Psychology 405: Psychometric Theory
Reliability Theory

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

May, 2020

Outline of Reliability Theory

1. Classical Test Theory
2. Generalizability approaches – ICC and raters
3. Item Response Theory: The new psychometrics?

Outline of Part III: the New Psychometrics

Classical Reliability

1. Classical model of reliability
   - Observed = True + Error
   - Reliability $= 1 - \frac{\sigma^2_{error}}{\sigma^2_{observed}}$
   - Reliability $= r_{xx} = r^2_{x_{domain}}$
   - Reliability as correlation of a test with a test just like it
2. Reliability requires variance in observed score
   - As σ^2_x decreases so will $r_{xx} = 1 - \frac{\sigma^2_{error}}{\sigma^2_{observed}}$
3. Alternate estimates of reliability all share this need for variance
   - 3.1 Internal Consistency
   - 3.2 Alternate Form
   - 3.3 Test-retest
   - 3.4 Between rater
4. Item difficulty is ignored, items assumed to be sampled at random

The "new psychometrics"

1. Model the person as well as the item
   - People differ in some latent score
   - Items differ in difficulty and discriminability
2. Original model is a model of ability tests
   - $p(correct|ability, difficulty, ...) = f(ability - difficulty)$
   - What is the appropriate function?
3. Extensions to polytomous items, particularly rating scale models

Classic Test Theory as 0 parameter IRT

Classic Test Theory considers all items to be random replicates of each other and total (or average) score to be the appropriate measure of the underlying attribute. Items are thought to be endorsed (passed) with an increasing probability as a function of the underlying trait. But if the trait is unbounded (just as there is always the possibility of someone being higher than the highest observed score, so is there a chance of someone being lower than the lowest observed score), and the score is bounded (from p=0 to p=1), then the relationship between the latent score and the observed score must be non-linear. This leads to the most simple of all models, one that has no parameters to estimate but is just a non-linear mapping of latent to observed:

$$p(correct_{ij}|\theta_i) = \frac{1}{1 + e^{-\theta_i}}. \quad (1)$$

Rasch model – All items equally discriminating, differ in difficulty

Slightly more complicated than the zero parameter model is to assume that all items are equally good measures of the trait, but differ only in their difficulty/location. The *one parameter logistic* (*1PL*) *Rasch model* (**?**) is the easiest to understand:

$$p(correct_{ij}|\theta_i, \delta_j) = \frac{1}{1 + e^{\delta_j - \theta_i}}. \quad (2)$$

That is, the probability of the i^{th} person being correct on (or endorsing) the j^{th} item is a logistic function of the difference between the person's ability (latent trait) (θ_i) and the item difficulty (or location) (δ_j). The more the person's ability is greater than the item difficulty, the more likely the person is to get the item correct.

Estimating the model

The probability of missing an item, q, is just 1 - p(correct) and thus the *odds ratio* of being correct for a person with ability, θ_i, on an item with difficulty, δ_j is

$$OR_{ij} = \frac{p}{1-p} = \frac{p}{q} = \frac{\frac{1}{1+e^{\delta_j-\theta_i}}}{1-\frac{1}{1+e^{\delta_j-\theta_i}}} = \frac{\frac{1}{1+e^{\delta_j-\theta_i}}}{\frac{e^{\delta_j-\theta_i}}{1+e^{\delta_j-\theta_i}}} = \frac{1}{e^{\delta_j-\theta_i}} = e^{\theta_i-\delta_j}. \tag{3}$$

That is, the odds ratio will be a exponential function of the difference between a person's ability and the task difficulty. The odds of a particular pattern of rights and wrongs over n items will be the product of n odds ratios

$$OR_{i1}\,OR_{i2}\ldots OR_{in} = \prod_{j=1}^{n} e^{\theta_i-\delta_j} = e^{n\theta_i}\,e^{-\sum_{j=1}^{n}\delta_j}. \tag{4}$$

Estimating parameters

Substituting P for the pattern of correct responses and Q for the pattern of incorrect responses, and taking the logarithm of both sides of equation 4 leads to a much simpler form:

$$ln\frac{P}{Q} = n\theta_i + \sum_{j=1}^{n} \delta_j = n(\theta_i + \bar{\delta}). \quad (5)$$

That is, the log of the pattern of correct/incorrect for the i^{th} individual is a function of the number of items * (θ_i - the average difficulty). Specifying the average difficulty of an item as $\bar{\delta} = 0$ to set the scale, then θ_i is just the logarithm of P/Q divided by n or, conceptually, the average logarithm of the p/q

$$\theta_i = \frac{ln\frac{P}{Q}}{n}. \quad (6)$$

Difficulty is just a function of probability correct

Similarly, the pattern of the odds of correct and incorrect responses across people for a particular item with difficulty δ_j will be

$$OR_{1j}OR_{2j}\ldots OR_{nj} = \frac{P}{Q} = \prod_{i=1}^{N} e^{\theta_i - \delta_j} = e^{\sum_{i=1}^{N}(\theta_i) - N\delta_j} \tag{7}$$

and taking logs of both sides leads to

$$ln\frac{P}{Q} = \sum_{i=1}^{N}(\theta_i) - N\delta_j. \tag{8}$$

Letting the average ability $\bar{\theta} = 0$ leads to the conclusion that the difficulty of an item for all subjects, δ_j, is the logarithm of Q/P divided by the number of subjects, N,

