Latent Trait Measurement Models for Binary Responses: Part 1

- Topics:
  - The Big Picture of Latent Trait Measurement Models
  - Review of Regression Models for Binary Outcomes
  - 1, 2, 3, and 4 Parameter IRT (and Rasch) Models
  - Item and Test Information (for Indexing Reliability)

Reviewing the Big Picture… of CTT

- **CTT** predicts the total: $\mathbf{Ytotal_s = TrueScore_s + error_s}$
  - Items are assumed exchangeable because their properties are not part of the model for creating a latent trait estimate (as total)
  - **Because the sum score serves AS the latent trait estimate**, it can be problematic to make comparisons across different forms
    - Item difficulty = mean of item (is sample-dependent)
    - Item discrimination = item–remainder correlation (is sample-dependent)
  - Estimates of reliability assume (without testing) unidimensionality; also tau-equivalence (alpha) or parallel items (Spearman-Brown)
    - Measurement error is (most often) assumed *constant* across the trait

- How do you make your instrument better?
  - Get more items. What kind of items? More.

Reviewing the Big Picture… of CFA

- **CFA predicts the ITEM response: $y_{is} = \mu_i + \lambda_i F_s + e_{is}$**
  - **<u>Linear</u> regression relating continuous item response to latent predictor F_s**
  - Both items AND subjects matter in predicting item responses
    - Item difficulty = intercept μ_i (in theory, sample independent)
    - Item discrimination = factor loading λ_i (in theory, sample independent)
  - The goal of the factors is to recreate the observed covariances among items, so **factors represent testable assumptions** about the pattern of item covariance
    - Responses should be unrelated after controlling for factors → local independence
    - But if not, error covariances could be used to capture unexpected multidimensionality!

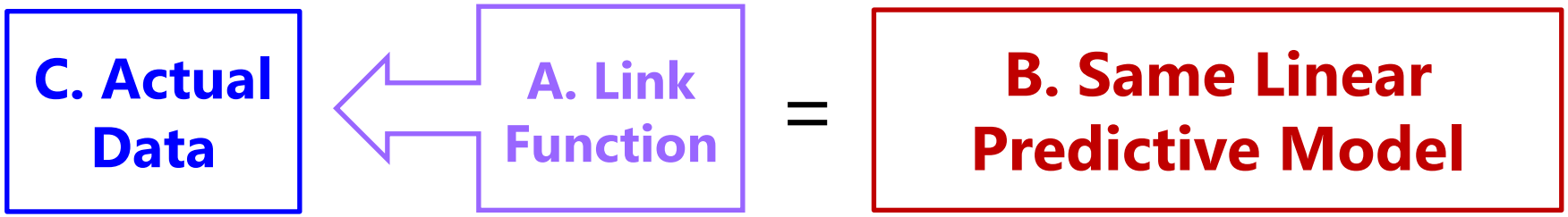
- **Because individual item responses are included:**
  - Items can vary in discrimination (→ Omega sum score reliability) and difficulty
  - To make your instrument better, you need MORE and BETTER items…
    - With higher standardized factor loadings → with greater information = λ_i^2 / **Var**(e_i)
- Measurement error is still assumed constant across the latent trait (one value)

From CFA to IRT and IFA

Outcome Type → ***Model Family Name***	**Observed Predictor x_i**	**Latent Predictor x_i**
Continuous outcomes → *"General Linear Model"*	"Linear" Regression	Confirmatory Factor Models
Discrete/categorical outcomes → *"Generalized Linear Model"*	Logistic/Probit/ Multinomial Regression	Item Response Theory and Item Factor Analysis

- Basis of Item Response Theory (IRT) and Item Factor Analysis (IFA) lies in models for discrete outcomes, which are called "**general*ized***" linear models
- Thus, IRT and IFA will be easier to understand after reviewing concepts from general*ized* linear models…
  - For more, see Lecture 2 and Examples 2a and 2b from this class

3 Parts of Generalized Linear Models



A. **Link Function**: Transformation of *conditional mean* to keep predicted outcomes within the bounds of the outcome

B. **Same Linear Model**: How the model linearly predicts the *link-transformed* conditional mean of the outcome

C. **Conditional Distribution**: How the outcome could be distributed given the possible values of the outcome

Generalized linear models work for many kinds of outcomes…

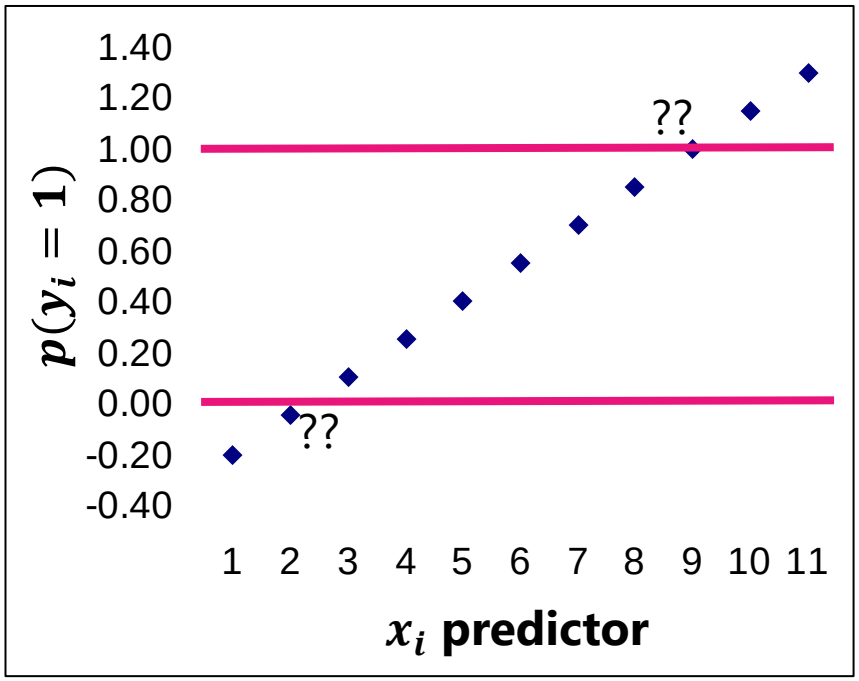
Here's how it works for binary outcomes

- Let's say we have a single binary (0 or 1) outcome… y_i *(i=person)*
- The mean of a binary outcome is the proportion of 1 values
  - So given each person's predictor values, the model tries to predict the **conditional mean**: the **probability of having a 1**: $\boldsymbol{p(y_i = 1)}$
    - The conditional mean has more possible values than the outcome!
  - General linear model: $\boldsymbol{p(y_i = 1) = \beta_0 + \beta_1(x_i) + e_i}$
    - $\boldsymbol{\beta_0}$ = expected probability of $\boldsymbol{y_i = 1}$ when all predictors = 0
    - $\boldsymbol{\beta_1}$ = expected change in $\boldsymbol{p(y_i = 1)}$ for per unit change in $\boldsymbol{x_i}$
    - $\boldsymbol{e_i}$ = difference between observed minus predicted **binary** values
  - Model becomes $\boldsymbol{y_i}$ = **(predicted probability of 1)** + $\boldsymbol{e_i}$
  - **What could possibly go wrong???**

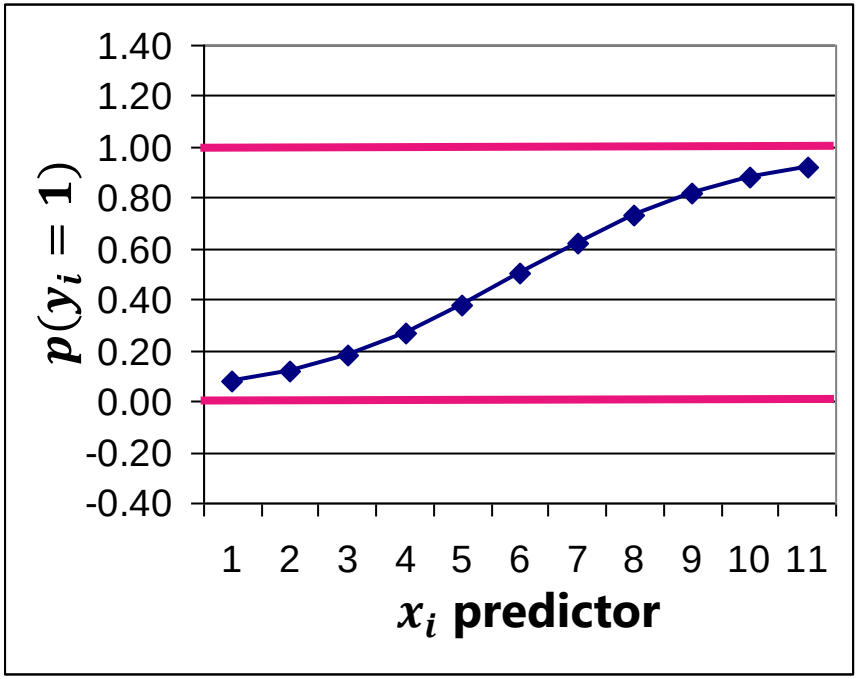
Normal GLM for Binary Outcomes?

- Problem #1: A **linear relationship** between x_i and y_i???
- Probability of a 1 is bounded between 0 and 1, but predicted probabilities from a linear model aren't going to be bounded
- Linear relationship needs to shut off → made nonlinear

We have this…

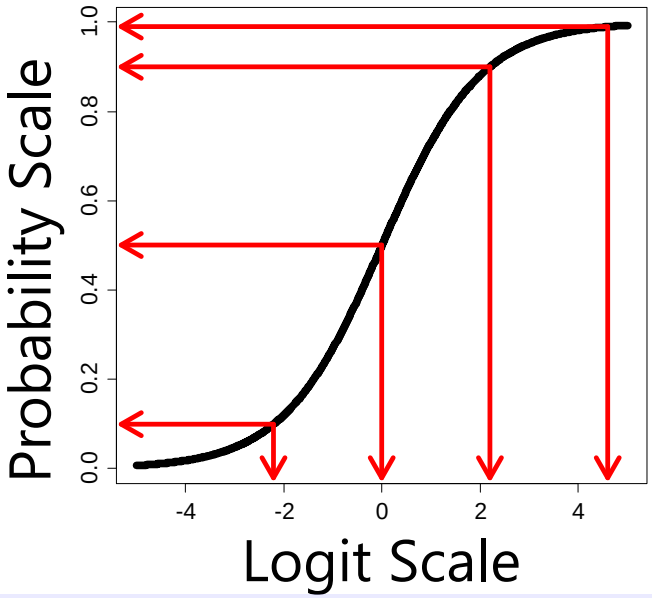


But we need this…



Generalized Models for Binary Outcomes

- Solution to #1: Rather than predicting $\boldsymbol{p(y_i = 1)}$ directly, the model transforms it into an unbounded variable using a **link function**:
  - Transform **probability** into **odds**: $\frac{p_i}{1-p_i} = \frac{\text{prob}(y_i=1)}{\text{prob}(y_i=0)}$
    - If $p(y_i = 1) = .7$ then $\text{Odds}(1) = 2.33$; $\text{Odds}(0) = 0.429$
    - But odds scale is skewed, asymmetric, and ranges 0 to +∞ → Not a good outcome!
  - **Take *natural log of odds* → called "logit" link:** $\textbf{Log}\left[\frac{p_i}{1-p_i}\right]$
    - If $p(y_i = 1) = .7$, then $\text{Logit}(1) = 0.846$; $\text{Logit}(0) = -0.846$
    - Logit scale is now symmetric about 0, range is ±∞ → Now a good outcome to predict!



Probability → "data scale"	Logit → "model scale"
0.99	4.6
0.90	2.2
0.50	0.0
0.10	−2.2

Can you guess what $p(.01)$ would be on the logit scale?

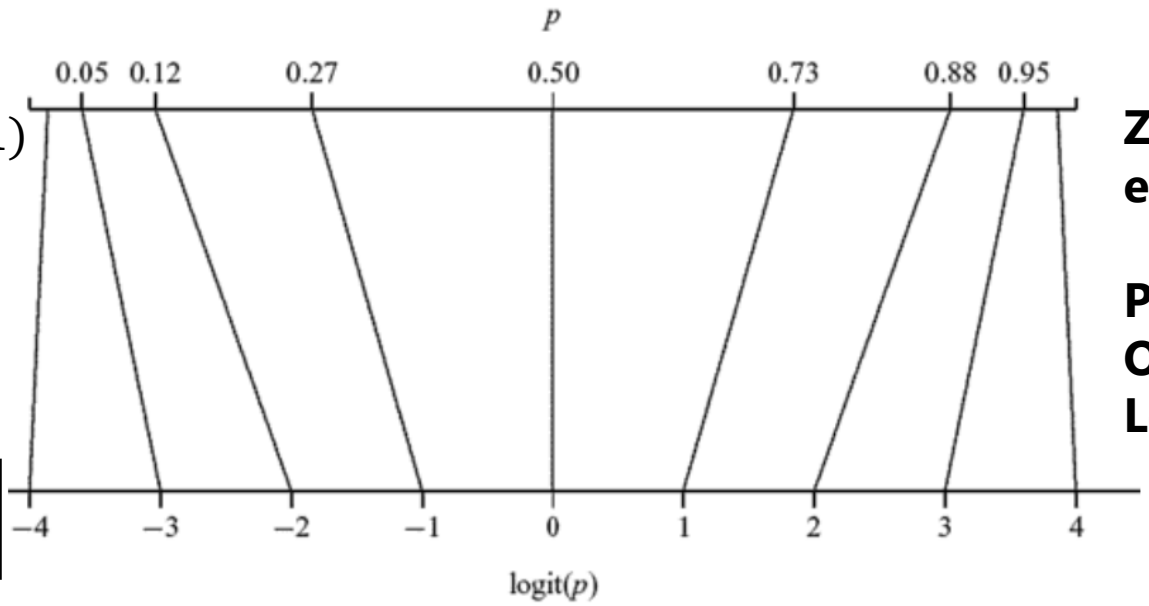
Solution to #1: Probability to Logits

- **A Logit link is a <u>nonlinear</u> transformation of probability:**
  - Equal intervals in logits are NOT equal intervals of probability
  - Logits range from ±∞ and are symmetric about prob = .5 (→ logit = 0)
  - Now we can use a linear model → The model will **linearly predict the expected logit**, which translates into a nonlinear prediction of probability → **the outcome conditional mean (probability) shuts off at 0 or 1 as needed**

<u>Probability:</u>
$p_i = p(y_i = 1)$

<u>Odds:</u> $\left[\frac{p_i}{1-p_i}\right]$

<u>Logit (log odds):</u>
$\text{Log}\left[\frac{p_i}{1-p_i}\right]$



Zero-point on each scale:

Prob = .5
Odds = 1
Logit = 0

Image borrowed from Figure 17.3 of: Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Normal GLM for Binary Outcomes?

