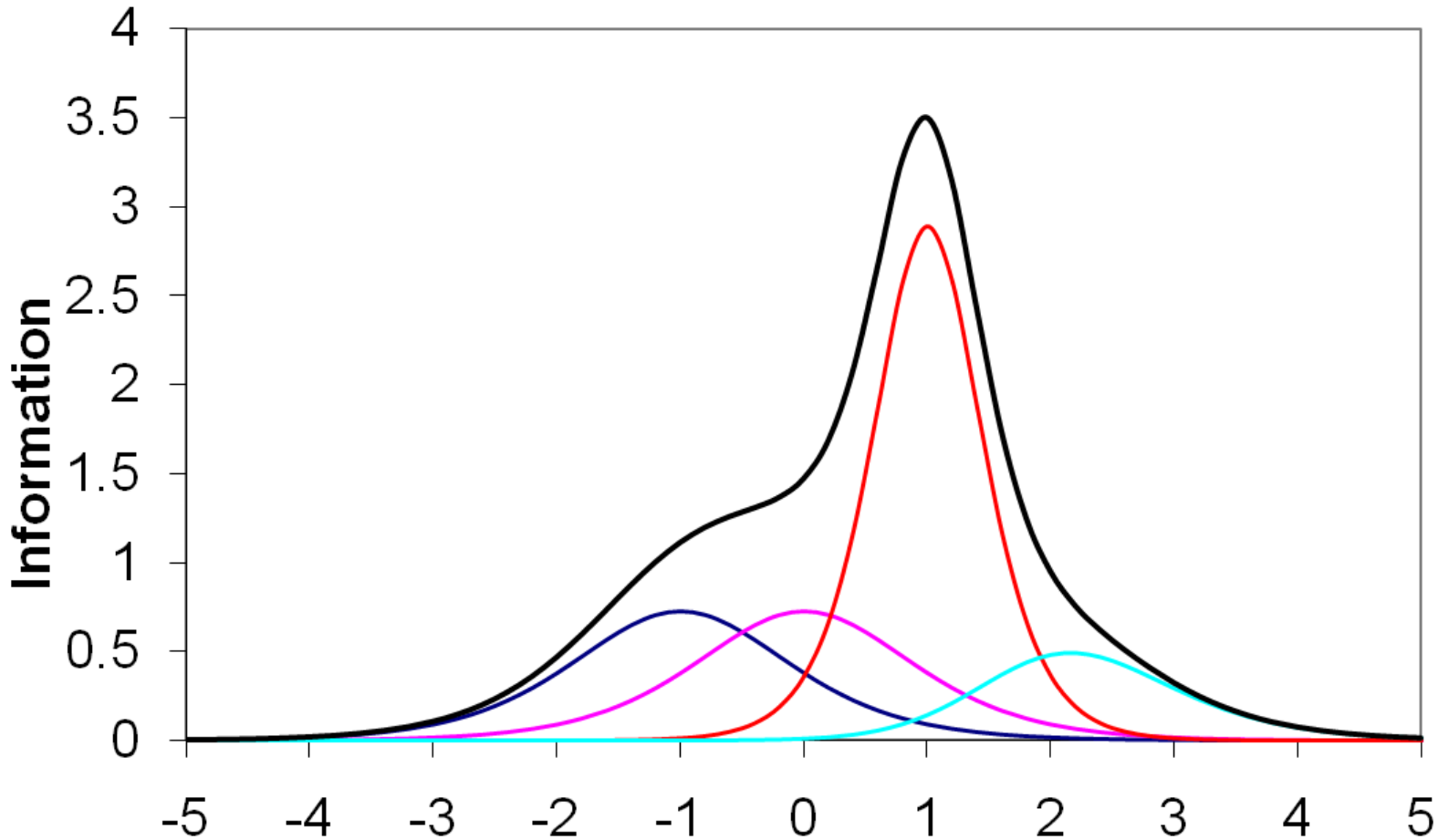
Par | Item	1	2	3	4
a	1	1	2	1
b	-1	0	1	2
c	0	0	0	0.2