$$\delta_j = \frac{ln\frac{Q}{P}}{N}. \tag{9}$$

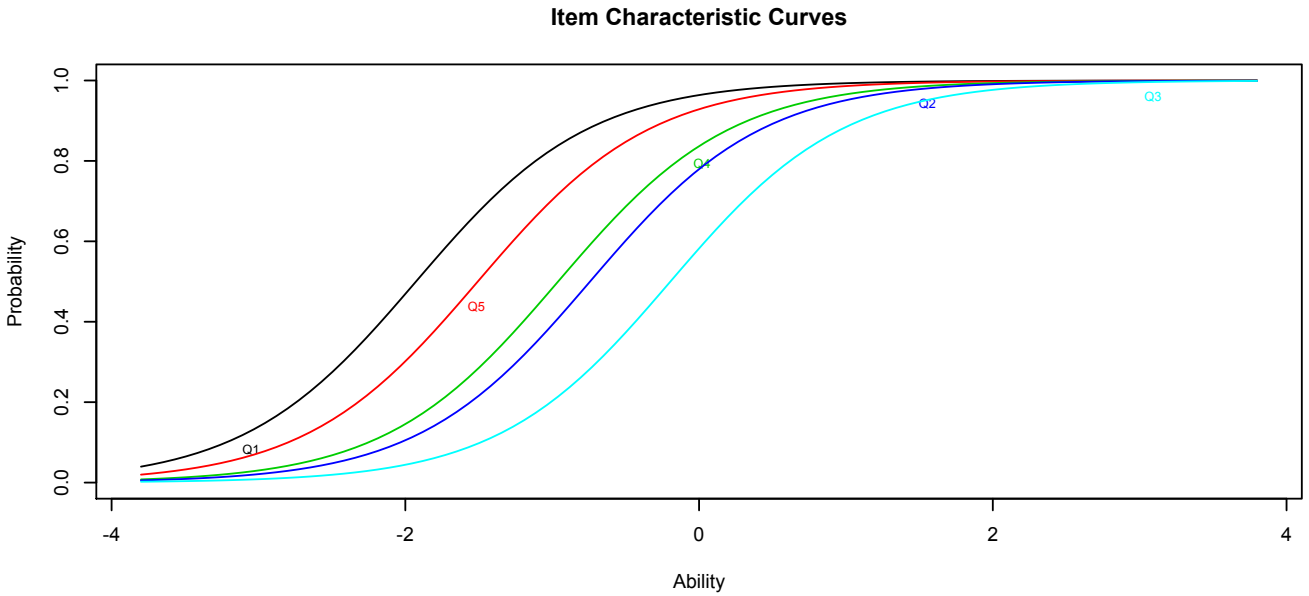
Rasch model in words

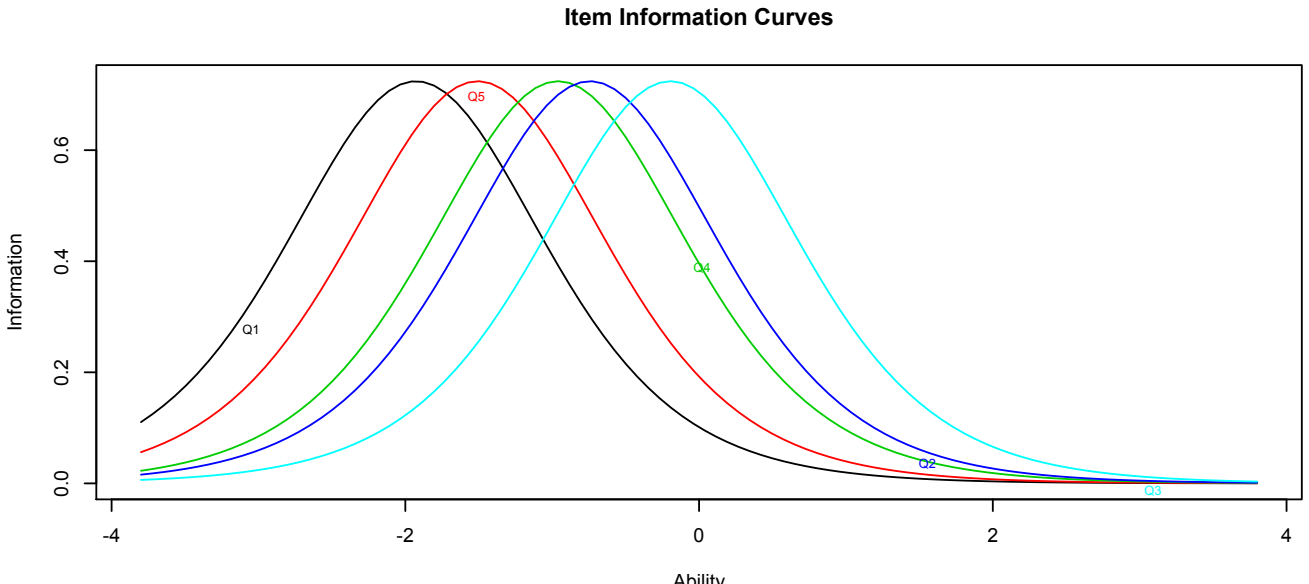
That is, the estimate of ability (Equation 6) for items with an average difficulty of 0 does not require knowing the difficulty of any particular item, but is just a function of the pattern of corrects and incorrects for a subject across all items.
Similarly, the estimate of item difficulty across people ranging in ability, but with an average ability of 0 (Equation 9) is a function of the response pattern of all the subjects on that one item and does not depend upon knowing any one person's ability. The assumptions that average difficulty and average ability are 0 are merely to fix the scales. Replacing the average values with a non-zero value just adds a constant to the estimates.

Rasch as a high jump

The independence of ability from difficulty implied in equations 6 and 9 makes estimation of both values very straightforward. These two equations also have the important implication that the number correct ($n\bar{p}$ for a subject, $N\bar{p}$ for an item) is monotonically, but not linearly related to ability or to difficulty.
That the estimated ability is independent of the pattern of rights and wrongs but just depends upon the total number correct is seen as both a strength and a weakness of the Rasch model. From the perspective of *fundamental measurement*, Rasch scoring provides an additive interval scale: for all people and items, if $\theta_i < \theta_j$ and $\delta_k < \delta_l$ then $p(x|\theta_i, \delta_k) < p(x|\theta_j, \delta_l)$. But this very additivity treats all patterns of scores with the same number correct as equal and ignores potential information in the pattern of responses.

Rasch estimates from ltm





The LSAT example from ltm

```
data(bock)
> ord <- order(colMeans(lsat6),decreasing=TRUE)
> lsat6.sorted <- lsat6[,ord]
> describe(lsat6.sorted)
> Tau <- round(-qnorm(colMeans(lsat6.sorted)),2)  #tau = estimates of threshold
> rasch(lsat6.sorted,constraint=cbind(ncol(lsat6.sorted)+1,1.702))
```

```
   var    n mean   sd median trimmed mad min max range  skew kurtosis   se
Q1   1 1000 0.92 0.27      1    1.00   0   0   1     1 -3.20     8.22 0.01
Q5   2 1000 0.87 0.34      1    0.96   0   0   1     1 -2.20     2.83 0.01
Q4   3 1000 0.76 0.43      1    0.83   0   0   1     1 -1.24    -0.48 0.01
Q2   4 1000 0.71 0.45      1    0.76   0   0   1     1 -0.92    -1.16 0.01
Q3   5 1000 0.55 0.50      1    0.57   0   0   1     1 -0.21    -1.96 0.02

> Tau
   Q1    Q5    Q4    Q2    Q3
-1.43 -1.13 -0.72 -0.55 -0.13

Call:
rasch(data = lsat6.sorted, constraint = cbind(ncol(lsat6.sorted) +
    1, 1.702))

Coefficients:
Dffclt.Q1  Dffclt.Q5  Dffclt.Q4  Dffclt.Q2  Dffclt.Q3     Dscrmn
   -1.927     -1.507     -0.960     -0.742     -0.195      1.702
```

Item information

When forming a test and evaluating the items within a test, the most useful items are the ones that give the most information about a person's score. In classic test theory, *item information* is the reciprocal of the squared *standard error* for the item or for a one factor test, the ratio of the item communality to its uniqueness:

$$I_j = \frac{1}{\sigma^2_{e_j}} = \frac{h^2_j}{1 - h^2_j}.$$

When estimating ability using IRT, the information for an item is a function of the first derivative of the likelihood function and is maximized at the inflection point of the *icc*.