- General linear model: $\boldsymbol{p(y_i = 1) = \beta_0 + \beta_1(x_i) + e_i}$
- If $\boldsymbol{y_i}$ is binary, then $\boldsymbol{e_i}$ can only be 2 things: $\boldsymbol{e_i = y_i - \hat{y}_i}$
  - If $\boldsymbol{y_i = 0}$ then $\boldsymbol{e_i}$ = (0 – predicted probability)
  - If $\boldsymbol{y_i = 1}$ then $\boldsymbol{e_i}$ = (1 – predicted probability)
- Problem #2a: So the residuals can't be normally distributed
- Problem #2b: The residual variance can't be constant over $\boldsymbol{\hat{y}_i}$ as in GLM because the **mean and variance are dependent**
  - Variance of binary variable: $\boldsymbol{Var(y_i) = p_i * (1 - p_i)}$

Mean and Variance of a Binary Variable

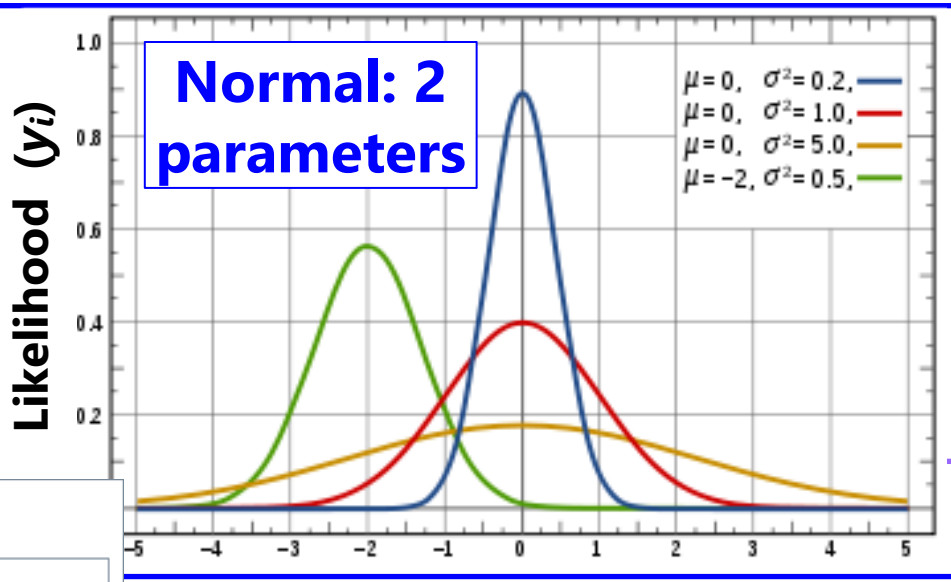
Mean (p_i)	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

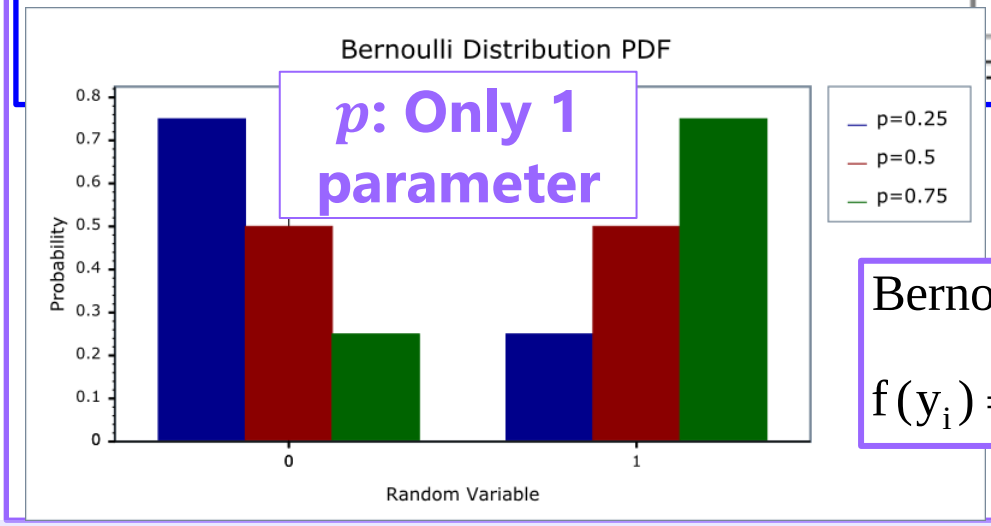
Solution to #2: Bernoulli Distribution

- Rather than using a **normal conditional distribution** for the outcome, we will use a **Bernoulli conditional distribution**

Univariate Normal PDF:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left[-\frac{1}{2} * \frac{(y_i - \mu)^2}{\sigma^2} \right]$$





Bernoulli PDF:

$$f(y_i) = (p)^{y_i} (1-p)^{1-y_i}$$

= $p(1)$ if y_i=1, $p(0)$ if y_i=0

Top image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution
Bottom image borrowed from: https://www.boost.org/doc/libs/1_70_0/libs/math/doc/html/math_toolkit/dist_ref/dists/bernoulli_dist.html

Predicted Binary Outcomes

- **Logit:** $\mathbf{Log}\left[\frac{p(y_i=1)}{p(y_i=0)}\right] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(x_i)$ ← **g(·) link**
  - Predictor effects are linear and additive like in regular regression, but $\boldsymbol{\beta}$ fixed effects describe changes to predicted **logit**

- **Odds:** $\left[\frac{p(y_i=1)}{p(y_i=0)}\right] = \exp(\boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_i)$

- **Probability:** $p(y_i = 1) = \frac{\exp(\boldsymbol{\beta_0}+\boldsymbol{\beta_1} x_i)}{1+\exp(\boldsymbol{\beta_0}+\boldsymbol{\beta_1} x_i)}$ ← **$g^{-1}(\cdot)$ inverse link**

  or equivalently $p(y_i = 1) = \frac{1}{1+\exp[-1(\boldsymbol{\beta_0}+\boldsymbol{\beta_1} x_i)]}$

- **Is "logistic regression" when using an observed x_i predictor; Is IRT/IFA when using a latent factor as the x_i predictor**
- Foreshadowing: IRT models are usually described using the probability formula, whereas IFA models use the logit formula

Converting Across the 3 Scales

- e.g., for $\mathbf{Log}\left[\frac{p(y_i=1)}{p(y_i=0)}\right] = \hat{y}_i = \beta_0 + \beta_1(x_i)$

Direction	Conditional Mean	Slope for x_i
Predicted logit outcome (i.e., given by "the **link**"):	$\hat{y}_i$	β_1
From logits to odds (or odds ratios for effect sizes):	Odds: $\exp(\hat{y}_i)$	Odds *ratio*: $\exp(\beta_1)$
From logits to probability (i.e., by the "**inverse link**"):	$\frac{\exp(\hat{y}_i)}{1+\exp(\hat{y}_i)}$	***Doesn't make any sense!***

- You can unlogit the model-predicted conditional mean all the way back into probability to express predicted outcomes, but **you can only unlogit the slopes back into odds ratios** (not all the way back to changes in probability)
- Order of operations: build predicted logit outcome, then logit → probability

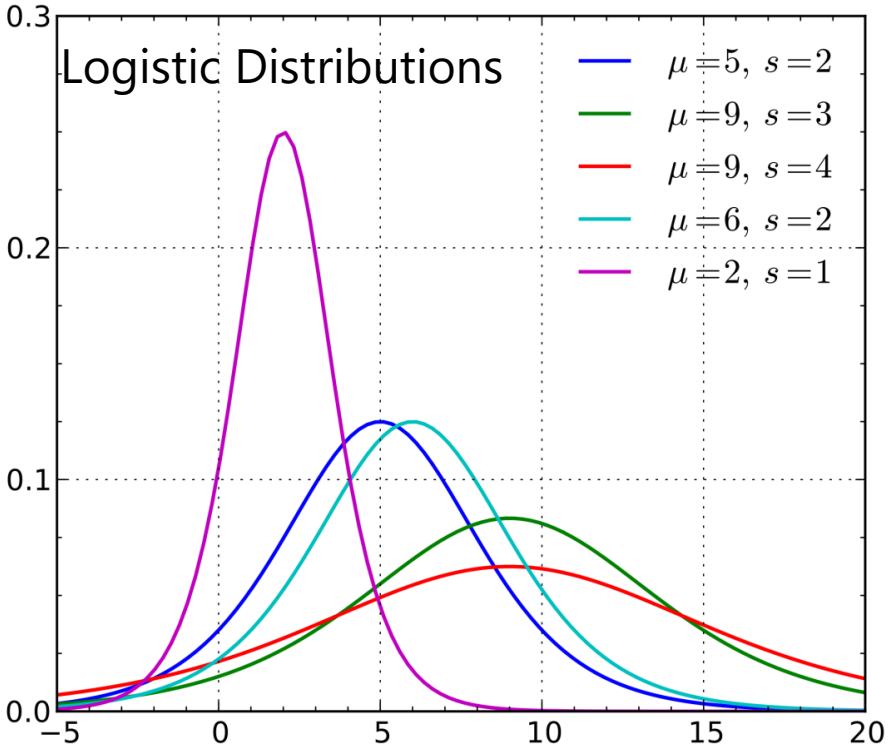
"Latent Responses" for Binary Data

- This model is sometimes expressed by calling the logit($\boldsymbol{y_i}$) an **underlying continuous ("latent") response of $\boldsymbol{y_i^*}$** instead:

Empty Model: $\boldsymbol{y_i^* = -threshold + e_i^*}$

$\boldsymbol{threshold = intercept\ \beta_0 * -1}$

  - In which $\boldsymbol{y_i = 1}$ if $(y_i^* > threshold)$, or $\boldsymbol{y_i = 0}$ if $(y_i^* \leq threshold)$



So **when predicting $\boldsymbol{y_i^*}$,**
$\boldsymbol{e_i^* \sim}$ Logistic$\left(0, \boldsymbol{\sigma^2_{e^*} = 3.29}\right)$

Logistic Distribution:

Mean = μ, Variance = $\frac{\boldsymbol{\pi^2}}{\boldsymbol{3}} s^2$,
where s = scale factor that allows for "over-dispersion" (must be fixed to 1 for binary responses for identification)

Image borrowed from: https://en.wikipedia.org/wiki/Logistic_distribution

Other Link Functions for Binary Data

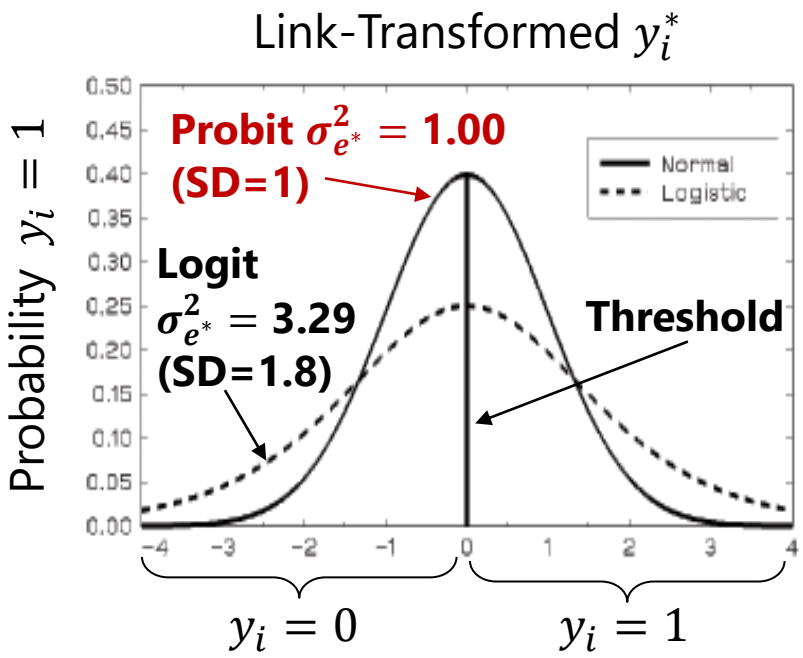
- The idea that a "latent" continuous variable underlies an observed binary response also appears in a "**Probit Regression**" model:

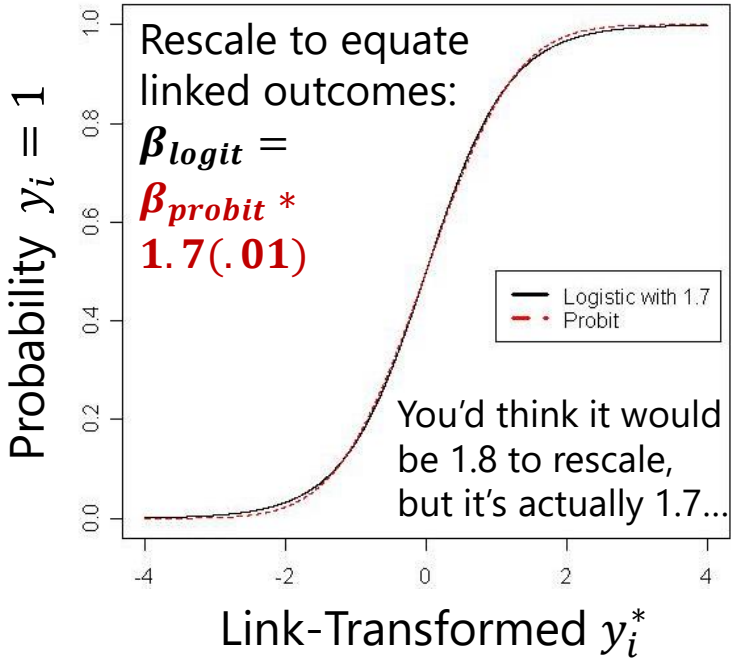
  - A ***probit*** link, such that now your model predicts a different transformed $\boldsymbol{y_i}$:
    $$\text{Probit}(y_i = 1) = \Phi^{-1}[p(y_i = 1)] = linear\ predictor$$ ← **g(·) link**
    - Φ = standard normal cumulative distribution function, so the link-transformed $\boldsymbol{y_i}$ **is the z-value** that corresponds to the location on standard normal curve **below** which the conditional mean probability is found (i.e., z-value for area to the left)
    - Requires integration to inverse link from probits to predicted probabilities

  - Same Bernoulli distribution for the conditional binary outcomes, in which residual variance is not separately estimated (so no e_i predicting original y_i)
    - Model scale: Probit can also predict "latent" response: $\boldsymbol{y_i^* = -threshold + e_i^*}$
    - But Probit says $\boldsymbol{e_i^* \sim Normal(0, \sigma^2_{e^*} = 1.00)}$, whereas logit $\boldsymbol{\sigma^2_{e^*} = \frac{\pi^2}{3} = 3.29}$

  - So given this difference in variance, probit coefficients are on a different scale than logit coefficients, and so their estimates won't match… however…

Probit vs. Logit: Should you care? Pry not.





- Other fun facts about probit:
  - **Probit** = "**ogive**" in the Item Response Theory (IRT) world
  - Probit has no odds ratios (because it's not based on odds)
  - Probit is the **only** option in IFA models using **limited-information estimation**!