Test Information

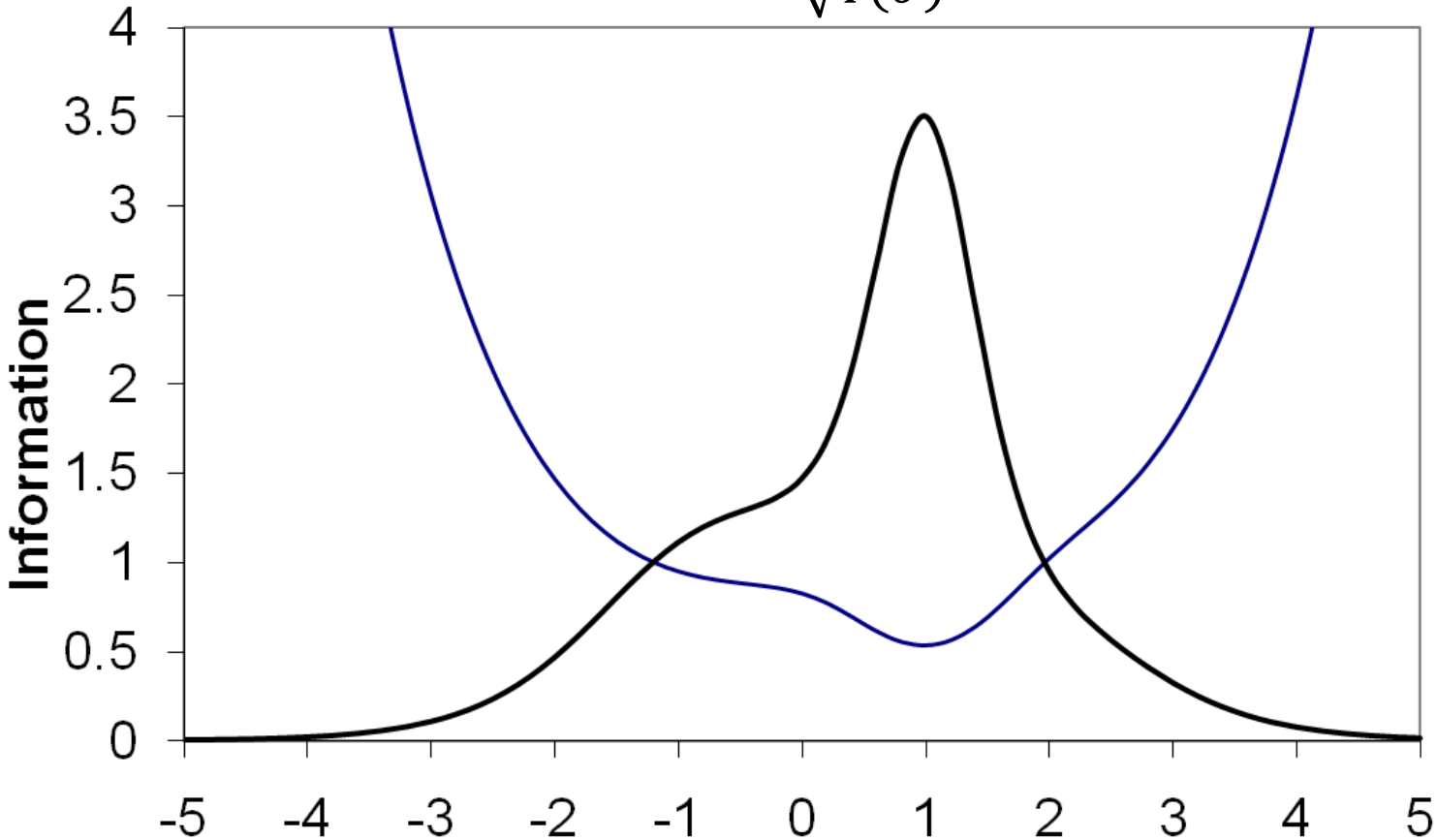
$$I(\theta) = \sum_i I_i(\theta)$$

Test information is the simple sum