Estimating item information

The information function for an item is

$$I(f, x_j) = \frac{[P_j'(f)]^2}{P_j(f)Q_j(f)} \tag{10}$$

For the 1PL model, P', the first derivative of the probability function $P_j(f) = \frac{1}{1+e^{\delta-\theta}}$ is

$$P' = \frac{e^{\delta-\theta}}{(1+e^{\delta-\theta})^2} \tag{11}$$

which is just P_jQ_j and thus the information for an item is

$$I_j = P_jQ_j. \tag{12}$$

That is, information is maximized when the probability of getting an item correct is the same as getting it wrong, or, in other words, the best estimate for an item's difficulty is that value where half of the subjects pass the item.

Elaborations of Rasch

1. Logistic or cumulative normal function
   - Logistic treats any pattern of responses the same
   - Cumulative normal weights extreme scores more
2. Rasch and 1PN models treat all items as equally discriminating
   - But some items are better than others
   - Thus, the two parameter model

$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{\alpha_i(\delta_j - \theta_i)}} \tag{13}$$

2PL and 2PN models

$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{\alpha_i(\delta_j - \theta_i)}} \tag{14}$$

while in the *two parameter normal* ogive (*2PN*) model this is

$$p(correct|\theta, \alpha_j, \delta) = \frac{1}{\sqrt{2\pi}} \int_{-\text{inf}}^{\alpha(\theta-\delta)} e^{-\frac{u^2}{2}} du \tag{15}$$

where $u = \alpha(\theta - \delta)$.
The information function for a two parameter model reflects the item discrimination parameter, α,

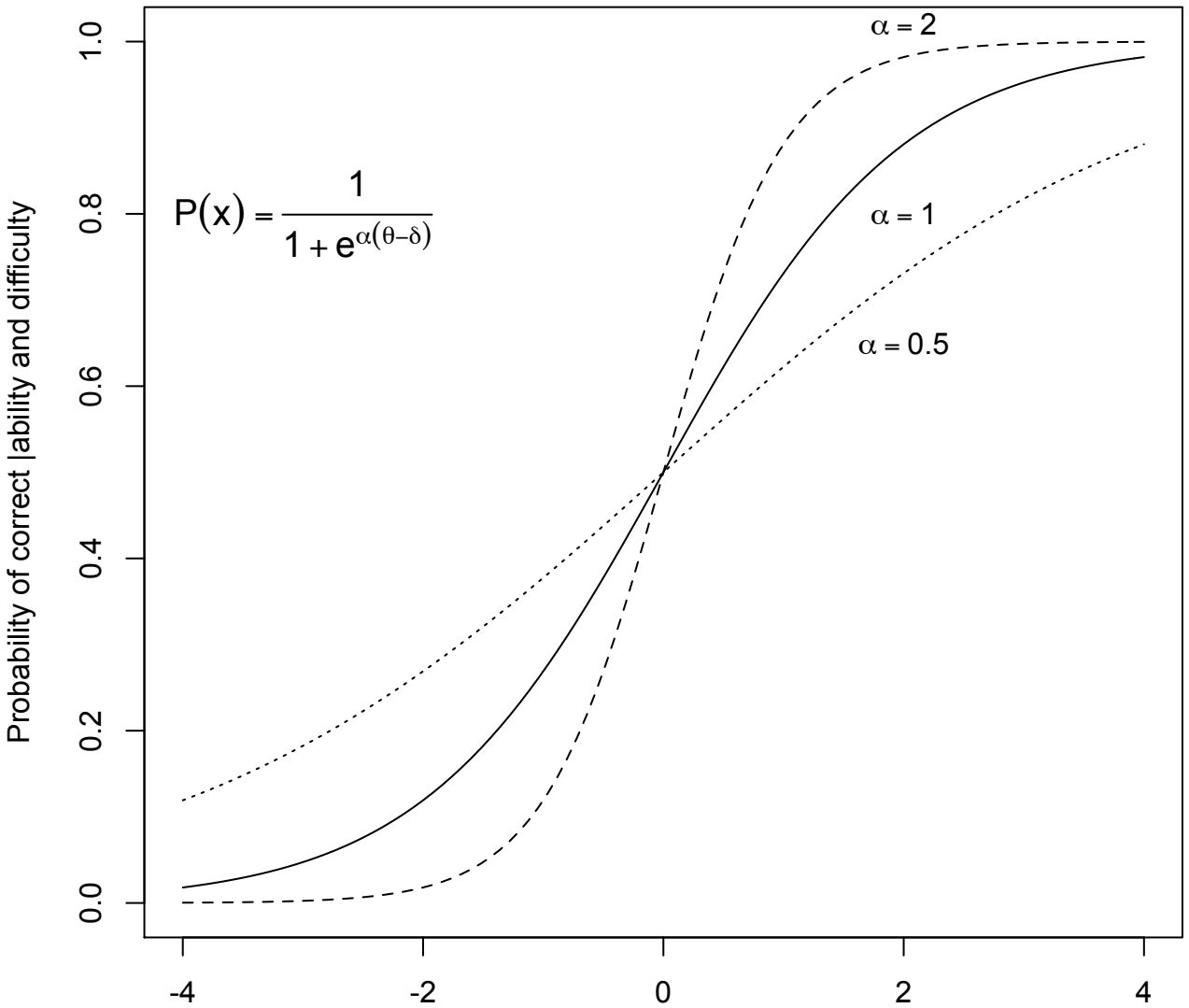
$$I_j = \alpha^2 P_j Q_j \tag{16}$$

which, for a 2PL model is

$$I_j = \alpha_j^2 P_j Q_j = \frac{\alpha_j^2}{(1 + e^{\alpha_j(\delta_j - \theta_j)})^2}. \tag{17}$$

The problem of non-parallel trace lines

2PL models differing in their discrimination parameter



Parameter explosion – better fit but at what cost

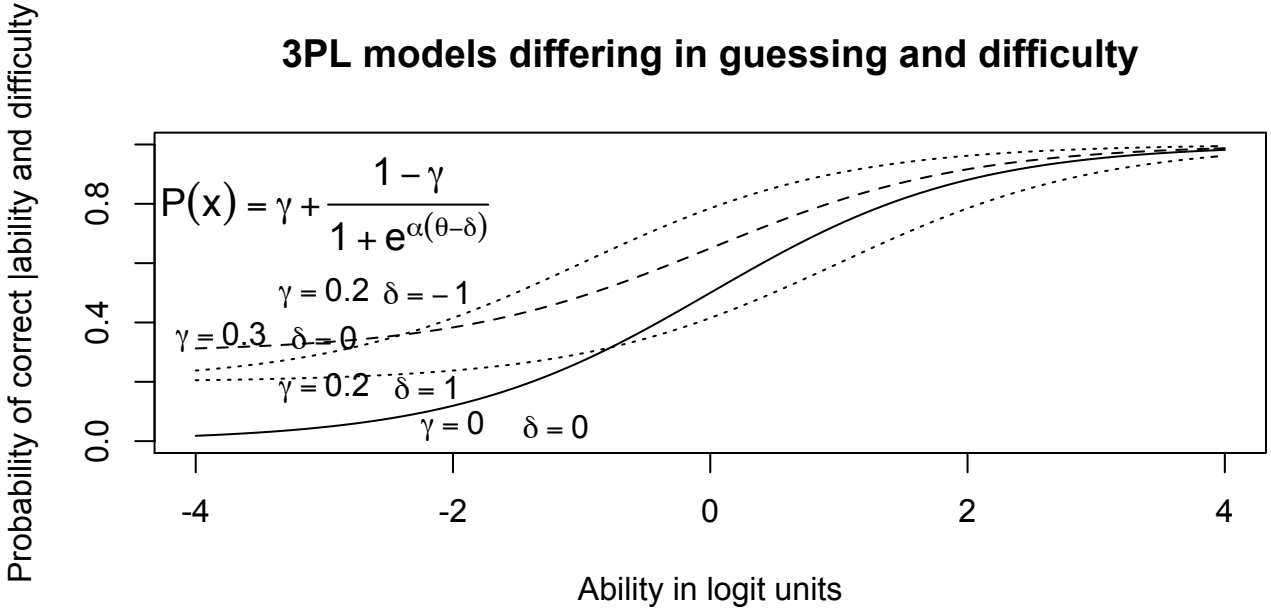
The 3 parameter model adds a guessing parameter.