- Both logit and probit assume **symmetry** of probability curve, but there are other *asymmetric* options as well: (complementary) log-log

Left image: exact source now unknown, but I think it was from Don Hedeker
Right image: borrowed from Jonathan Templin

How IRT/IFA are the same as CFA

- **NOW BACK TO YOUR REGULARLY SCHEDULED MEASUREMENT CLASS**
- **IRT/IFA** = confirmatory measurement model in which latent traits are the model predictors (so you decide which items measure which traits)
  - Like CFA, **both items and subjects matter** because their properties are included in the measurement model (item difficulty and discrimination; subject F)
  - Item discrimination means the same thing in IRT and IFA, but they differ in how location of the item on the trait is indexed (item "difficulties" versus "thresholds")

- After controlling for a subject's latent trait value ($\boldsymbol{F}$ is now called **theta, $\boldsymbol{\theta}$**), the item responses should be uncorrelated (also called local independence)
  - The ONLY reason item responses are correlated is a (unidimensional) theta
  - Otherwise, we CAN fit confirmatory multidimensional factor models instead, and then responses are independent after controlling for ALL the thetas
  - As in CFA, can be violated by other types unaccounted for multidimensionality or dependency (e.g., "specific" method factors for common stems as "testlets")
    - Error covariances must be specified using method factors when using ML estimation

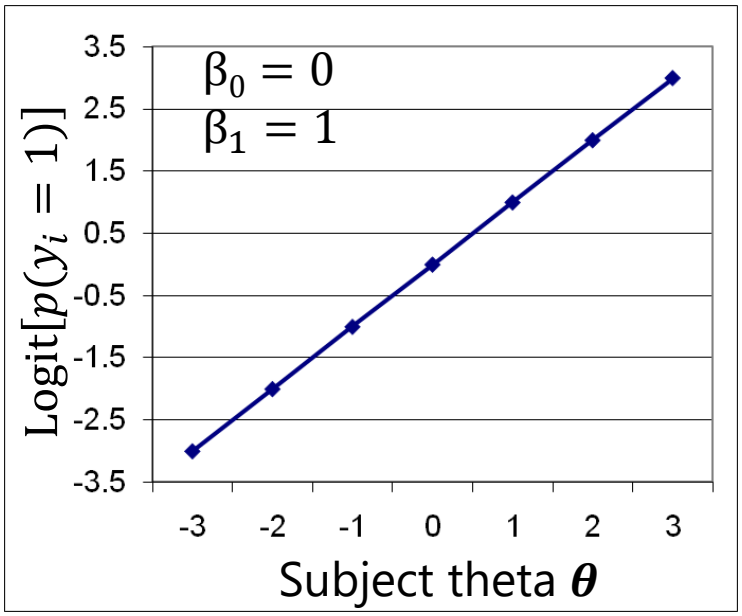
How IRT/IFA are *different from* CFA

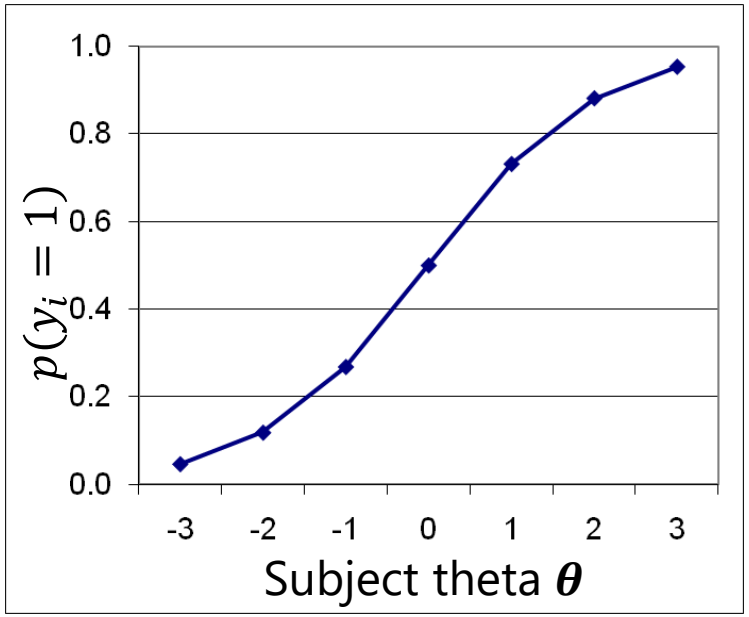
- IRT/IFA uses the same family of **link functions** (transformations) as in generalized models, it's just that the predictor is latent instead of observed
  - IRT/IFA = logistic/probit regression instead of linear regression
  - Predictor = Latent factor/trait in IRT/IFA = "**theta**" $\boldsymbol{\theta}$, and its slopes are still supposed to predict the associations of the item responses, just like in CFA

- **IRT/IFA specifies a <u>nonlinear</u> relationship between binary, ordinal, or nominal item responses and the latent trait (now called "theta" $\boldsymbol{\theta}$)**
  - Probability is bounded between 0 and 1, so the effect (slope) of theta must be nonlinear, so it will shut off towards the extremes of theta (as an S-shaped curve)
  - Errors cannot have constant variance across theta or be normally distributed
  - Full-information estimators use logit ($\sigma^2_{e^*} = 3.29$) or probit ($\sigma^2_{e^*} = 1.00$) link functions, but limited-information estimators only have probit ($\sigma^2_{e^*} = 1.00$)
    - **Logit = 1.7*Probit**, so the predicted probabilities are equivalent either way
    - Probit in IRT models is called "ogive" (as discussed in Embretson & Reise)

Nonlinear Prediction by θ in IRT/IFA

- The relationship between theta and the probability of response=1 is **"nonlinear" → a monotonic s-shaped logistic curve** whose shape and location are dictated by the estimated item parameters
  - **Linear** prediction of the **logit → nonlinear** prediction of **probability**





- Btw, it may be that other kinds of non-linear relationships could be more appropriate and thus fit better → These are "non-parametric" IRT models

Item Response Theory (IRT) = Item Factor Analysis (IFA) Models

Mplus can do ALL of these model/estimator combinations:	Model form: with **discrimination** and **difficulty** parameters	Model form: with **loadings** and **threshold** parms
Full-information estimation via Maximum Likelihood ("Marginal ML") → *uses original item responses*	**"IRT"** (Mplus gives only for binary responses)	**"?"** (Mplus gives for all models)
Limited-information estimation via Weighted Least Squares ("WLSMV") → *uses item response summary*	**"?"** (Mplus gives only for binary responses)	**"IFA"** (Mplus gives for all models)

- CFA assumes normally distributed, continuous item responses, but "**CFA** models for **categorical** responses" = **IRT** and **IFA** models

- These different names are used to reflect the combination of how the model is specified and how it is estimated, but it's the same core model
  - Btw, R Lavaan only has limited-information estimation for these models…

Model Format in IRT and IFA

- Item Factor Analysis (IFA) models look very similar to CFA, but Item Response Theory (IRT) models look quite different
- Partly due to predicting logits/probits (IFA) vs. probability (IRT):
  - **Logit:** $\mathbf{Log}\left[\frac{p(y_i=1)}{p(y_i=0)}\right] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_i}$
  - **Probability:** $p(y_i = 1) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$
- Partly due to different model parameterizations (stay tuned)
- The IFA and IRT model parameters are just re-arrangements of each other for common cases, but historically have been estimated differently (full vs. limited information) and for different purposes
  - Mplus provides both kinds of output for binary data, but only IFA output for categorical data (we will calculate IRT version)
- We'll start with IRT for binary responses, then move to IFA…
  - IRT parameterization is (arguably) more useful (and more direct → reliability)

Simplest IRT Model:
Rasch Model for Binary (0/1) Responses

- **Rasch model** as originally described (in which $\boldsymbol{\theta}$ **variance is estimated**):
  - **Logit**: $\text{Log}\left[\frac{p(y_{is}=1)}{p(y_{is}=0)}\right] = \theta_s - b_i$
  - **Probability**: $p(y_{is}=1) = \frac{\exp(\theta_s - b_i)}{1+\exp(\theta_s - b_i)}$

y_{is} is 0 or 1 response to item i for subject s

1.7 may go inside exp( ) if predicting logits so model parms stay in probit scale

  - $\boldsymbol{\theta_s}$ **= subject trait** → most likely latent trait score (**theta**, a **random effect**) for subject *s* given their pattern of item responses
  - $\boldsymbol{b_i}$ **= "item difficulty"** → **location** on latent trait (estimated as a **fixed effect**) (like an intercept, but it's actually "difficulty" now!)

Random = parameter has distribution; Fixed = no distribution

- Probability of $y_{is} = 1$ depends on subject trait (theta) vs. item difficulty:
  - If trait > difficulty, then logit > 0, and probability > .50
  - If difficulty > trait, then logit < 0, and probability < .50

Fundamentals of IRT

- **Back in CTT**, scores only have meaning relative to the persons in the same sample, and thus **sample norms** are needed to interpret a person's score
  - "I got a 12. Is that good?"
    "Well, that puts you into the 90th percentile."
    "Great!"
  - "I got a 12. Is that good?"
    "Well, that puts you into the 10th percentile."
    "Doh!"
  - Same score in both cases, but different reference groups!

- **In IRT**, the properties of items and subjects are placed along the same underlying latent continuum= "**conjoint scaling**"
  - This concept can be illustrated using **construct maps that order both subjects** in their trait levels **and items** in their difficulty/severity…

A Latent Continuum of Pokémon Knowledge

Trait theta θ_s has meaning based on the items *at that location*: subject theta = item difficulty at which $p(y_{is} = 1) = .50$

Subject Trait Side

(Expert) Hugh

X

Cass

(Starter) Bladimir

Daddy

X

Mommy

(Novice) Grandma

Item Difficulty Side

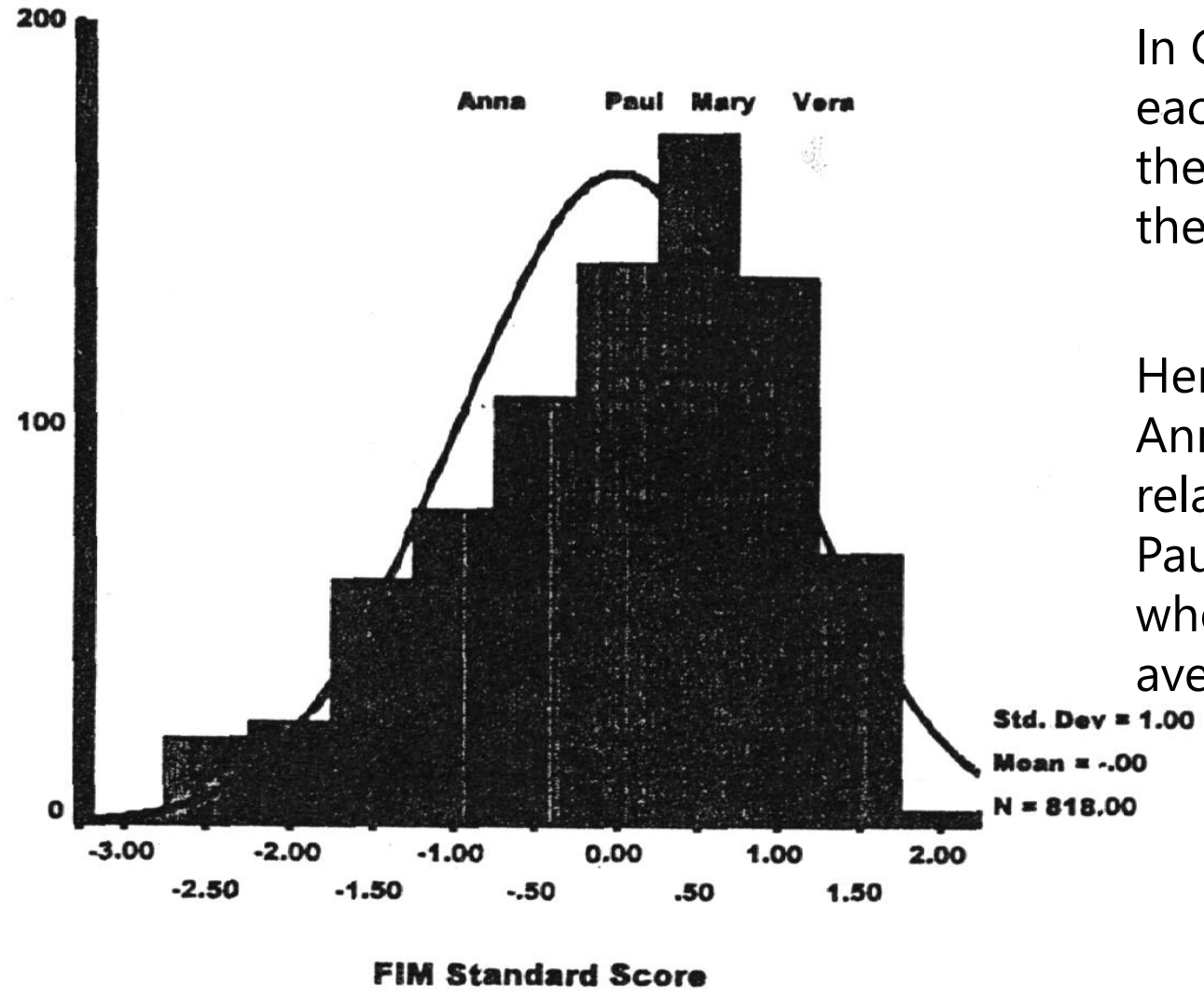







All images borrowed from The Google

Norm-Referenced Measurement in CTT

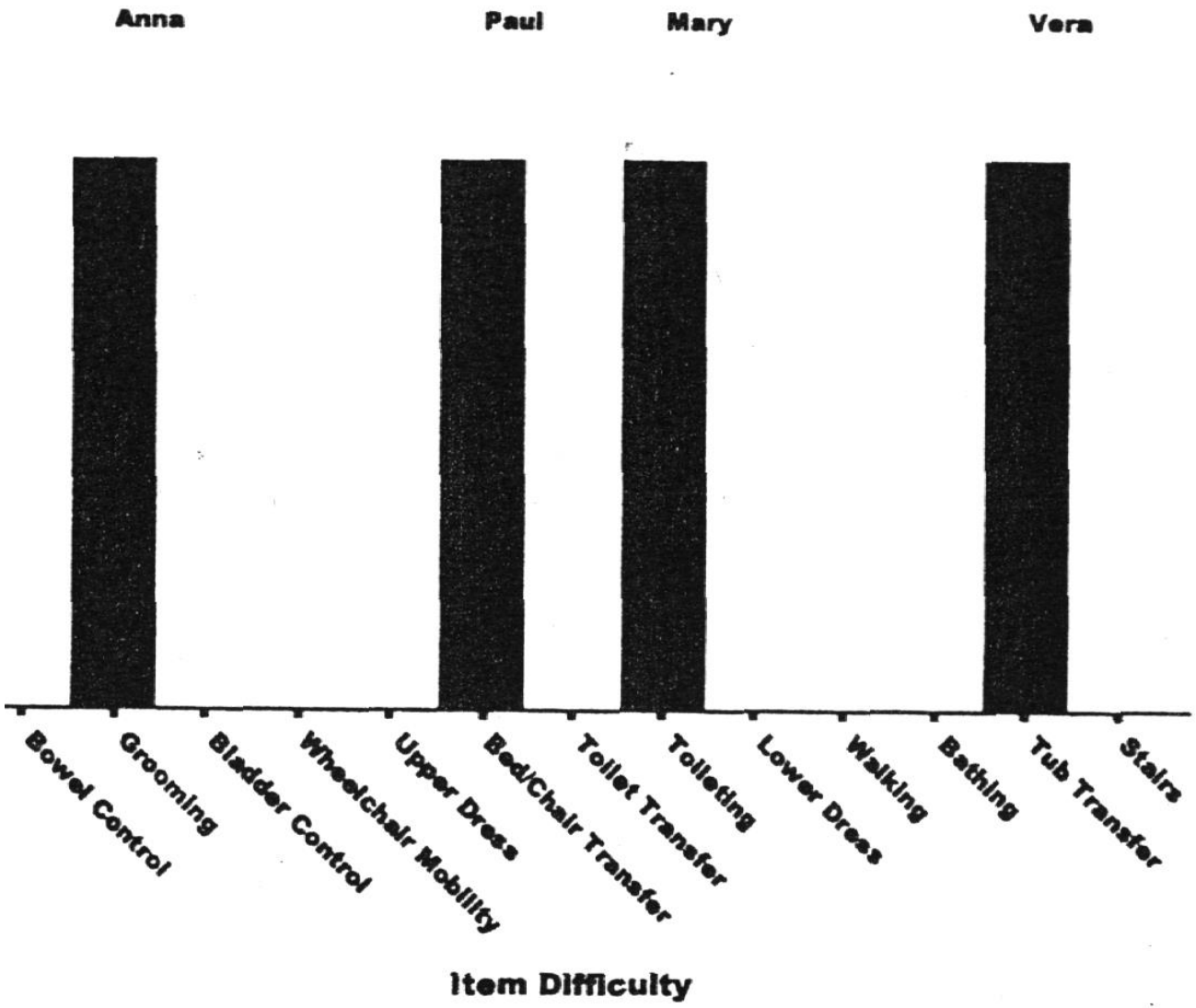


In CTT, the ability level of each subject is relative to the abilities of the rest of the test sample

