Whats' beyond Concerto: An introduction to the R package *catR*

Session 1:

Overview of basic notions (CAT, IRT, dichotomous IRT models)

The Psychometrics Centre, Cambridge, June 10th, 2014

Outline:

1. CAT principles
2. Basic notions of item response theory (IRT)
   (a) Assumptions of IRT
   (b) IRT models
   (c) Model calibration
   (d) Proficiency estimation
   (e) Item and test information
3. Back to CAT

1. CAT principles

Like most psychometric methods, CAT aims at estimating some person latent trait (proficiency, ability...) by means of the responses provided to a set of items

Two main administration schemes:

- linear testing: all test takers receive the same set of items (possibly with different item ordering)
- adaptive testing: items are selected iteratively and administered in order to optimally estimate each test taker's proficiency

Computerized adaptive testing (CAT): adaptive administration with computer-based routines and item banks

1. CAT principles

Item bank: collection of items from which one can sample and administer items in a CAT framework

An item bank must be calibrated before a CAT is performed

Item calibration: estimation of item parameters arising from some pre-selected item response model (see later)

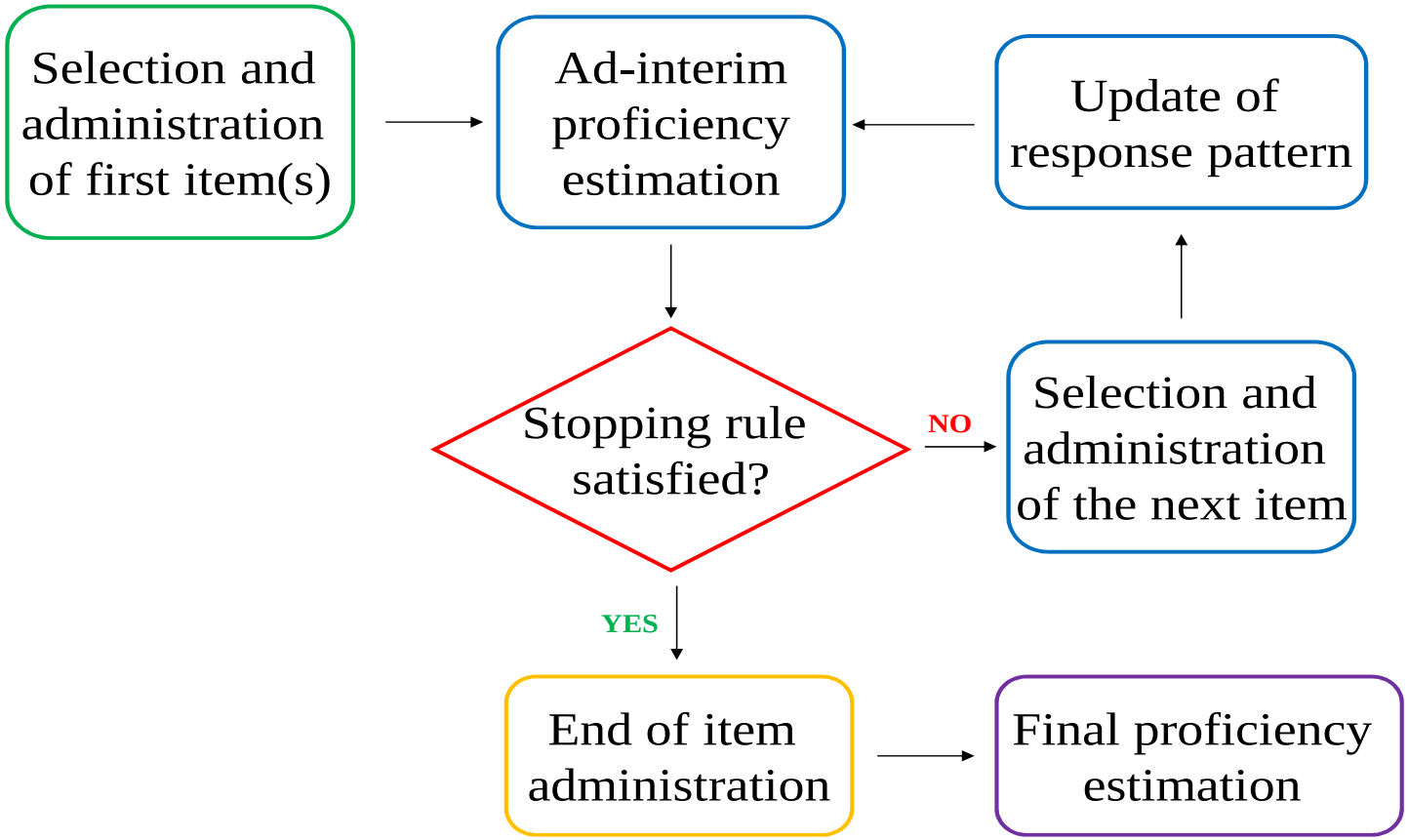
In linear testing, items may not be calibrated in advance (calibration after response patterns are sampled)