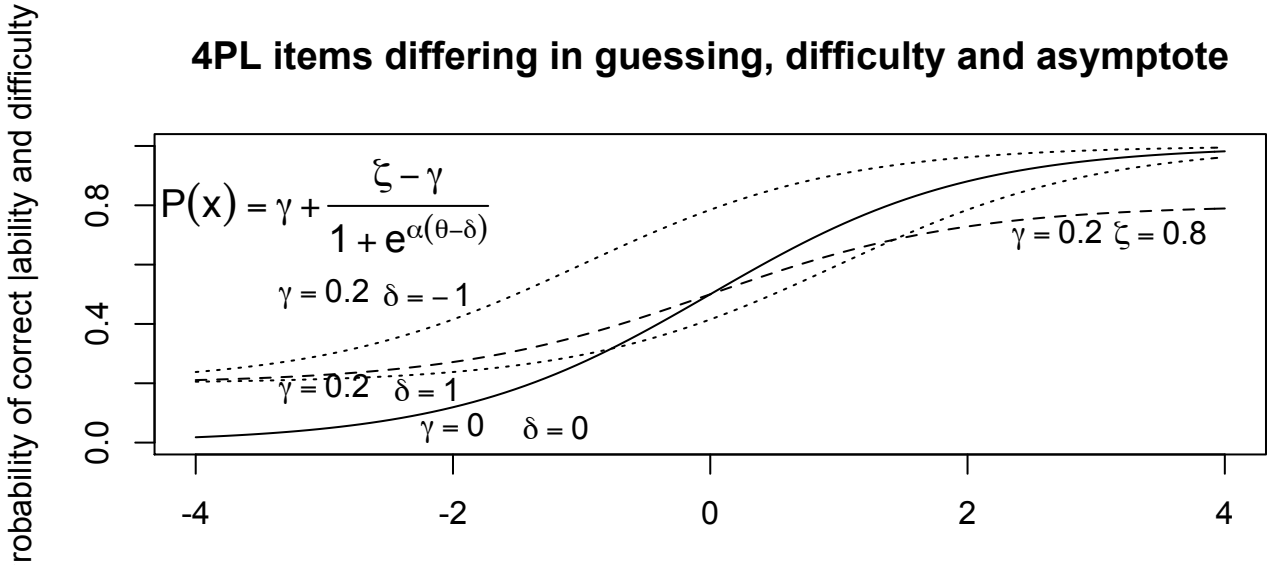
$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j, \gamma_j) = \gamma_j + \frac{1-\gamma_j}{1+e^{\alpha_i(\delta_j-\theta_i)}} \tag{18}$$

And the four parameter model adds an asymtotic parameter

$$P(x|\theta_i, \alpha, \delta_j, \gamma_j, \zeta_j) = \gamma_j + \frac{\zeta_j-\gamma_j}{1+e^{\alpha_j(\delta_j-\theta_i)}}. \tag{19}$$

frame


3PL models differing in guessing and difficulty
Probability of correct |ability and difficulty
$P(x) = \gamma + \frac{1-\gamma}{1+e^{\alpha(\theta-\delta)}}$
$\gamma = 0.2$ $\delta = -1$
$\gamma = 0.3$ $\delta = 0$
$\gamma = 0.2$ $\delta = 1$
$\gamma = 0$ $\delta = 0$
0.0
0.4
0.8
-4
-2
0
2
4
Ability in logit units


4PL items differing in guessing, difficulty and asymptote
robability of correct |ability and difficulty
$P(x) = \gamma + \frac{\zeta-\gamma}{1+e^{\alpha(\theta-\delta)}}$
$\gamma = 0.2$ $\delta = -1$
$\gamma = 0.2$ $\zeta = 0.8$
$\gamma = 0.2$ $\delta = 1$
$\gamma = 0$ $\delta = 0$
0.0
0.4
0.8
-4
-2
0
2
4

Personality items with monotone trace lines

A typical personality item might ask "How much do you enjoy a lively party" with a five point response scale ranging from "1: not at all" to "5: a great deal" with a neutral category at 3. An alternative response scale for this kind of item is to not have a neutral category but rather have an even number of responses. Thus a six point scale could range from "1: very inaccurate" to "6: very accurate" with no neutral category
The assumption is that the more sociable one is, the higher the response alternative chosen. The probability of endorsing a 1 will increase monotonically the less sociable one is, the probability of endorsing a 5 will increase monotonically the more sociable one is.