Here, we would say that Anna is functioning relatively worse than Paul, Mary, and Vera, who are each above average (which is 0)

Image from Embretson & Reise (2000)

Item-Referenced Measurement in IRT



Each subject's theta level reflects the type of activity they are predicted to be able to do on their own with prob = **.50**

Given theta, the model can predict the probability of accomplishing each task (even if not administered!)

Image from Embretson & Reise (2000)

Interpretation of Theta Latent Traits

- **Theta estimates are 'sample-free' and 'scale-free'**
  - Theta estimate does not depend on who else was measured
  - Theta estimate does not depend on which items were given
    - AFTER calibrating all items to same metric, can get a subject's location on latent metric regardless of which *particular* items were given

- However: although the theta estimate does not depend on the particular items given, its ***standard error*** does
  - Extreme thetas without many items of comparable difficulty will not be estimated that well → large SE (flat likelihood)
  - Likewise, items of extreme difficulty without many subjects of comparable traits will not be estimated that well → large SE

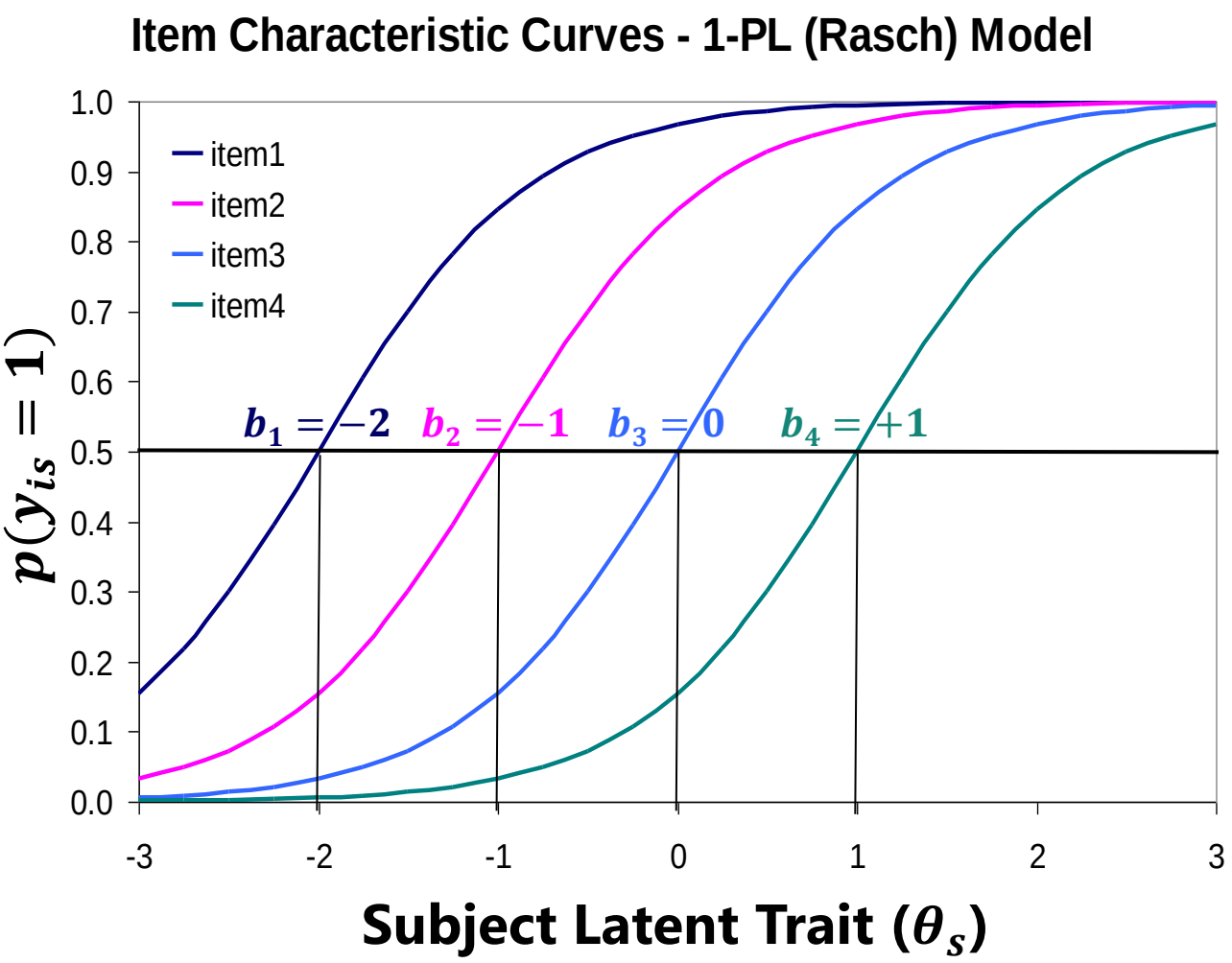
Another version: The 1PL Model

- **The "Rasch" model is a rescaled version of the One-Parameter Logistic IRT model → "1PL"**
  - **Logit:** $\text{Log}\left[\frac{p(y_{is}=1)}{p(y_{is}=0)}\right] = a(\theta_s - b_i)$
  - **Probability:** $p(y_{is}=1) = \frac{\exp[a(\theta_s - b_i)]}{1+\exp[a(\theta_s - b_i)]}$

In the "Rasch" model, a is fixed = 1 while theta's variance is estimated; in the 1PL, a is estimated and theta's variance is fixed = 1 (and optional 1.7 → probit)

  - $\boldsymbol{a}$ **= "item discrimination"** = relation of item to latent trait = slope of curve at probability = .50 (at inflection, its max slope) = **fixed effect**
  - The 1-PL model has "$\boldsymbol{a}$" and not "$\boldsymbol{a_i}$" – that's because $\boldsymbol{a}$ is assumed constant across items (and thus, the 1 parameter that is estimated for each item is still difficulty $\boldsymbol{b_i}$ as a **fixed effect** (no distribution)
  - If using the probit link function, the predicted outcome is the z-score for the area to the left under the normal curve for that predicted probability

1-PL (→Rasch) Model Predictions



b_i = item difficulty location on latent trait at which probability = .50

a = discrimination slope at prob = .50, (logit = 0, which is point of inflection)

Note: **equal a terms** means the ICCs will never cross → this is "Specific Objectivity"

Can you guess what's next?
2-Parameter Logistic Model (2PL)

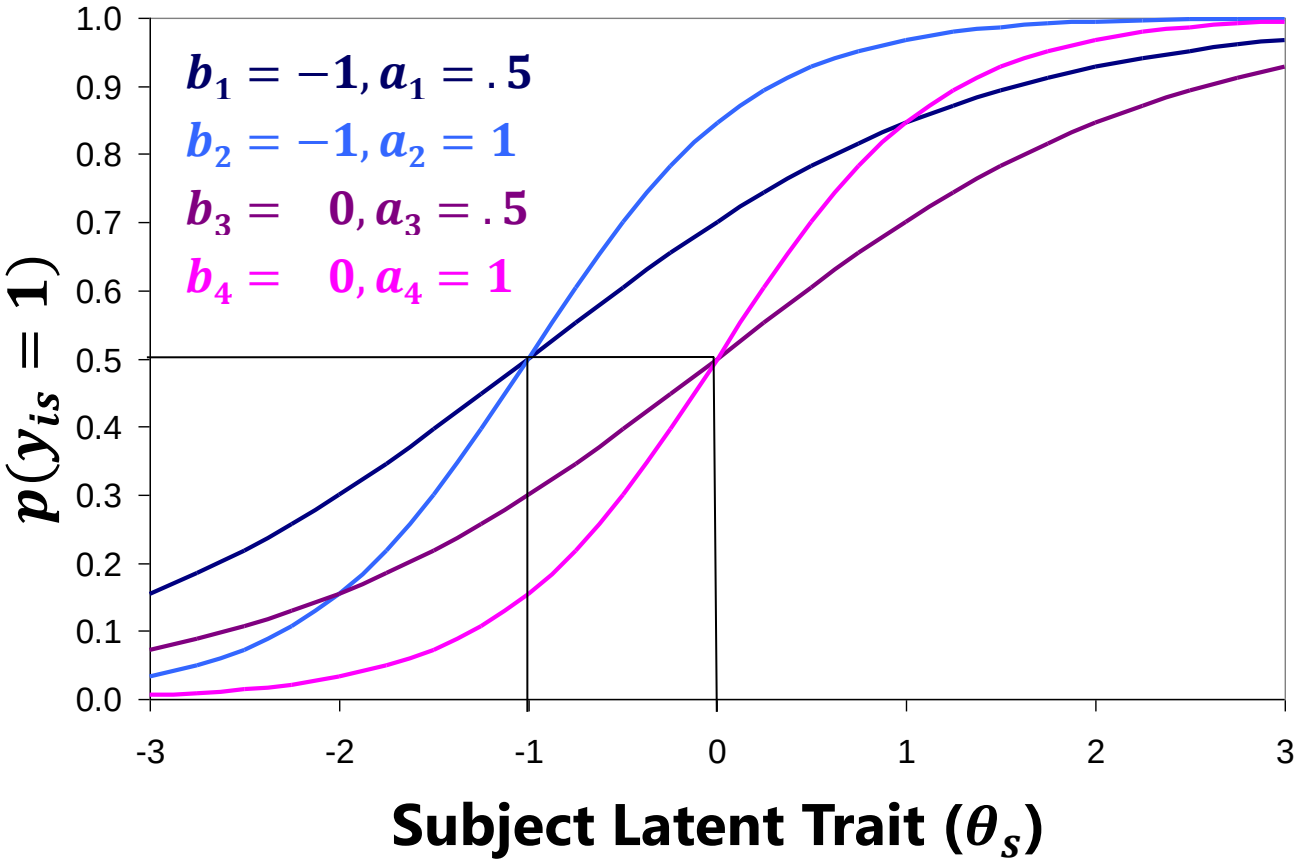
- The **1-PL (Rasch)** model assumes tau-equivalence → **equal discrimination**
- The **2-PL frees this constraint** by changing "a" to "a_i" (as **fixed effects**):
  - **Logit:** $\text{Log}\left[\frac{p(y_{is}=1)}{p(y_{is}=0)}\right] = a_i(\theta_s - b_i)$
  - **Probability:** $p(y_{is}=1) = \frac{\exp[a_i(\theta_s - b_i)]}{1+\exp[a_i(\theta_s - b_i)]}$

Relative to a logit-link model, parameters from a **probit-link** (ogive) model will be smaller by a factor of ~1.7

  - a_i = **"item discrimination"** = relation of **each item** to latent trait
    = slope of curve at probability = .50 (at inflection, its max slope)
  - b_i is still item difficulty (location where probability = .50)
  - Note that a_i is a **linear** slope for theta θ predicting the **logit of** $y_{is} = 1$ but a **nonlinear** slope for theta θ predicting the **probability of** $y_{is} = 1$

Item Characteristic Curves: 2PL Model

b_i = **difficulty** = location on latent trait at which $p_i = .50$ (or logit = 0)
a_i = **discrimination** slope at $p_i = .50$ (at the point of curve inflection)



Note: **unequal a_i**
→ curves cross
→ violates "**Specific Objectivity**"

At Theta $\theta_s = -1$:
Items **3** and **4** are harder than **1** and **2**
→ lower prob of 1