of item information at a particular θ



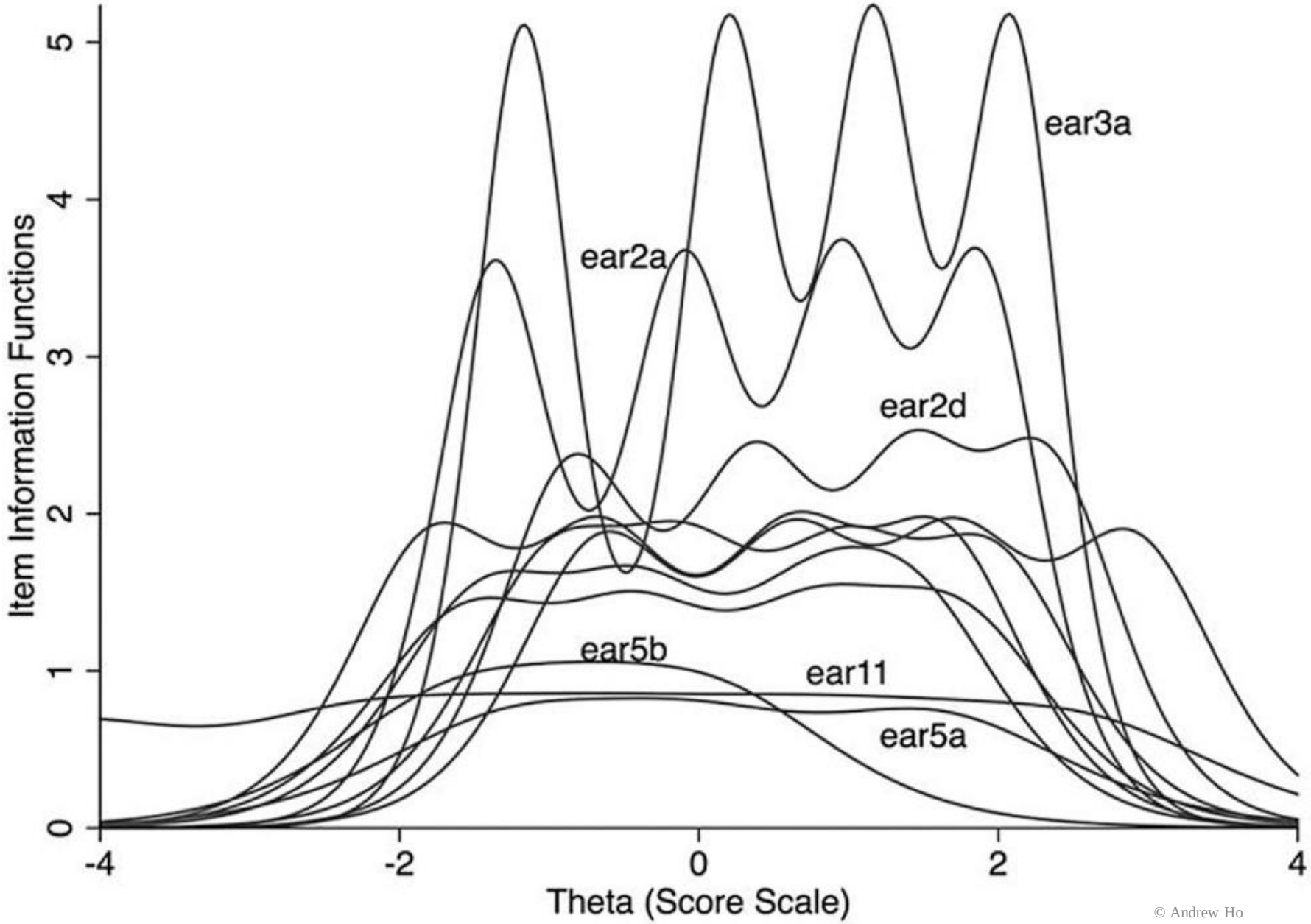
Conditional Standard Error of Measurement (CSEM)