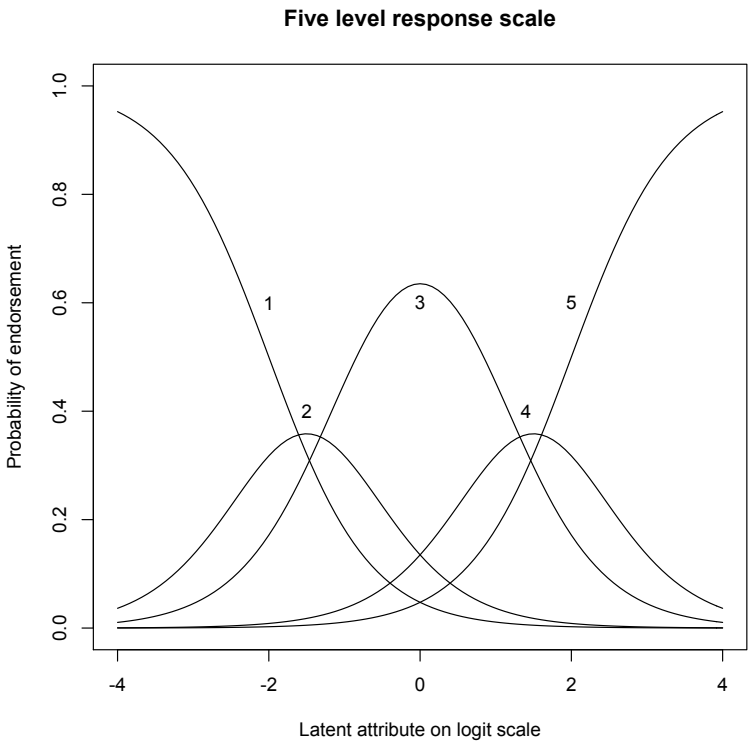
Threshold models

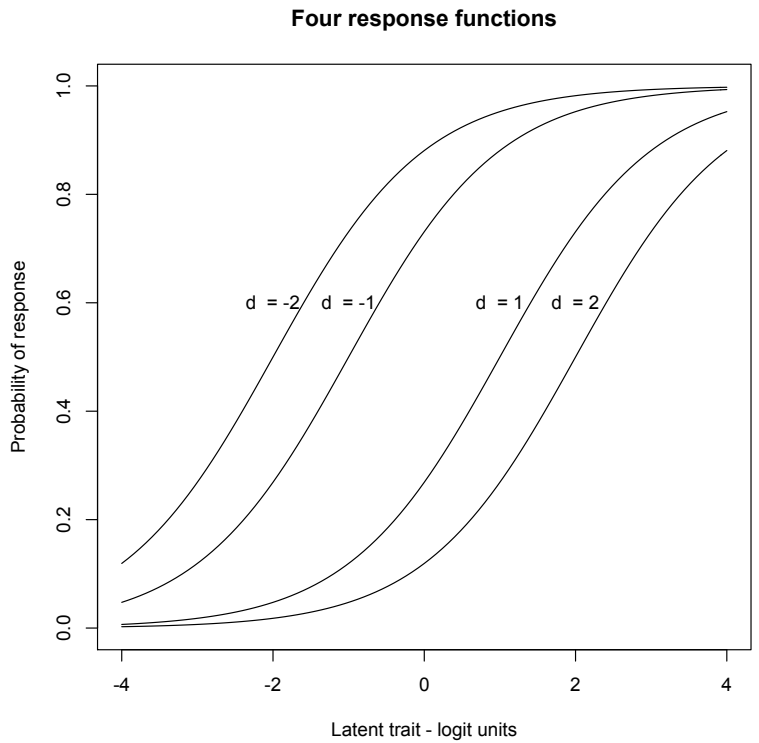
For the 1PL or 2PL logistic model the probability of endorsing the k^{th} response is a function of ability, item thresholds, and the discrimination parameter and is

$$P(r = k|\theta_i, \delta_k, \delta_{k-1}, \alpha_k) = P(r|\theta_i, \delta_{k-1}, \alpha_k) - P(r|\theta_i, \delta_k, \alpha_k) = \frac{1}{1 + e^{\alpha_k(\delta_{k-1} - \theta_i)}} - \frac{1}{1 + e^{\alpha s_k(\delta_k - \theta_i)}} \quad (20)$$

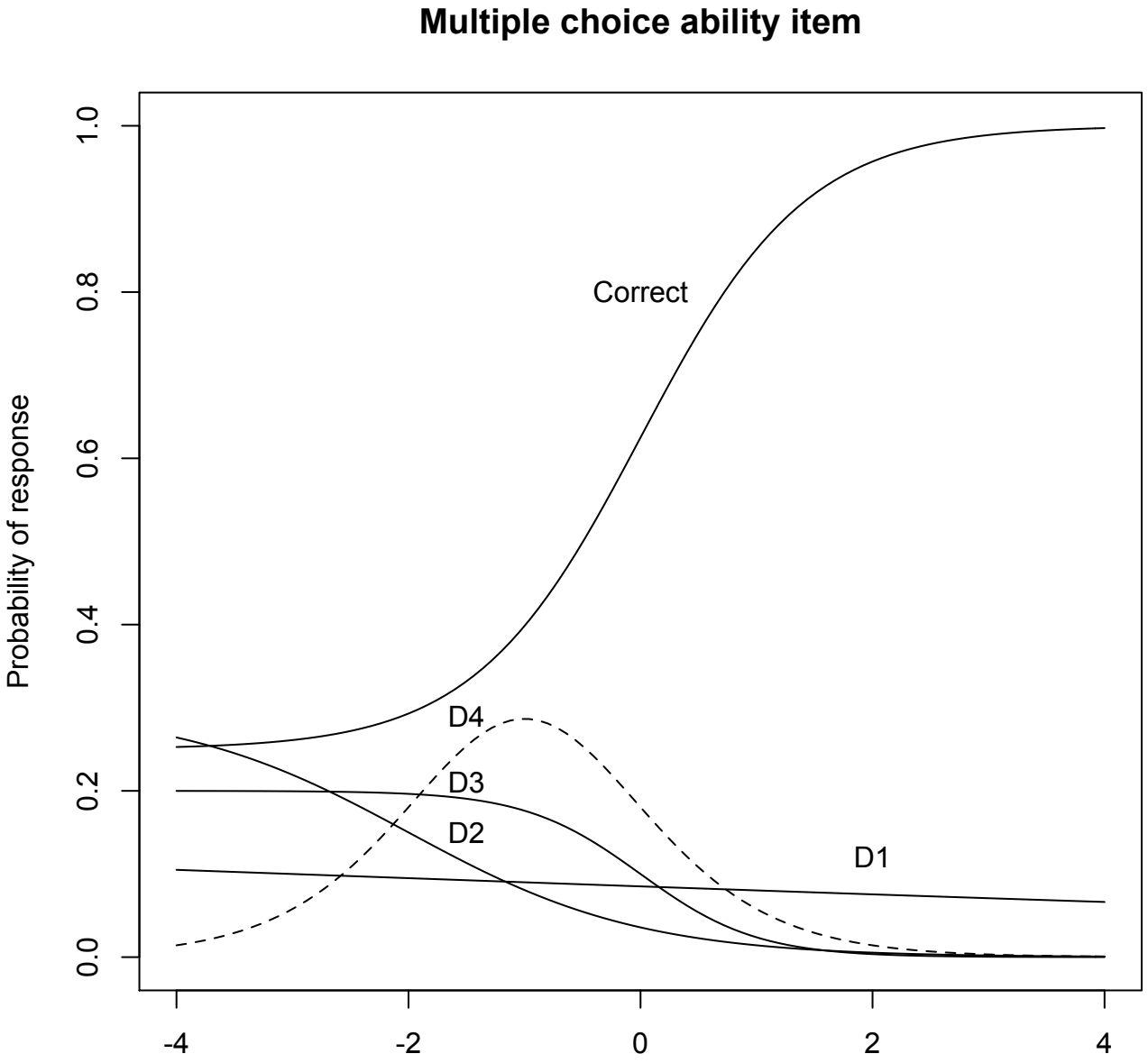
where all b_k are set to $b_k = 1$ in the 1PL Rasch case.