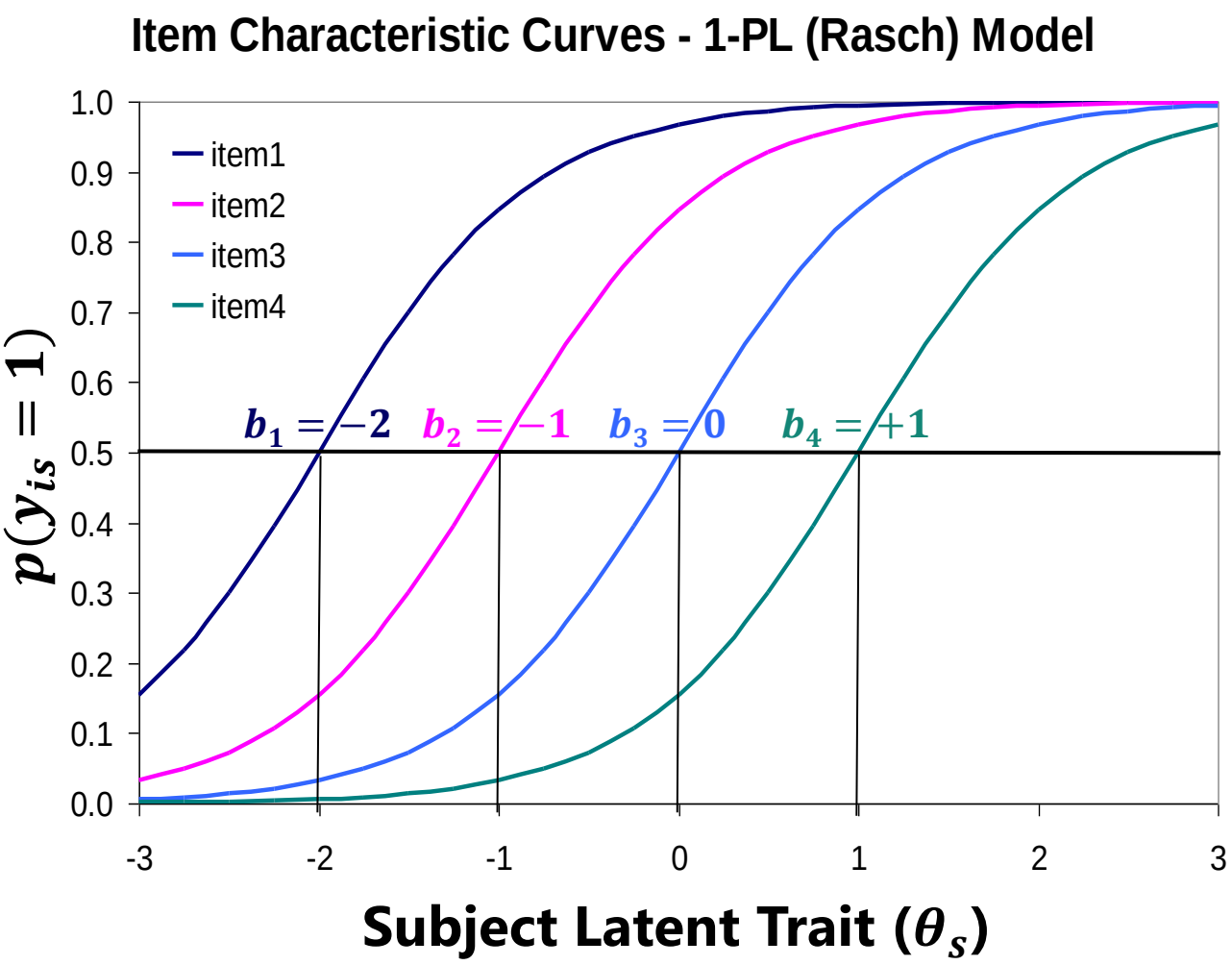
At Theta $\theta_s = +2$:
Item **1** is now harder than Item **4** → lower prob of 1

"IRT" vs. "Rasch"

- According to most **IRT** people, a "Rasch" model is just an IRT model with item discrimination $\boldsymbol{a_i}$ held equal across items (a tau-equivalent model)
  - Rasch = 1-PL where $\boldsymbol{b_i}$ item difficulty is the only item parameter
  - Slope = discrimination $\boldsymbol{a_i}$ = strength of relation of item to latent trait theta θ_s
  - In Rasch, $a = 1$ and theta variance = ?; In 1PL, $a =$? and theta variance = 1
  - *"Items may not be equally 'good', so why not just let their slopes vary?"*

- According to strict **Rasch** believers, the 2PL and rest of IRT are bananas
  - Rasch models have specific properties that are lost once you allow the item curves to cross (by using **item-varying $\boldsymbol{a_i}$**) → "**Loss of Specific Objectivity**"
    - Under the Rasch model, subjects are ordered the same in terms of predicted responses regardless of which item difficulty location you're looking at
    - Under the Rasch model, items are ordered the same in terms of predicted responses regardless of what level of subject theta you're looking at
    - **$\boldsymbol{a_i}$ item discrimination represents a theta*item interaction** → the item curves cross, so the ordering of subjects or items is no longer invariant, and this is "bad"
  - *"Items should not vary in discrimination if you know your construct!"*

1-PL (→Rasch) Model Predictions



b_i = item difficulty location on latent trait at which probability = .50

a = discrimination slope at prob = .50, (logit = 0, which is point of inflection)

Note: **equal a terms** means the ICCs will never cross → this maintains "**Specific Objectivity**"

2PL IRT vs. 1PL IRT (Rasch): What Goes into Theta

- In Rasch/1PL models, **the sum score is a "sufficient statistic"** for theta
  - For example, given 5 items ordered in difficulty from easiest to hardest, each of these response patterns where **3/5 are correct** would yield the **same estimate of theta**:

    1 1 1 0 0 (most consistent)
    0 1 1 1 0
    0 0 1 1 1
    1 0 1 0 1 (???)
    …. (and so forth)

- In 2PL (logit or probit) models, **items with higher discrimination (a_i) count more** towards theta (and theta SE will be lower with higher a_i items)
  - It not only matters **how many** items you got correct, but **which ones**
  - Rasch believers don't like this idea, because then the ordering of subjects on latent trait theta is dependent on the item properties

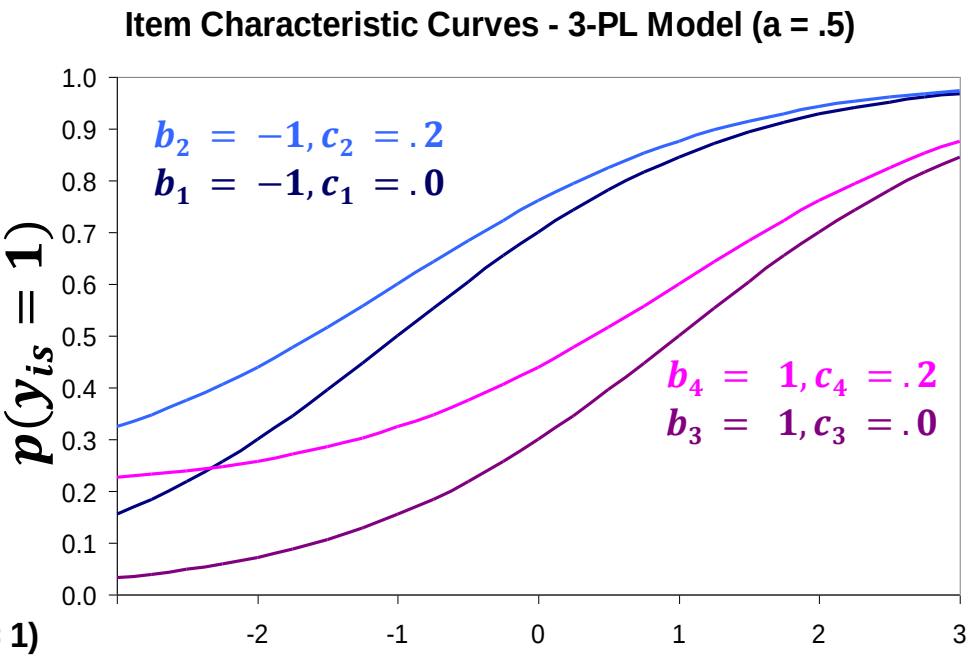
Yet Another Model for Binary Responses: 3-Parameter Logistic Model (3PL)

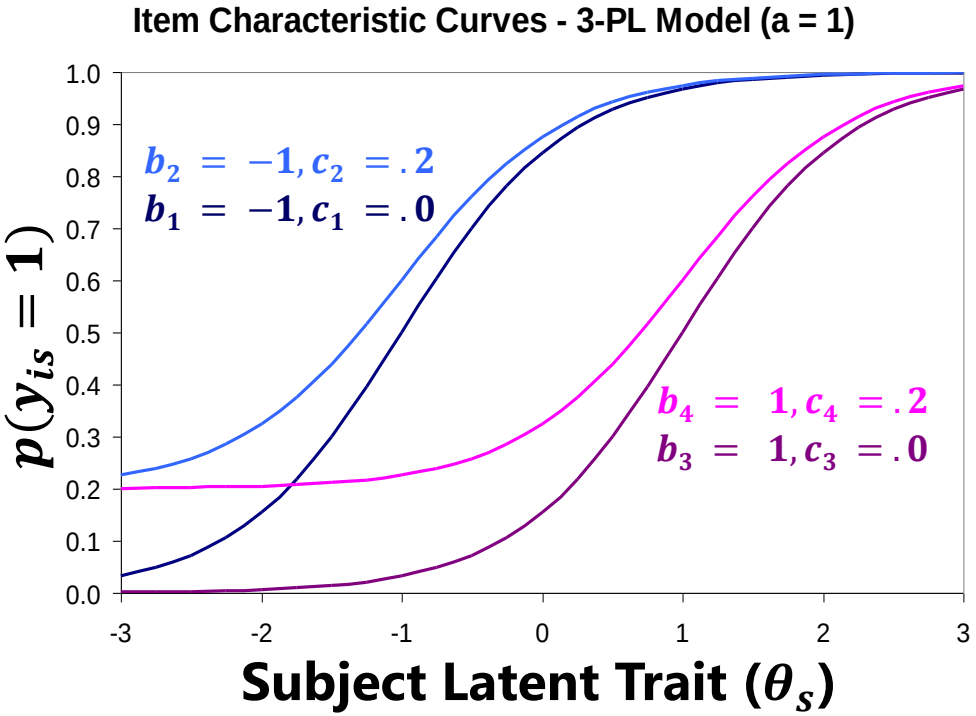
$$p(y_{is} = 1) = c_i + (1 - c_i)\left[\frac{\exp[a_i(\theta_s - b_i)]}{1 + \exp[a_i(\theta_s - b_i)]}\right]$$

- $\boldsymbol{b_i}$ = item difficulty → still is location (a fixed effect), but not at prob = .50
  - Higher values → more difficult items (lower probability of a 1)
- $\boldsymbol{a_i}$ = item discrimination → still is slope at b_i (a fixed effect)
  - Higher values = more discriminating items = better items *at its location*
- $\boldsymbol{c_i}$ = item lower asymptote → "guessing" (where $\boldsymbol{c_i > 0}$; is a fixed effect)
  - Lower bound of probability of 1 independent of theta
  - e.g., would be around .25 given 4 equally guess-able multiple-choice responses
  - Could estimate a common c across items as an alternative (but is not often done)
- Probability starts at guessing c_i then depends on theta θ_s, a_i, and b_i
  - 3-PL model is available starting in Mplus 7.4; c_i is labeled as $2
  - Require LOTS of subjects because c_i parameters are hard to estimate—you must have enough low theta subjects to determine what the probability of guessing is likely to be

Top: Items with **lower** discrimination ($a_i = .5$)

Below: Items with **higher** discrimination ($a_i = 1$)





Note that item difficulty b_i values are still at the point of inflection, but if $c_i > 0$, that's not at prob = .50 → expected prob at b_i is higher by c_i as: $prob = (1 + c_i)/2$

One Last Model for Binary Responses: 4-Parameter Logistic Model (4PL)

$$p(y_{is} = 1) = c_i + (d_i - c_i)\left[\frac{\exp[a_i(\theta_s - b_i)]}{1 + \exp[a_i(\theta_s - b_i)]}\right]$$

- $\boldsymbol{b_i}$ = item difficulty → location (not at prob = .50)
- $\boldsymbol{a_i}$ = item discrimination → slope (at location)
- $\boldsymbol{c_i}$ = item lower asymptote → "guessing"
- $\boldsymbol{d_i}$ = item upper asymptote → "**carelessness**" (so $d_i < 1$)
  - Maximum probability to be achieved independent of trait (theta θ_s)
  - Could be carelessness or unwillingness to endorse the item no matter what

All item parameters remain **fixed effects**

- Probability starts at "guessing" c_i, tops out at "carelessness" d_i, then in between depends on theta θ_s, a_i, and b_i
  - 4-PL model in Mplus 7.4 onward; c_i and d_i are labeled as \$2 and \$3
  - But good luck estimating it! May need to use a common c and d instead

Anchoring: Model Identification in IRT

- As in CFA, we have a latent trait (a pretend predictor) without a scale: **so we need to give each theta θ_s a mean and a variance**
  - This is called "**anchoring**" in IRT → CFA calls it "**model identification**"
  - As in CFA, there are 2 equivalent options: *Anchor by Subjects* or *Anchor by Items*

- **Anchor by subjects**: Fix theta θ_s mean = 0 and theta θ_s variance = 1
  - Is "z-score" (standardized factor) model identification used in CFA
  - All item difficulties b_i and item discriminations a_i are then estimated
  - In Rasch model, the common a would be estimated but equal across items

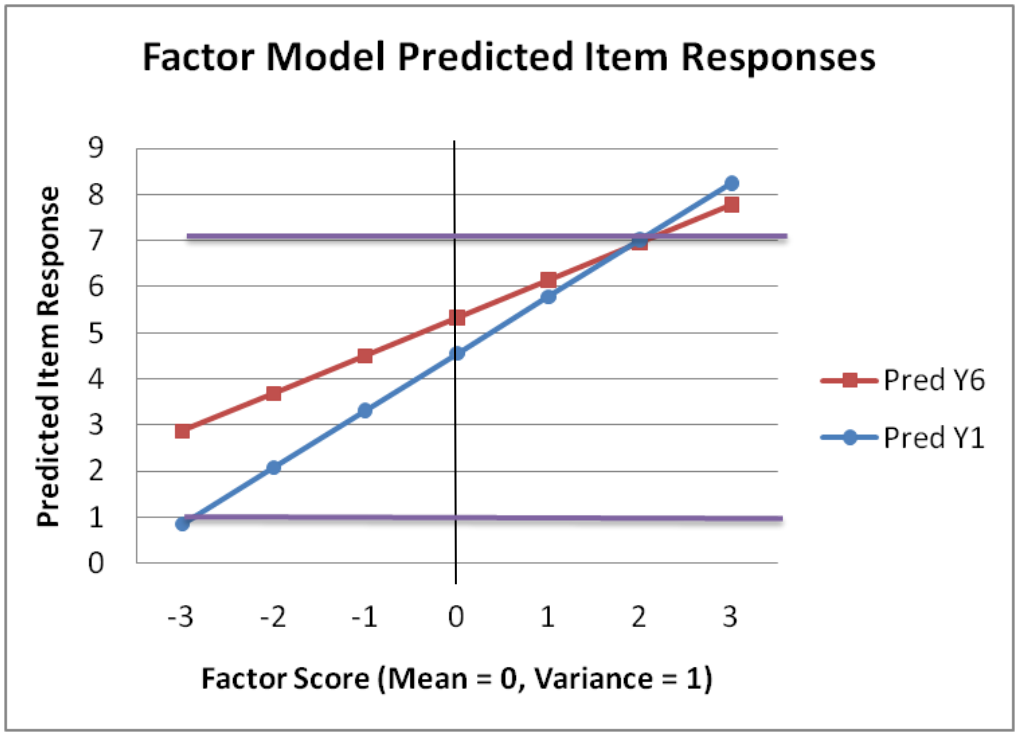
- **Anchor by items**: Fix one item difficulty b_i = 0 and one item a_i = 1
  - Is "marker item" approach to model identification used in CFA
  - Mean and variance of theta θ_s are estimated instead
  - Fixing mean of item difficulty = 0 is another way (more common in Europe)

- Big picture: as in CFA, the numerical scale doesn't matter, all that matters is that subjects and items are on the same scale → "conjoint scaling"