1. CAT principles

Any CAT process can be skematically split in four steps:

1. starting step: first item(s) is (are) selected and administered, and item responses are collected (no proficiency estimation)
2. test step: iterative process of proficiency estimation, next item selection and administration
3. stopping step: end of adaptive process once some stopping rule is satisfied
4. final step: final proficiency estimation and output return

1. CAT principles



1. CAT principles

All aspects of a CAT will be further described...

... but first we need more insight on IRT!

2. Basic notions of item response theory (IRT)

IRT aims at providing models that describe the probability of each possible item response, given test taker's proficiency level and item characteristics (i.e. item parameters)

Number and type of item parameters vary from one model to another (see later)

In this session one restricts to dichotomous item responses: only correct or incorrect responses!

2. (a) Assumptions of IRT

Classical IRT models require three assumptions (assumed to hold in this workshop):

1. Unidimensionality: each test taker's proficiency level is characterized by a single latent trait (i.e. only one latent dimension is targeted by the test)
2. Local independence: at given proficiency level, item responses are independent (i.e. all responses of one test taker are assumed to be independent from each other)
3. Monotonicity: the probability of answering an item correctly is a monotone (i.e. non-decreasing) function of the proficiency level

2. (a) Assumptions of IRT

Unidimensionality assumption can be relaxed by introducing multidimensional IRT models (see e.g., Reckase, 2009)

Local independence may not hold when e.g. items are nested in a common stimulus such as in testlets, for which specific models exist (Wainer, Bradlow, & Wang, 2007)

Monotonicity assumption can also be relaxed with very specific IRT models (see e.g., van der Linden & Hambleton, 1997)

2. (b) IRT models

With dichotomous item responses and the three previous assumptions, most common IRT models are:

- Rasch or one-parameter logistic (1PL) model
- two-parameter logistic (2PL) model
- three-parameter logistic (3PL) model
- (four-parameter logistic (4PL) model)

θ: proficiency level

j: item of interest

X_j: item response with

$$X_j = \begin{cases} 1 & \text{for a correct response} \\ 0 & \text{for an incorrect response} \end{cases}$$

2. (b) IRT models

Rasch (1PL) model (Rasch, 1960):

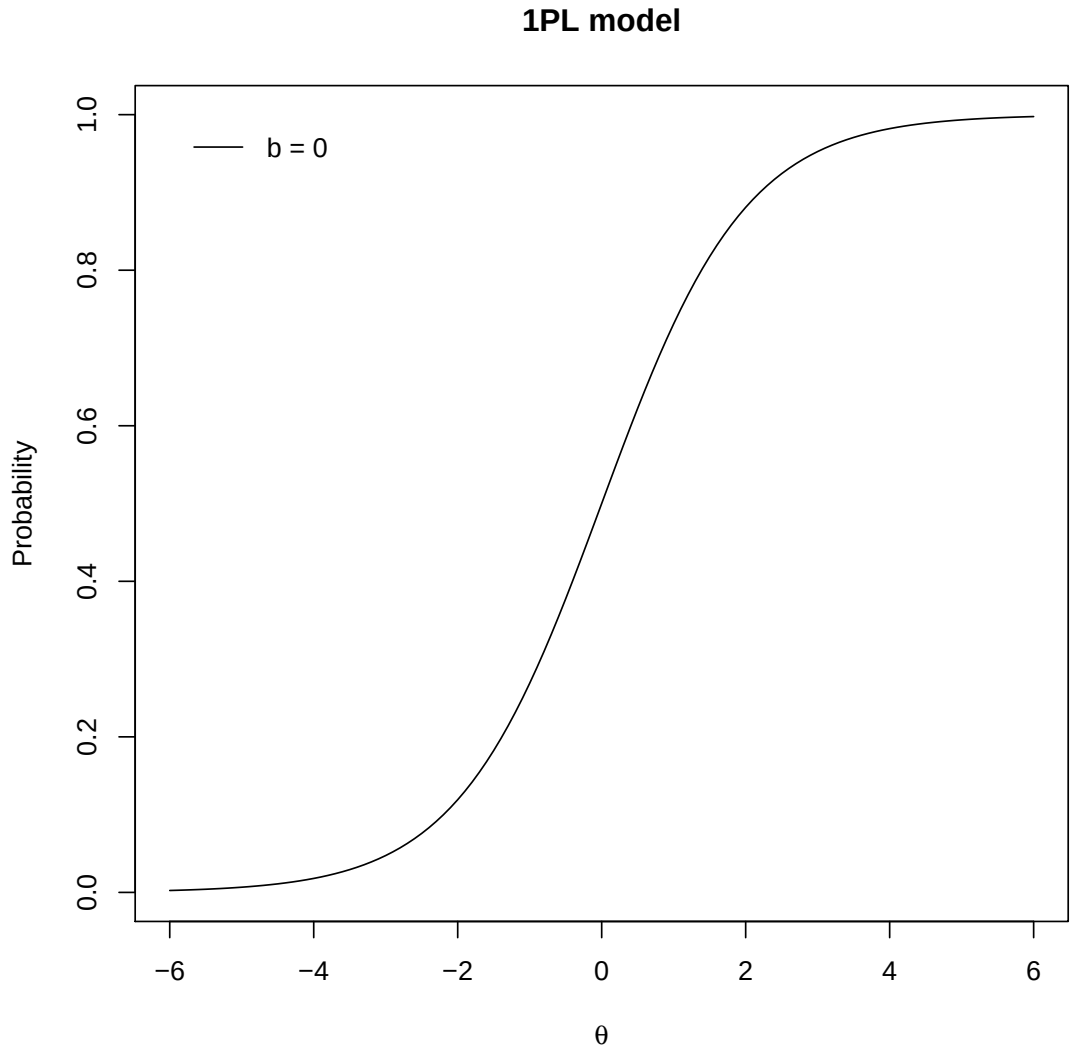
$$Pr(X_j = 1|\theta, b_j) = \frac{\exp\left[D\left(\theta - b_j\right)\right]}{1 + \exp\left[D\left(\theta - b_j\right)\right]}$$