Responses to a multiple choice polytomous item



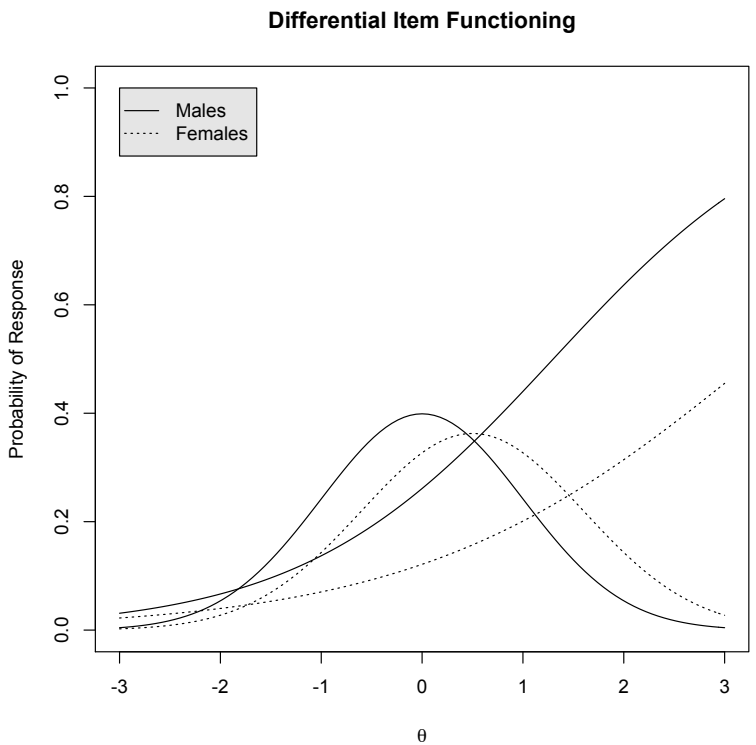


Differences in the response shape of mulitple choice items



Differential Item Functioning

1. Use of IRT to analyze item quality
   - Find IRT difficulty and discrimination parameters for different groups
   - Compare response patterns



FA and IRT

If the correlations of all of the items reflect one underlying latent variable, then factor analysis of the matrix of tetrachoric correlations should allow for the identification of the regression slopes (α) of the items on the latent variable. These regressions are, of course just the factor loadings. Item difficulty, δ_j and item discrimination, α_j may be found from factor analysis of the tetrachoric correlations where λ_j is just the factor loading on the first factor and τ_j is the normal threshold reported by the `tetrachoric` function (**???**).

$$\delta_j = \frac{D\tau}{\sqrt{1-\lambda_j^2}}, \qquad \alpha_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{21}$$

where D is a scaling factor used when converting to the parameterization of *logistic* model and is 1.702 in that case and 1 in the case of the normal ogive model.

FA and IRT

IRT parameters from FA

$$\delta_j = \frac{D\tau}{\sqrt{1-\lambda_j^2}}, \qquad \alpha_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{22}$$

FA parameters from IRT

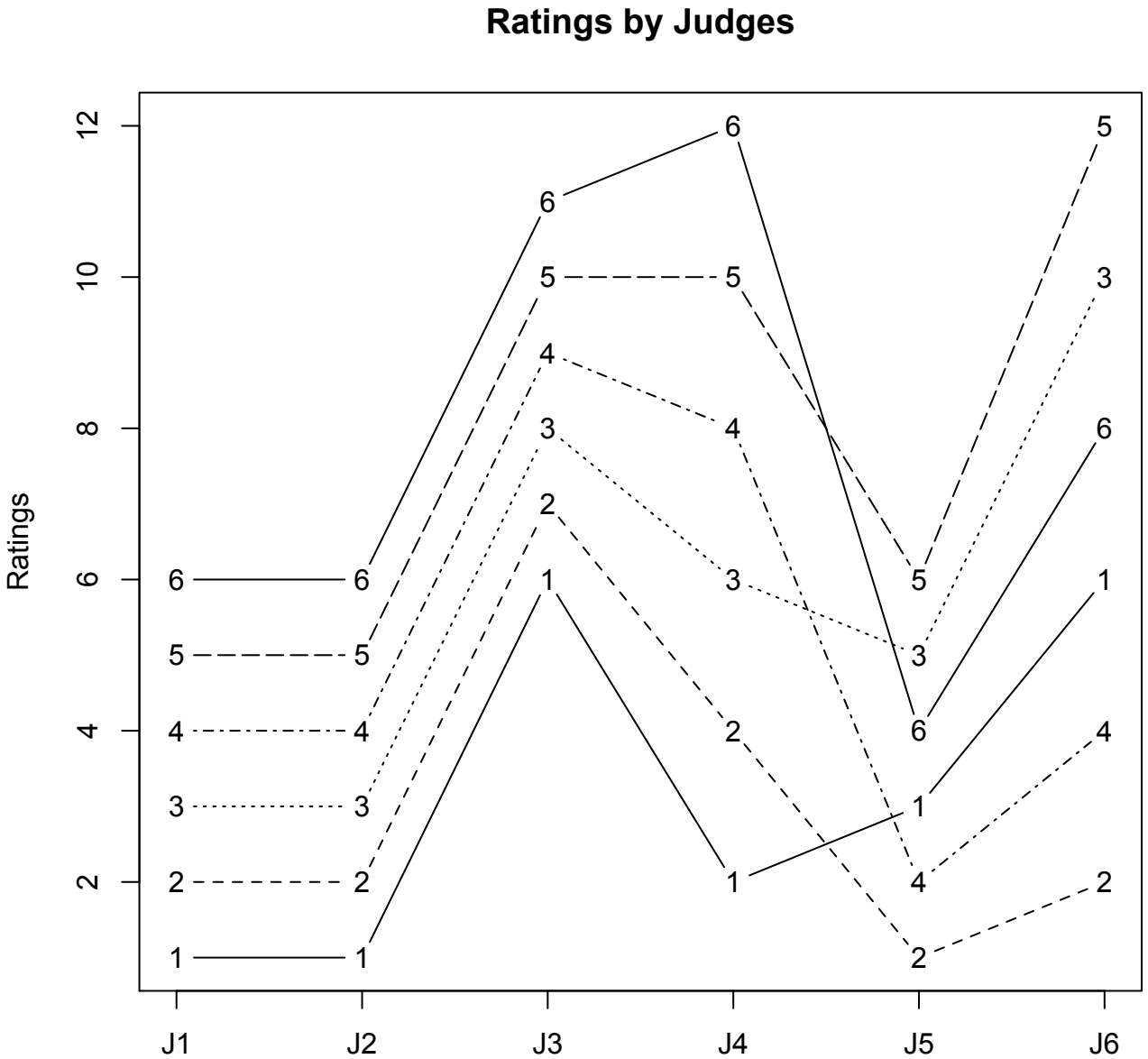
$$\lambda_j = \frac{\alpha_j}{\sqrt{1+\alpha_j^2}}, \qquad \tau_j = \frac{\delta_j}{\sqrt{1+\alpha_j^2}}.$$