Information: Reliability in IRT Models

- "**Information**" ≈ reliability → measurement precision
- In **CFA models** (continuous y_{is}), item-specific "information" is rarely referred to, because standardized loadings cover it:
  - How good is my item → how much information is in it?
    - How much of its variance is "true" (shared with the factor) relative to how much of its variance is "error"?
    - **Information = unstandardized loading^2 / error variance**
  - Note that information is assumed **constant** across trait values in CFA
    - Items with a greater proportion of true variance are better, the end
    - So the instrument's "information function" **is FLAT** across trait values in CFA
  - How do I make my test better?
    - **More items with more information** (with stronger factor loadings)
  - Sum of information across items = **Test information function**
    - Test information function will also be flat across trait values in CFA

Item Information in CFA Models



$$y_{6s} = 5.32 + 0.82(F_s) + e_{6s}$$
$$e_{6s}^2 = 1.67$$

$$y_{1s} = 4.55 + 1.23(F_s) + e_{1s}$$
$$e_{1s}^2 = 1.53$$

Info $y_6 = 0.82^2 / 1.67 = .401$

Info $y_1 = 1.23^2 / 1.53 = .998$

Std $y_{6s} = 3.48 + 0.54(F_s) + e_{6s}$

Std $y_{1s} = 2.60 + 0.71(F_s) + e_{1s}$

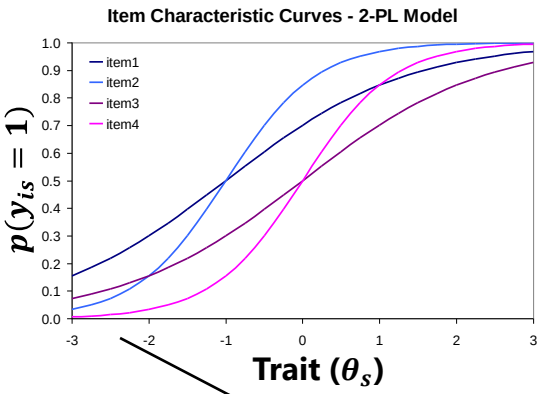
- CFA has a **linear slope (factor loading)** → predicts the same increase in the $\boldsymbol{y_{is}}$ item response for per unit higher $\boldsymbol{F_s}$ (all across levels of F_s)
- $\boldsymbol{y_1}$ **has more information than** $\boldsymbol{y_6}$ (and a higher standardized factor loading), so y_1 is better than y_6, period(t) (for all possible factor scores)

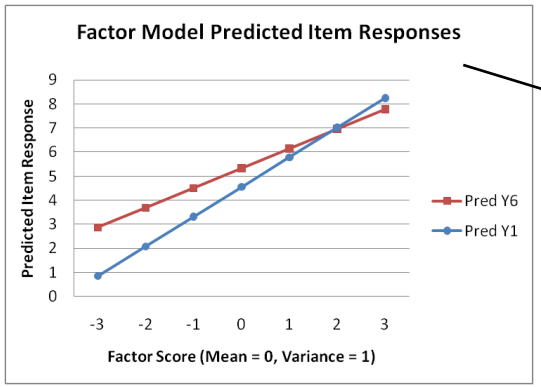
Test Information in IRT Models

- IRT test information can be converted to a reliability metric as follows:
  - **Reliability = information / (information+1)**
    - **Information of 4 converts to reliability of .80**
    - **information of 9 converts to reliability of .90**
- This formula comes from classical test theory:
  - Reliability = true var / (true var + error var)
  - Reliability = 1 / (1 + error var), where error var = 1/info
  - Reliability = 1 / 1 + (1/info) → info / (info+1)

- An analog of overall model-based reliability (e.g., omega) could be formed by summing reliabilities for each possible theta, weighted by the number of subjects at each level of theta, but (to me) that's missing the point…
- Because the slopes relating Theta to the probability of an item response are non-linear, this means that **reliability must VARY over theta**
  - **So FOR WHOM is your test sufficiently reliable??**

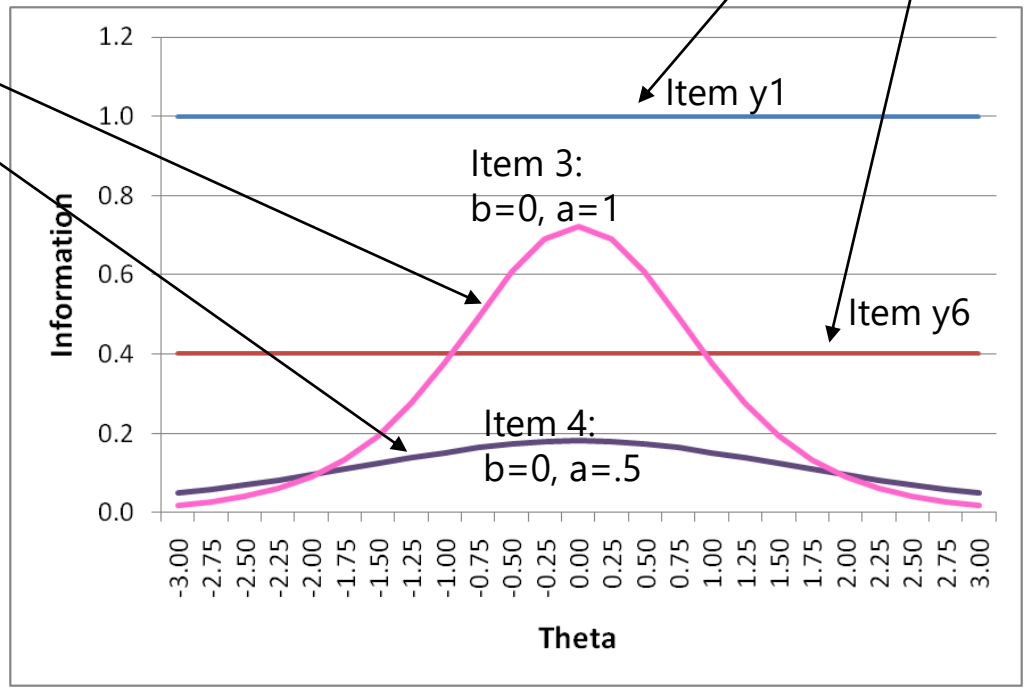
Item Information in CFA vs. IRT





CFA Item Information Functions

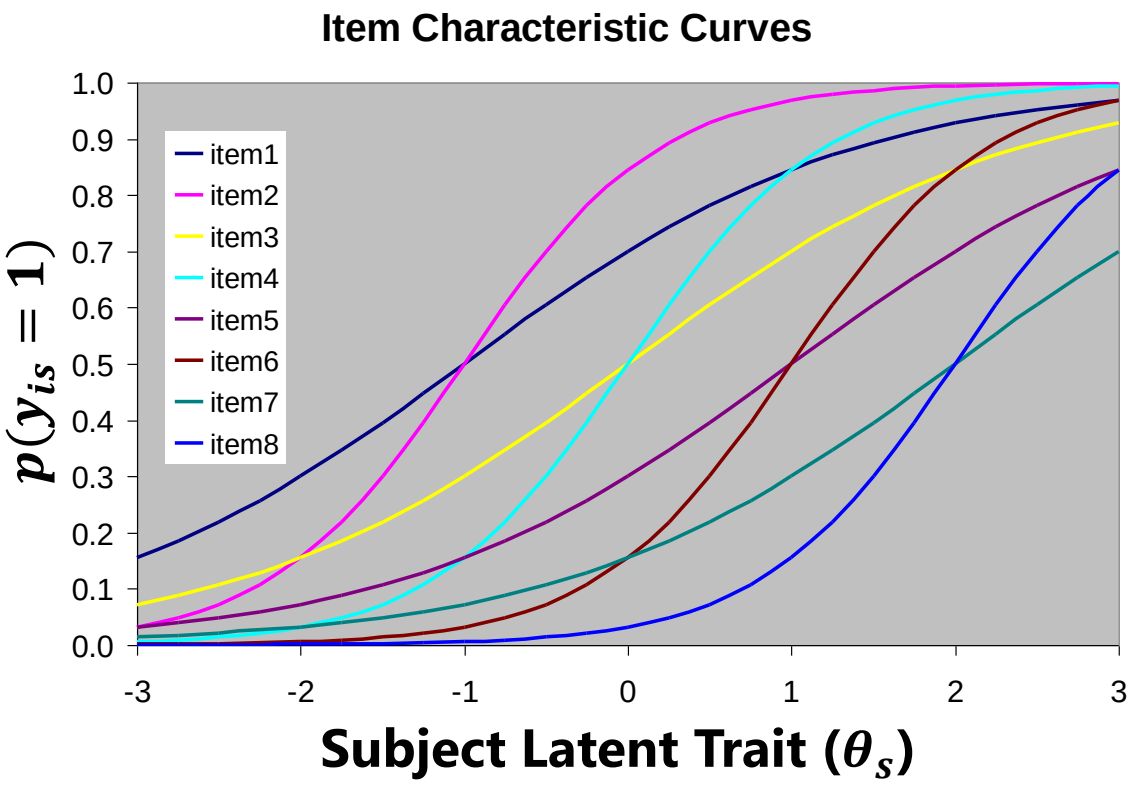
IRT Item Information Functions



If theta variance =1, then *at a given theta value*, binary item information $= a^2 * p(y_{is} = 1) * p(y_{is} = 0)$

Effects of Item Parameters on Item Characteristic Curves

Item	1	2	3	4	5	6	7	8
a discrimination	0.5	**1.0**	0.5	**1.0**	0.5	**1.0**	0.5	**1.0**
b difficulty	-1.0	-1.0	0.0	0.0	1.0	1.0	2.0	2.0

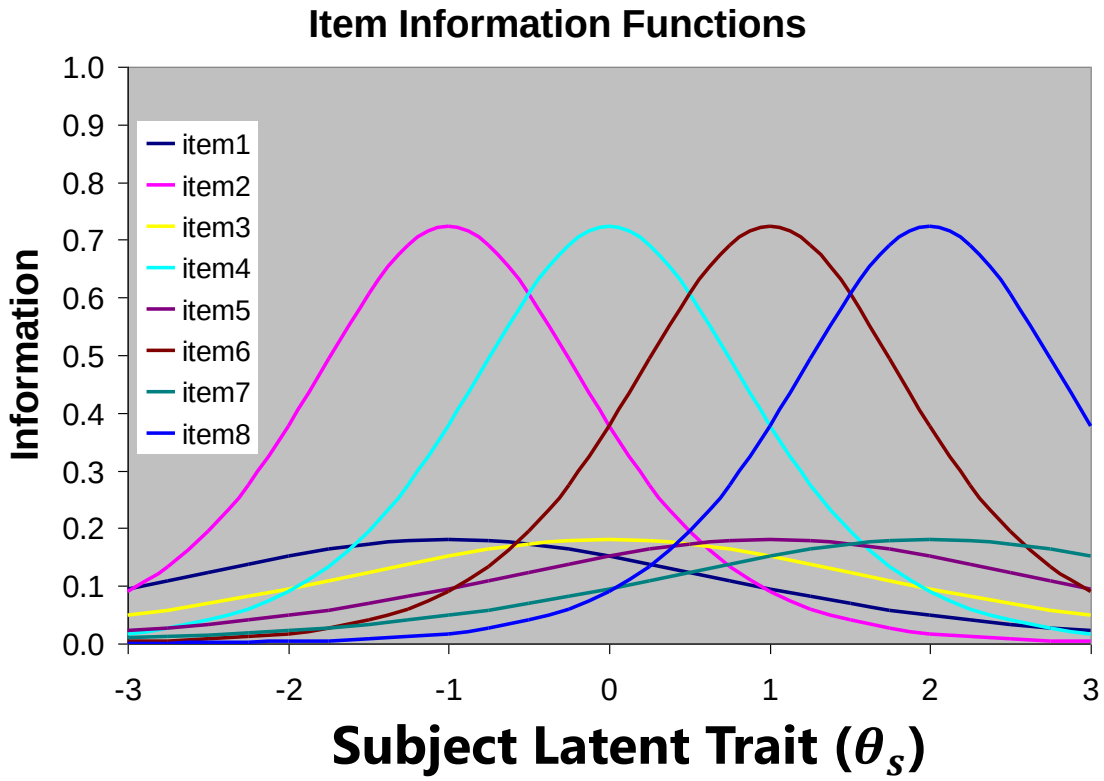


An important result of the non-linear slopes in an IRT model is that the **slope stops working** (so reliability decreases) as you move away from the item difficulty location.

In the **CFA** model with linear slopes, **the slope never stops working** (at least in theory).

Effects of Item Parameters on Item Information Curves

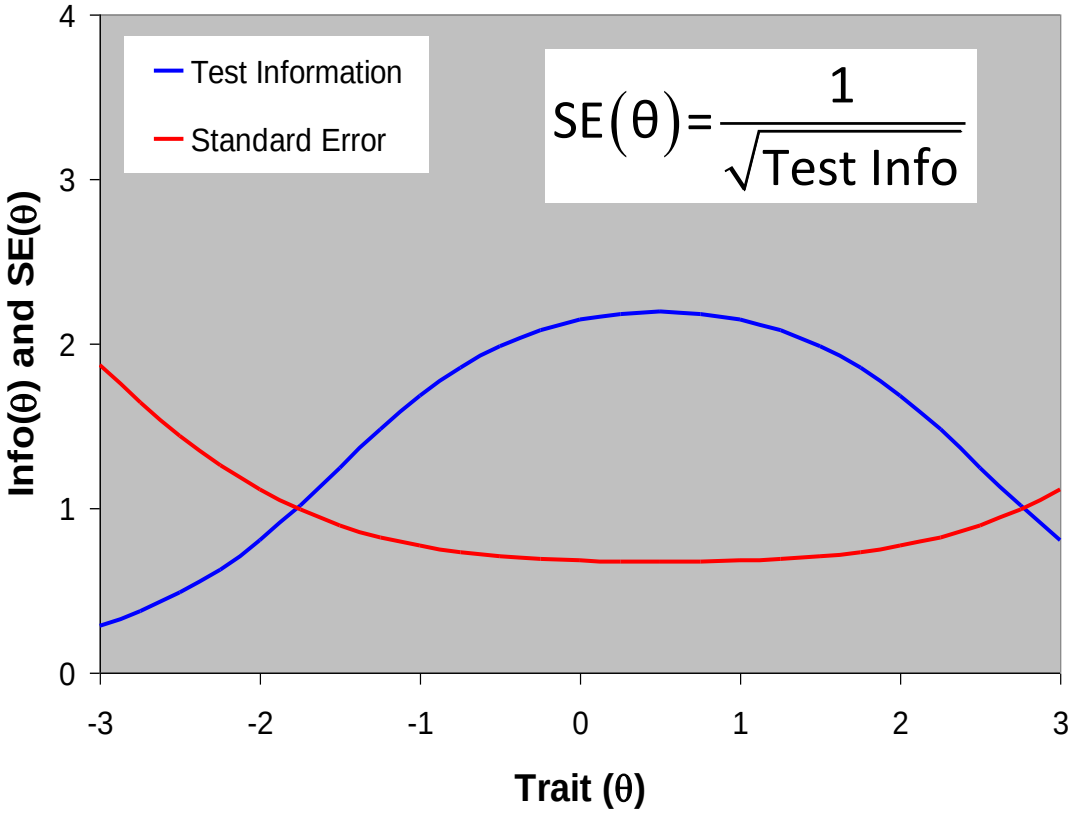
Item	1	2	3	4	5	6	7	8
a discrimination	0.5	**1.0**	0.5	**1.0**	0.5	**1.0**	0.5	**1.0**
b difficulty	-1.0	-1.0	0.0	0.0	1.0	1.0	2.0	2.0



Information (reliability) is maximized around the item difficulty location.