b_j: item difficulty level

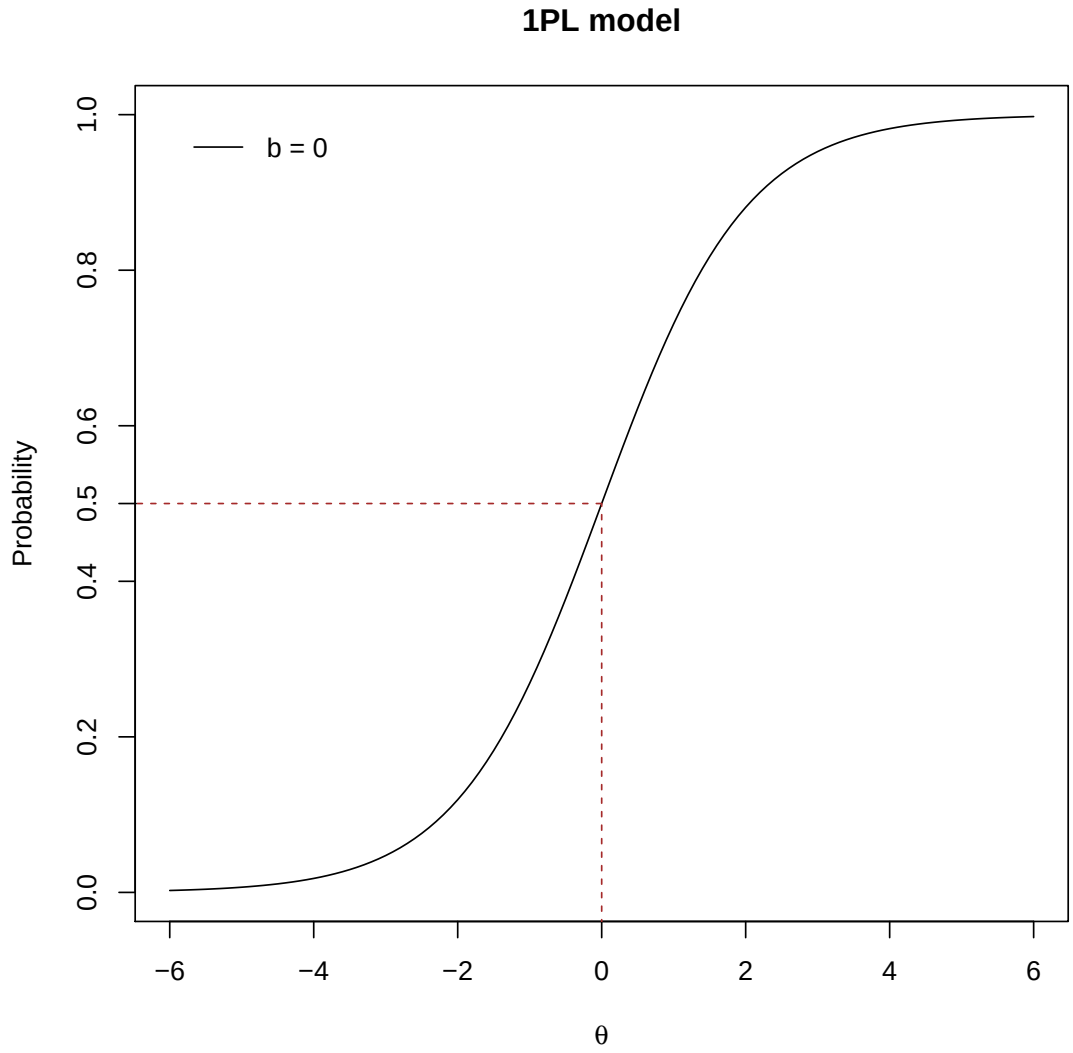
D: scaling constant (set to 1 for logistic metric and 1.7 for normal metric; see Haley, 1952)