$$SE(\hat{\theta}|\theta) = \frac{1}{\sqrt{I(\theta)}}$$

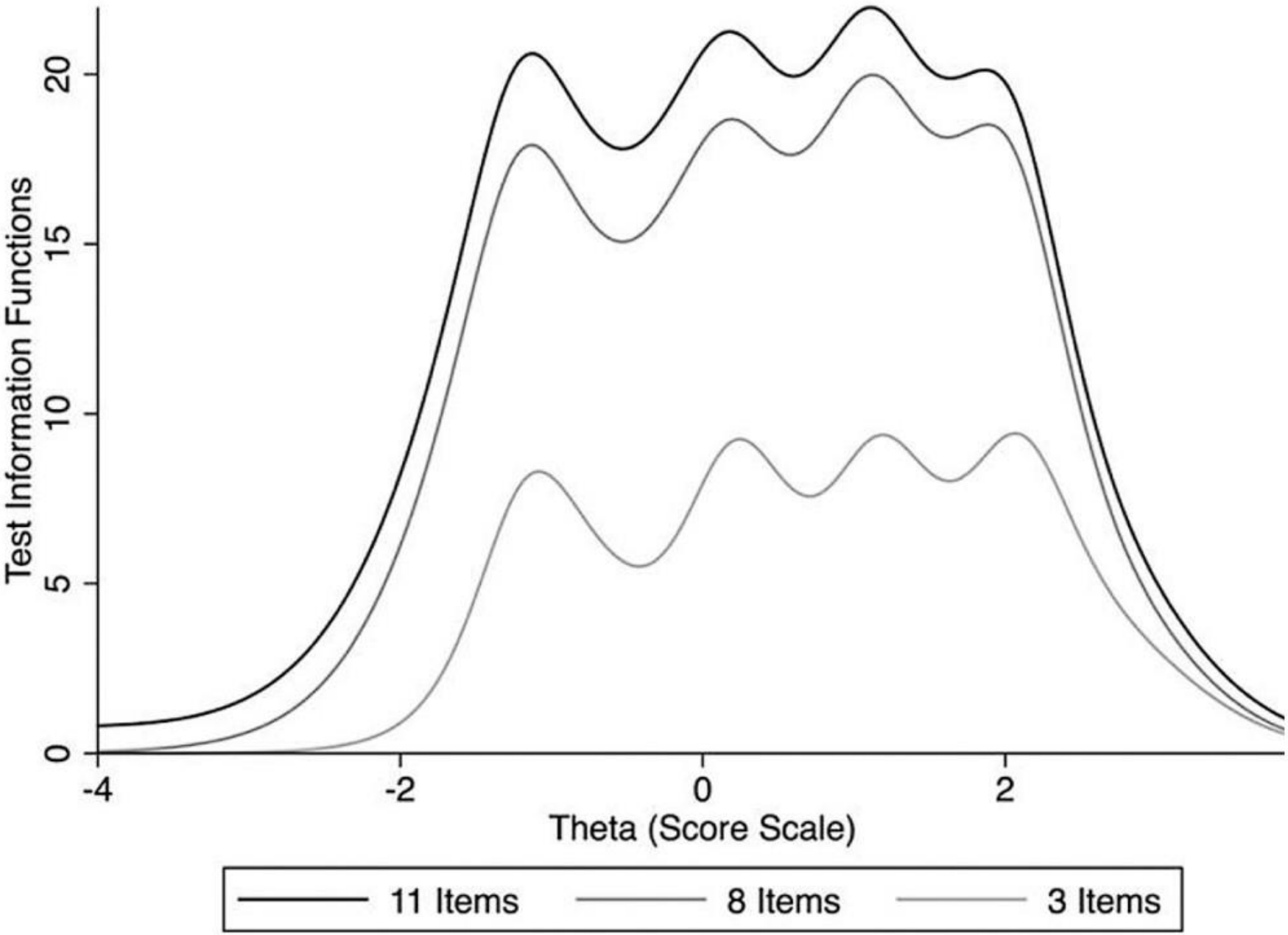


This "U-shaped" IRT CSEM above contrasts with the CSEM for simple sum scoring or %-correct scoring . If conventional scores are a proportion (a simplification), the error is binomial $\sqrt{\frac{\phi(1-\phi)}{n_i}}$. Is conventional error greatest for central or extreme scores?