Items with greater a_i item discrimination values have greater absolute information.

Test Information (and SE) by Theta



If you sum all the item information curves, you get a **test information** curve that describes how reliable your set of items is over the range of the trait (Theta).

Test Information is very useful to know—it can tell you where the holes are in your measurement precision, it and guides you in adding/removing items.

There is no single "ideal" test information function—only what is optimal for **your** purposes of measurement. Here are a few examples….

Another Example of (Not-So-Good) Test Information

But test info only gets up to ~2…

(Uh oh!)

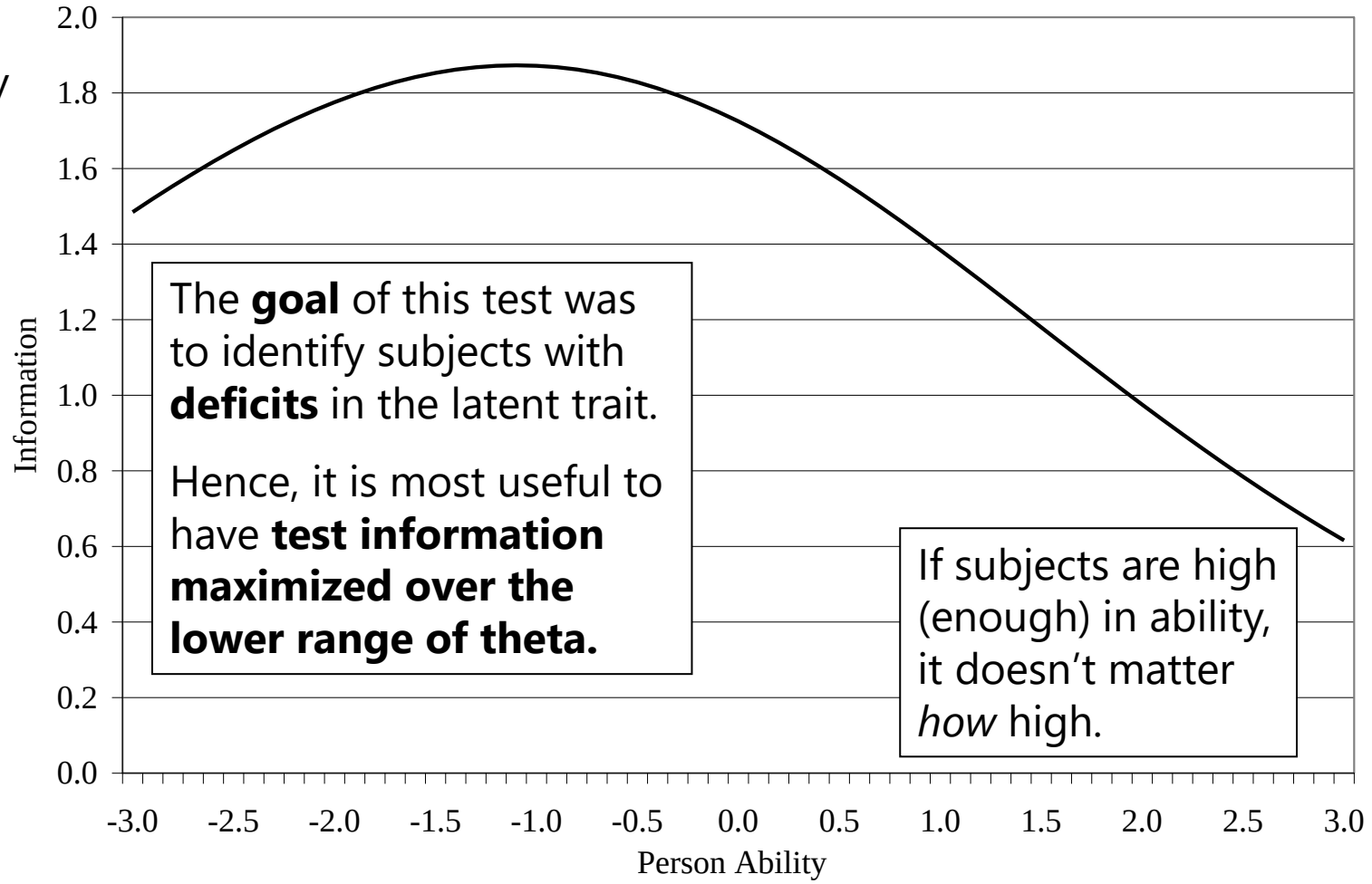
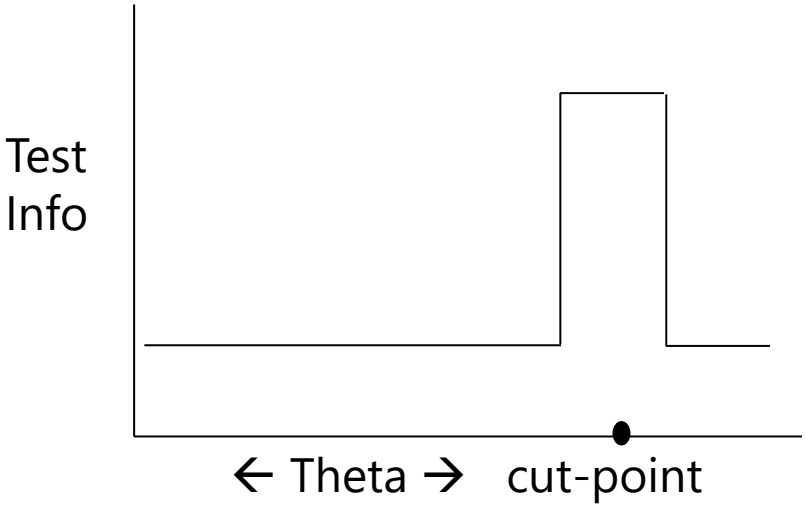


Image from my dissertation (the "done" kind)!

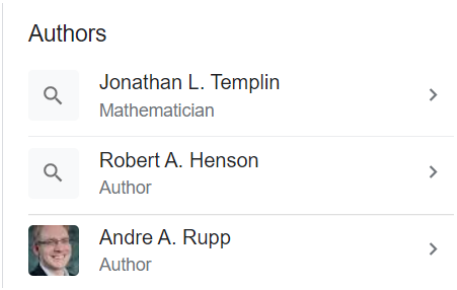
Other Shapes of Test Information

- If the goal is to measure a trait across subjects equally well, and you expect people to be normally distributed, then your best bet is to create a test with information highest in the middle (where most people are likely to be)
- If your goal is to identify individuals below or above a cut-point, however, your **test information function** should ideally look more like this:

  - Want to **maximize sensitivity near the cut-point region,** and not waste time measuring people well who are far away from the cut-point
  - If **classifying subjects** is the goal of measurement, however, you might be better off with a different family of latent trait models in which Theta is already a categorical "attribute" instead: Diagnostic Classification Models, as covered by the book **Diagnostic Measurement** …



According to The Google:



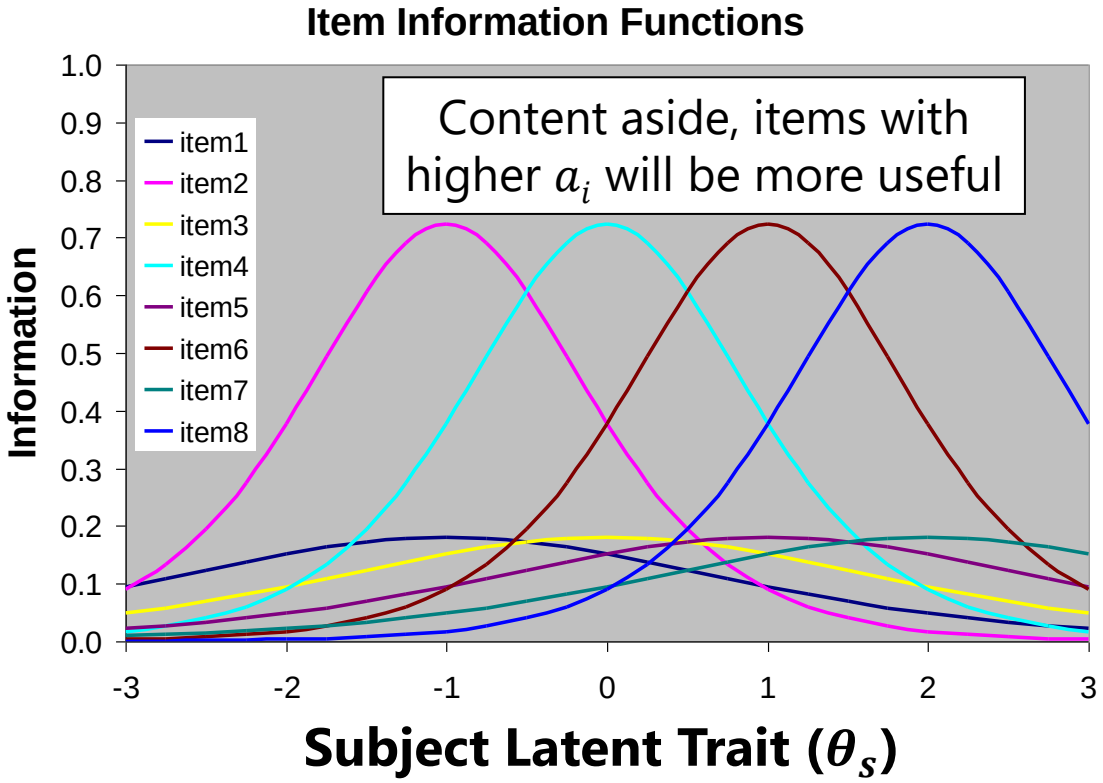
How to Improve Your Reliability

- In CTT, because item properties are not part of the model, items are seen as exchangeable, and more items is better
  - Thus, *any* new item is *equally* better for the model

- In CFA and IRT, more items is still better…
  - **In CFA, the question is "how much better"?**
    - This depends on the standardized loading; intercepts are not important
    - Specifies a **linear relationship** between theta and the item responses, so "for whom" isn't relevant—a better item is **better for everyone equally**
  - **In IRT, the question is "how much better, and for whom?"**
    - Depends on the discrimination (a_i slope) and the difficulty (b_i location), respectively (difficulties are important, and are always estimated)
    - Because of the **nonlinear relationship** between theta and the item responses, items are **only useful for thetas in the middle of their S-curves**

Effects of Item Parameters on Item Information Curves

Item	1	2	3	4	5	6	7	8
a discrimination	0.5	**1.0**	0.5	**1.0**	0.5	**1.0**	0.5	**1.0**
b difficulty	-1.0	-1.0	0.0	0.0	1.0	1.0	2.0	2.0



In addition to a_i item discrimination, though, you want to make sure you are covering the range of difficulty where you want to measure your subjects best.

IRT and Adaptive Testing: *Fewer Items Can Actually Be Better*

- In a normal distribution of the latent trait and a comparable distribution of item difficulty, **extreme subjects are usually measured less well** (higher SE).

- For fixed-item tests, more items is generally better, but one can get the same precision of measurement with fewer items by using **adaptive tests with items of targeted levels of difficulty**. Different forms across subjects are given to maximize efficiency.