the irt.fa function

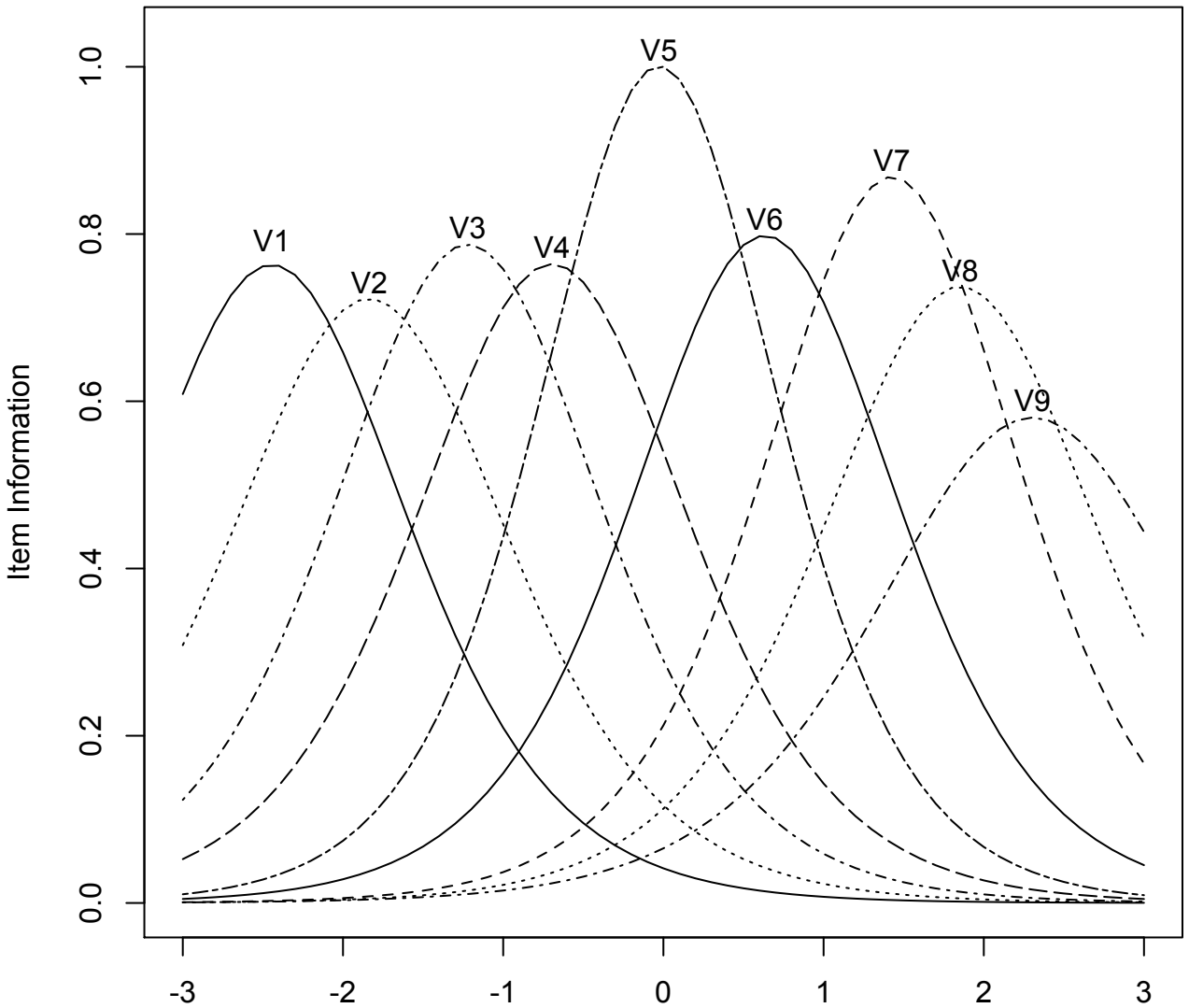
```
> set.seed(17)
> items <- sim.npn(9,1000,low=-2.5,high=2.5)$items
> p.fa <-irt.fa(items)
```

```
Summary information by factor and item
 Factor =  1
                -3   -2   -1    0    1    2     3
V1            0.61 0.66 0.21 0.04 0.01 0.00  0.00
V2            0.31 0.71 0.45 0.12 0.02 0.00  0.00
V3            0.12 0.51 0.76 0.29 0.06 0.01  0.00
V4            0.05 0.26 0.71 0.54 0.14 0.03  0.00
V5            0.01 0.07 0.44 1.00 0.40 0.07  0.01
V6            0.00 0.03 0.16 0.59 0.72 0.24  0.05
V7            0.00 0.01 0.04 0.21 0.74 0.66  0.17
V8            0.00 0.00 0.02 0.11 0.45 0.73  0.32
V9            0.00 0.00 0.01 0.07 0.25 0.55  0.44
Test Info     1.11 2.25 2.80 2.97 2.79 2.28  0.99
SEM           0.95 0.67 0.60 0.58 0.60 0.66  1.01
Reliability 0.10 0.55 0.64 0.66 0.64 0.56 -0.01
```