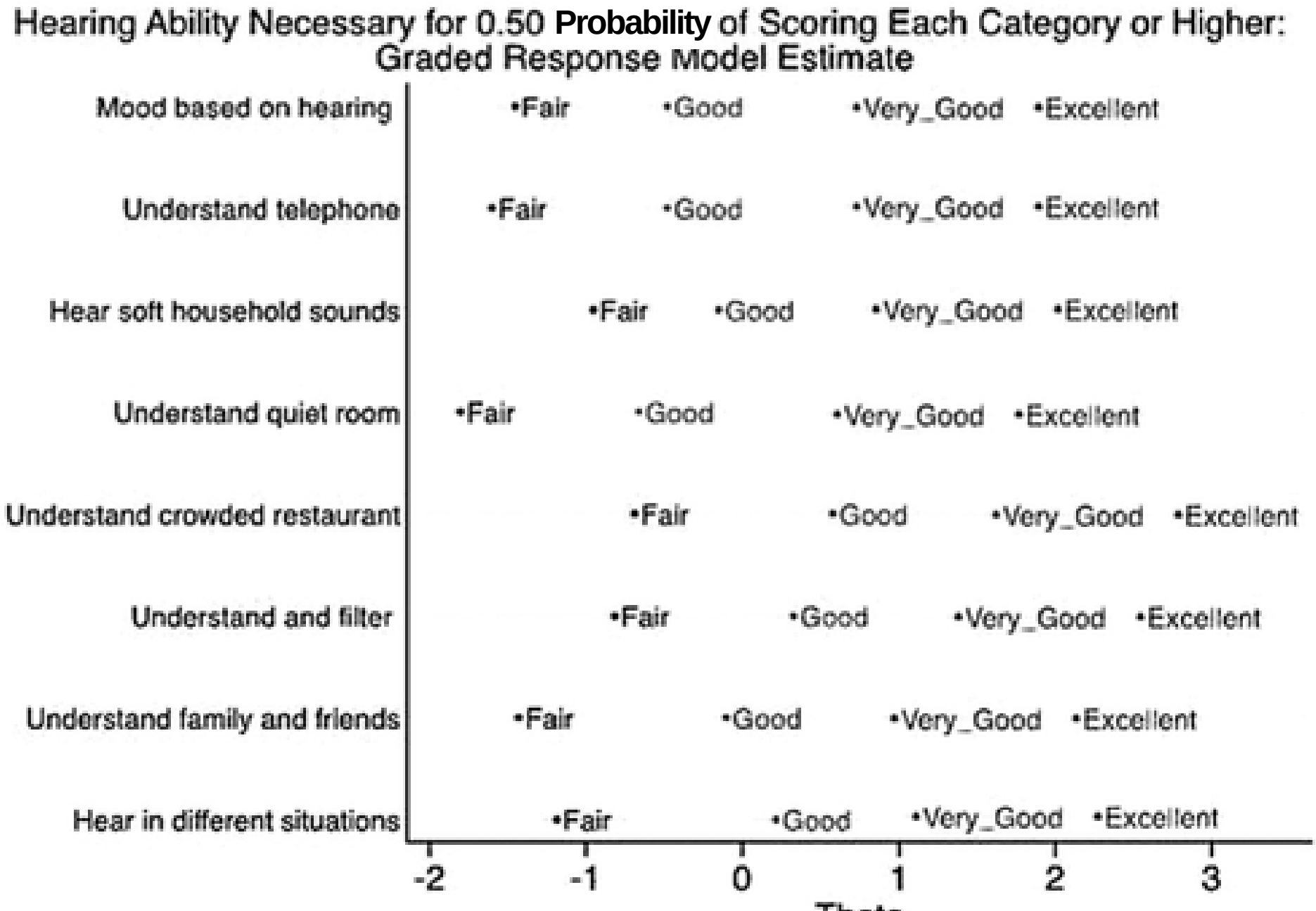
Step 7: Efficient Measurement - Item Information



Step 7: Efficient Measurement – Test Information



Step 7: Efficient Measurement - Item Maps



Step 8: Correlation and Prediction

Table 3. Correlation Coefficients for Transformed Variables in Analysis 1.[a]

	WRS′		PTA′′	
	Left (n = 54)	Right (n = 52)	Left (n = 53)	Right (n = 55)
3-item scale	0.29	0.45	0.31	0.35
95% CI	0.03-0.52	0.11-0.58	0.15-0.60	0.13-0.59
6-item scale	0.37	0.50	0.40	0.42
95% CI	0.20-0.64	0.26-0.68	0.24-0.66	0.25-0.67
8-item scale	0.40	0.48	0.42	0.43
95% CI	0.05-0.54	0.15-0.61	0.16-0.62	0.14-0.60
11-item scale	0.39	0.49	0.39	0.43
95% CI	0.10-0.57	0.18-0.62	0.19-0.63	0.18-0.62

[a]All correlations were significant at $P < .05$. WRS′: transformed word recognition scores, $x' = -\log(102 - x)$. Higher scores indicate higher levels of auditory effectiveness. PTA′′: pure tone audiometry, $x = \log(x)$. Higher scores indicate lower levels of auditory effectiveness.

An 8-Step Plan

Step 1: Content and Cognition

Step 2: Scoring and Scaling

Step 3: Correlation and Reliability

Step 4: Classical Item Diagnostics

Step 5: Latent Structure Analysis

Step 6: Item Response Theory (IRT)

Step 7: IRT for Efficient Measurement

Step 8: Correlation and Prediction

How do we "develop and validate" a measure?

We say, "*$\boldsymbol{X}$ is important. No one thinks of $\boldsymbol{X}$. Existing measures of $\boldsymbol{X}$ are off the mark. $\boldsymbol{X}$ matters more than everything else. If only we paid attention to $\boldsymbol{X}$.*" If that is your argument, I suggest this research agenda:

1. Establish the **theoretical construct**
   - This measure should exist.
2. Establish the **latent structure**
   - This components of the measure relate as expected.
3. Establish **reliability**
   - The score you estimate should be precise.
4. Establish **predictions** and **intercorrelations**
   - These scores predict outcomes it should. They also predict outcomes things better than, over and above, other scores.
5. Establish **usefulness**
   - Using these scores achieves the intended purposes.