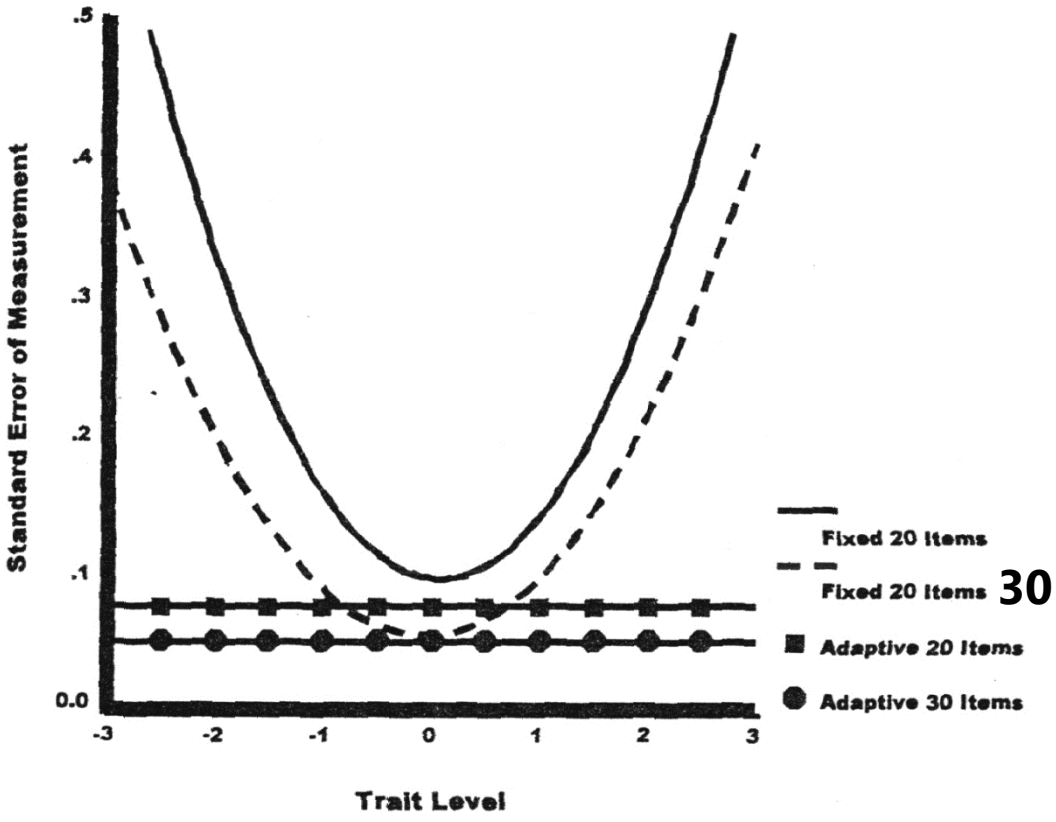
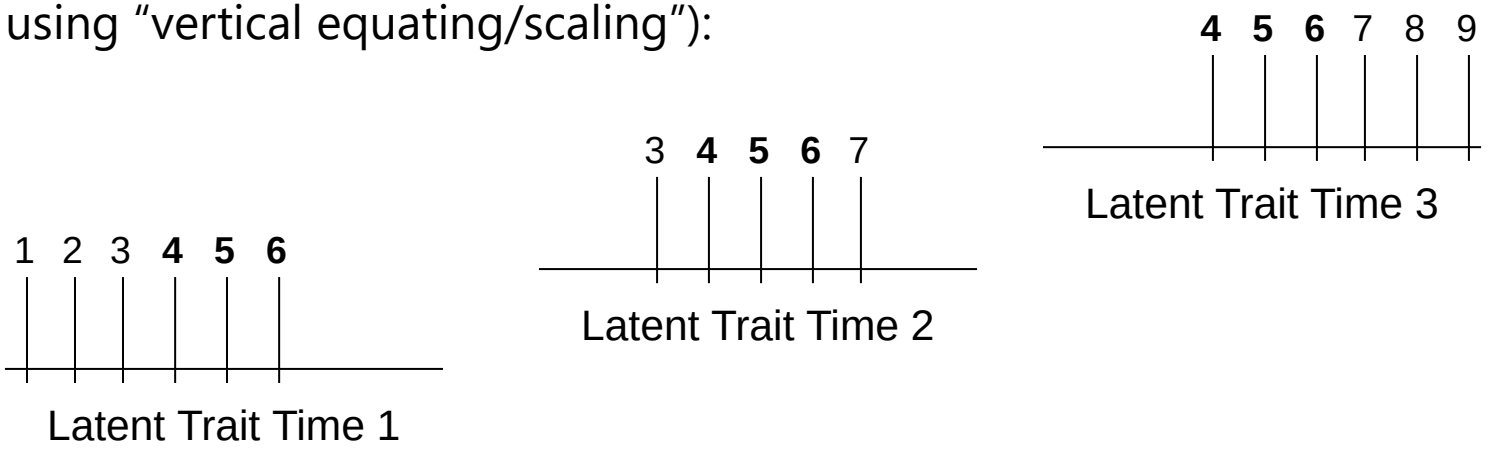
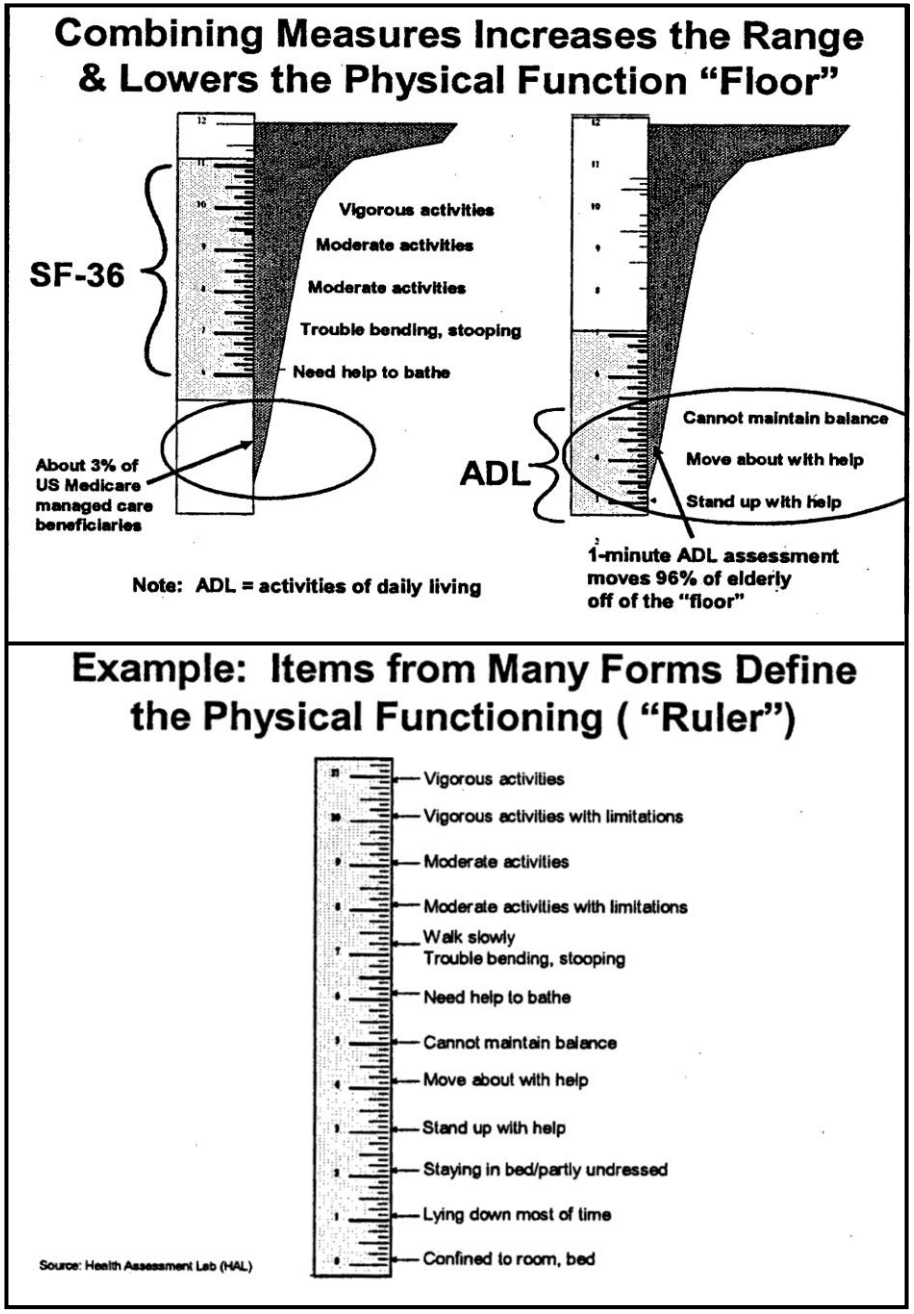


Image from Embretson & Reise (2000)

IRT (and CFA) Help Measure Change AND Maintain Sensitivity across Samples

- **Theta is scaled and interpreted relative to the items**, not relative to the other persons in the sample (in 2PL, is item difficulty at *prob* = .50)
  - This means you can give different forms over time and still compare thetas
  - MUST have some "**linking items**" → common set of items across occasions
  - Although this property is helpful when dealing with "accidental" alternative forms (e.g., changed response options, dropped items), linking items can be used advantageously as well
  - Here, **we grow a test over time** within a sample (i.e., using "vertical equating/scaling"):





Linking Thetas across Tests

SF-36: measure of *higher* physical functioning

ADL: measure of *lower* physical functioning

So don't choose: Administer a core set of linking items from both tests to a single sample

Linking items then form a common metric

- More precision than single test
- Allows for comparisons across groups or studies

See Mungas & Reed (2000) for an example of linking over forms

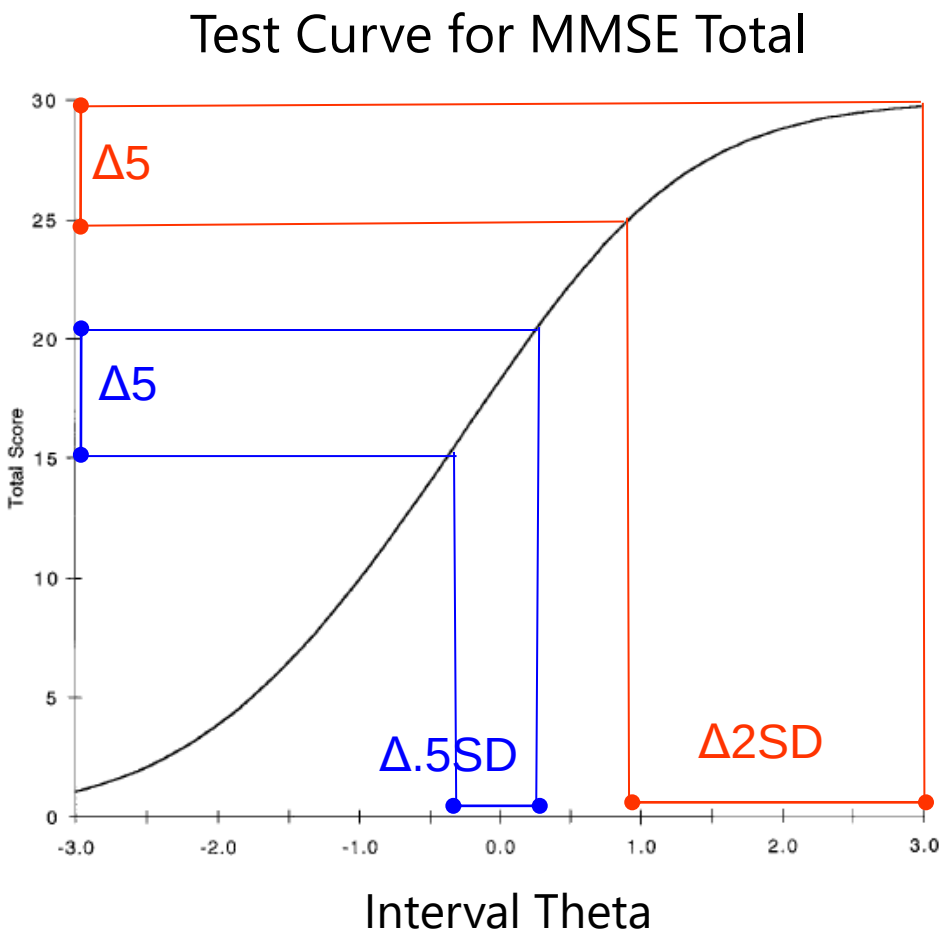
Another Benefit of IRT (and CFA)

- **IRT**: If the model fits, the scale of **theta is linear/interval**
  - Supports mathematical operations that assume interval measurement
  - Same ordering of subjects as in raw scores, but the distances between subjects may be different, especially at the ends (due to less reliability)

- **CTT**: **Sum scores** have an **ordinal** relationship to the latent trait at best
  - Does not support operations that assume interval measurement, which can bias tests of mean differences, regression slopes, etc.
  - Spurious interactions can result in tests of mean differences if groups differ in how well they are measured (i.e., floor and ceiling effects)

- Bottom line: Measurement matters for testing everyday hypotheses, NOT just when fitting measurement models for specific issues

Example from Mungas & Reed (2000)



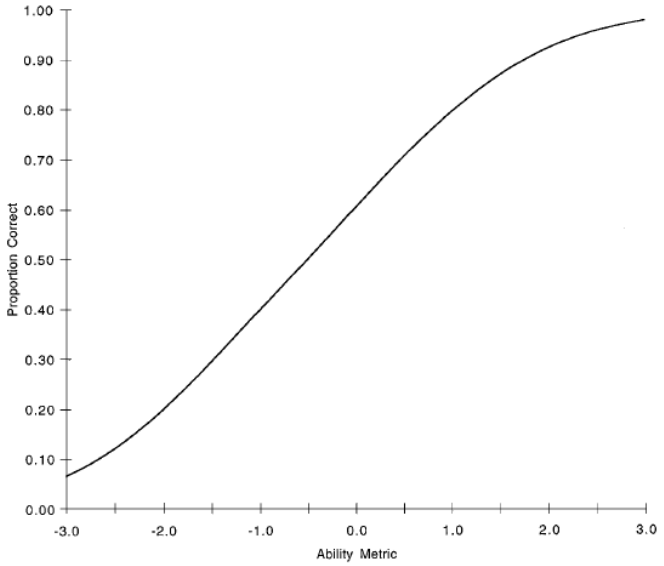
The bottom and top of the MMSE total score (ordinal) are "squished" relative to the latent trait scale (interval).

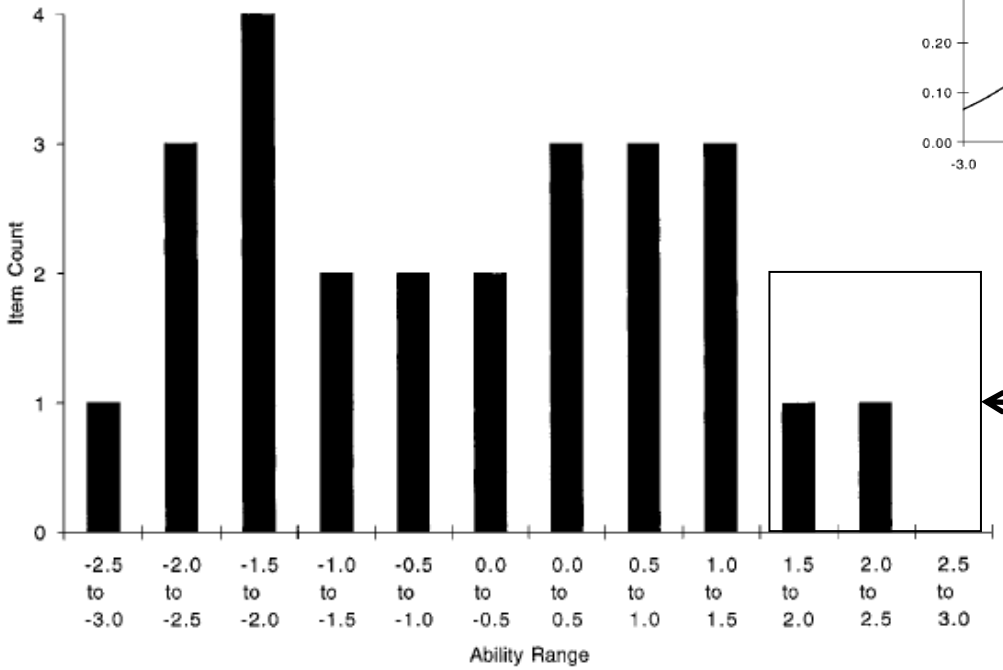
This means that one-unit changes along the MMSE total do not really have the same meaning across the latent trait, which makes many kinds of comparisons problematic.

Example from Mungas & Reed (2000)

Right: They combined 3 tests to get better measurement, as shown in the test curve →

Below: Items at each trait location contribute to scale's capacity to differentiate persons in ability at each point in the continuum.





There is a hole near the top, which explains the flattening of the curve (less information there).

IRT and IFA Part 1: Summary

- **CFA** models use ≥1 latent factors to predict **continuous** item responses (as in linear regression)
- **IFA/IRT** models use ≥1 latent factors to predict **categorical** item responses (as in binary, ordinal, or nominal regression)
  - IFA models will look more like CFA models – stay tuned for Part 2!
  - IRT models look strange, but their $\boldsymbol{b}$ location parameters are (arguably) more useful than the IFA versions (which is why I start with IRT)
    - At a minimum, items differ in $\boldsymbol{b}$ = trait location (as difficulty/severity) → 1PL or Rasch
    - Could also allow different $\boldsymbol{a}$ discrimination (as max slope) across items → 2PL
    - Could also allow different $\boldsymbol{c}$ lower or $\boldsymbol{d}$ upper asymptotes → 3PL or 4PL

- Because latent traits (factors, variables, now called theta $\boldsymbol{\theta}$) have a **nonlinear** relation to the probability of a response:
  - Items are **most useful** for trait levels **at their $\boldsymbol{b}$ location** parameter
  - **Reliability** (as "test information") **must vary over the latent trait**, because it depends on how many (and how good) items you have at each location!