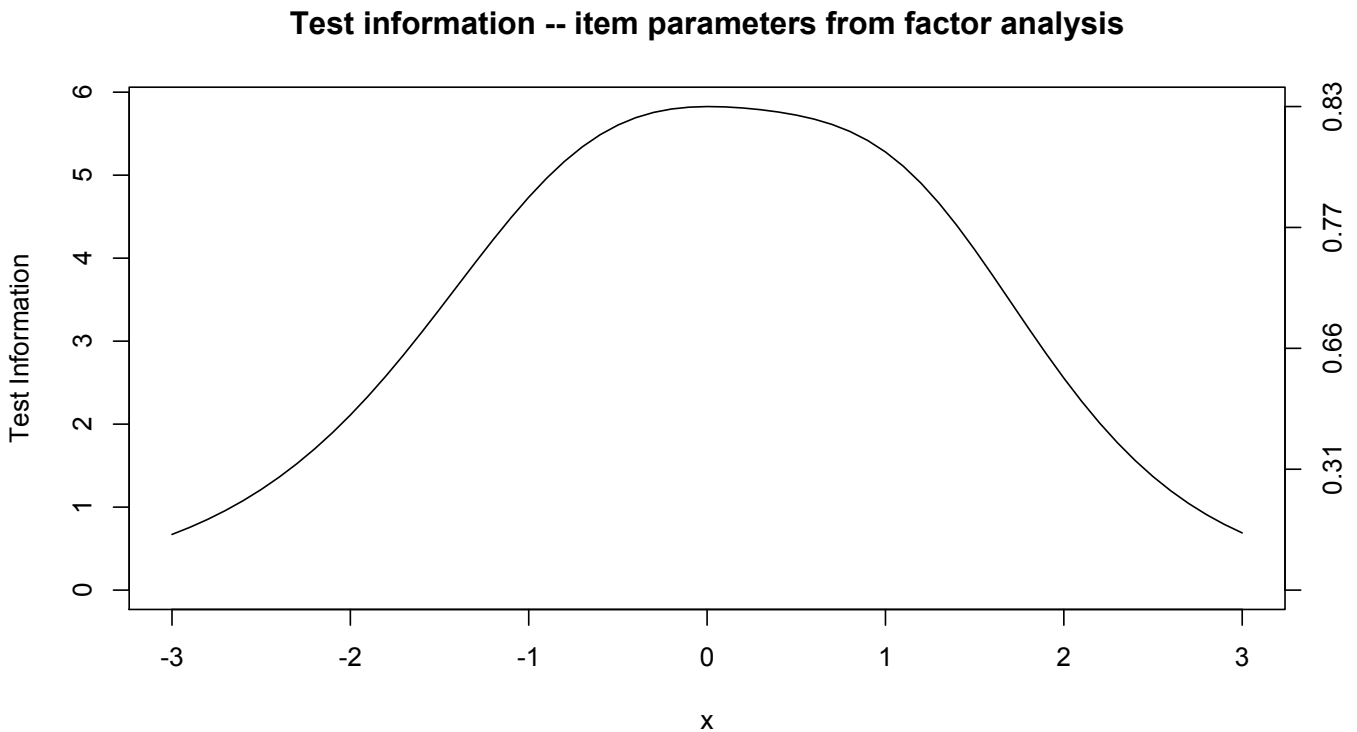
Item Characteristic Curves from FA



Item information from FA



Test Information Curve



Comparing three ways of estimating the parameters

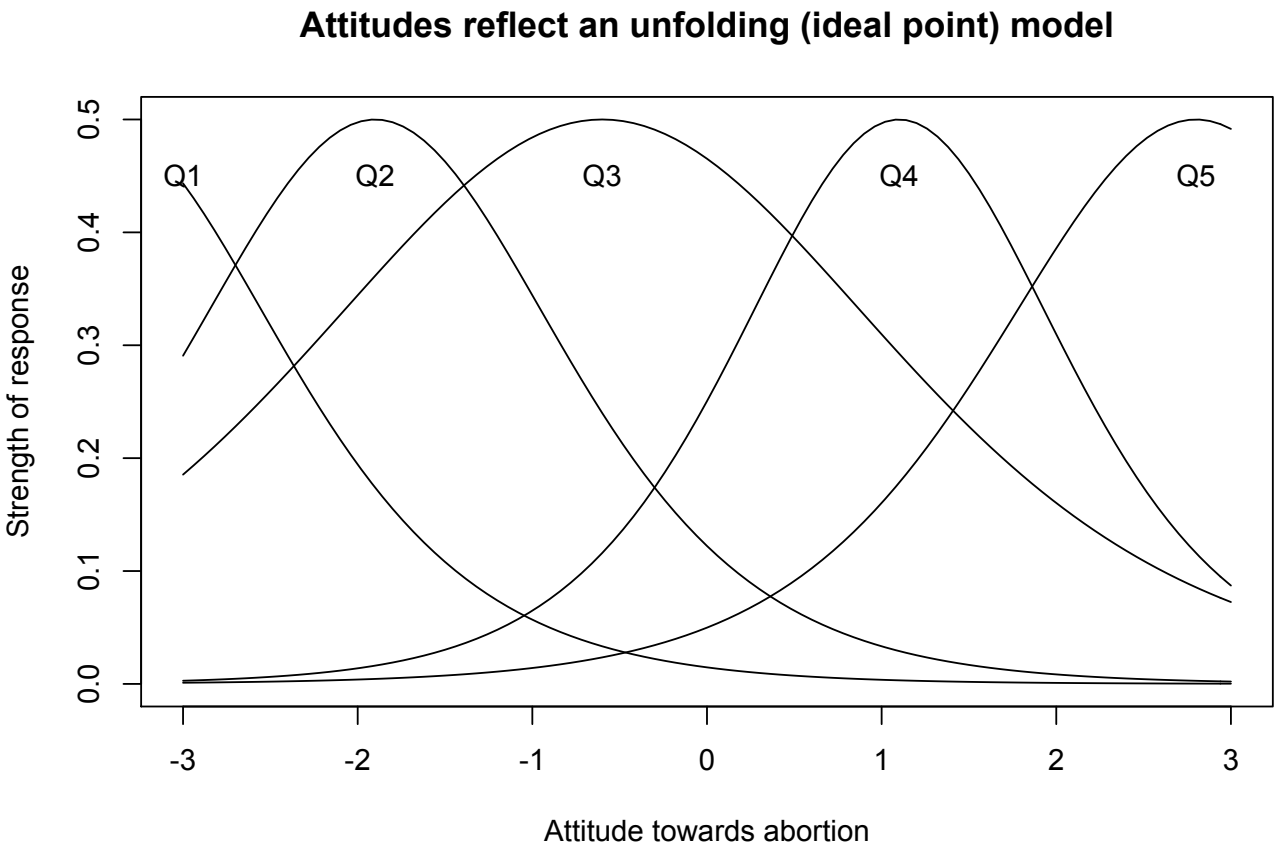
```
set.seed(17)
items <- sim.npn(9,1000,low=-2.5,high=2.5)$items
p.fa <- irt.fa(items)$coefficients[1:2]
p.ltm <- ltm(items~z1)$coefficients
p.ra <- rasch(items, constraint = cbind(ncol(items) + 1, 1))$coefficients
a <- seq(-2.5,2.5,5/8)
p.df <- data.frame(a,p.fa,p.ltm,p.ra)
round(p.df,2)
```

```
           a Difficulty Discrimination X.Intercept.   z1 beta.i beta
Item 1 -2.50      -2.45           1.03         5.42 2.61   3.64    1
Item 2 -1.88      -1.84           1.00         3.35 1.88   2.70    1
Item 3 -1.25      -1.22           1.04         2.09 1.77   1.73    1
Item 4 -0.62      -0.69           1.03         1.17 1.71   0.98    1
Item 5  0.00      -0.03           1.18         0.04 1.94   0.03    1
Item 6  0.62       0.63           1.05        -1.05 1.68  -0.88    1
Item 7  1.25       1.43           1.10        -2.47 1.90  -1.97    1
Item 8  1.88       1.85           1.01        -3.75 2.27  -2.71    1
Item 9  2.50       2.31           0.90        -5.03 2.31  -3.66    1
```

Attitudes might not have monotone trace lines

1. *Abortion is unacceptable under any circumstances.*
2. *Even if one believes that there may be some exceptions, abortions is still generally wrong.*
3. *There are some clear situations where abortion should be legal, but it should not be permitted in all situations.*
4. *Although abortion on demand seems quite extreme, I generally favor a woman's right to choose.*
5. *Abortion should be legal under any circumstances.*

Ideal point models of attitutude



IRT and CTT don't really differ except

1. Correlation of classic test scores and IRT scores $> .98$.
2. Test information for the person doesnt't require people to vary
3. Possible to item bank with IRT
   - Make up tests with parallel items based upon difficulty and discrimination
   - Detect poor items
4. Adaptive testing
   - No need to give a person an item that they will almost certainly pass (or fail)
   - Can tailor the test to the person
   - (Problem with anxiety and item failure)