$Pr(X_j = 1|\theta, b_j)$ has a logistic shape and b_j controls the location of the response probability curve (item response function)

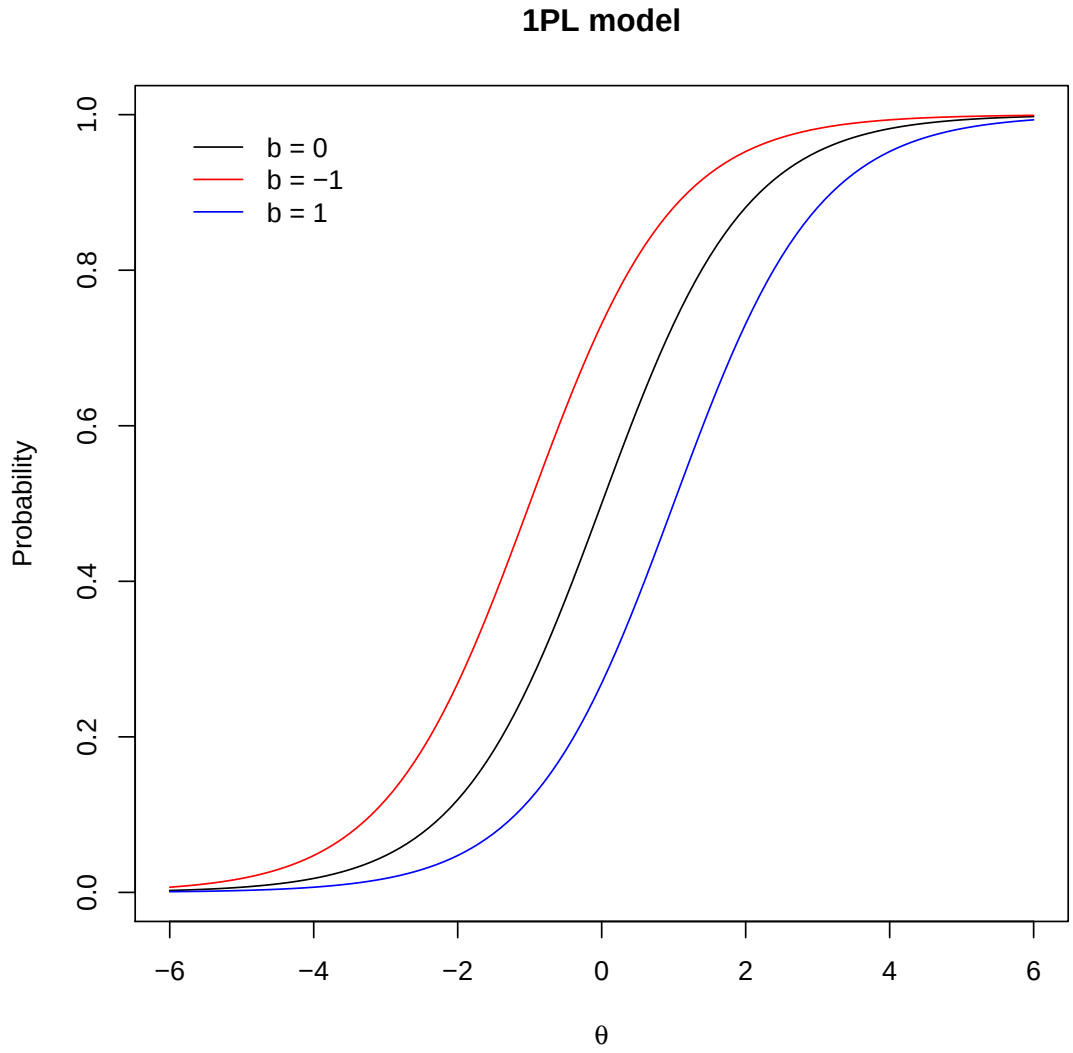
2. (b) IRT models



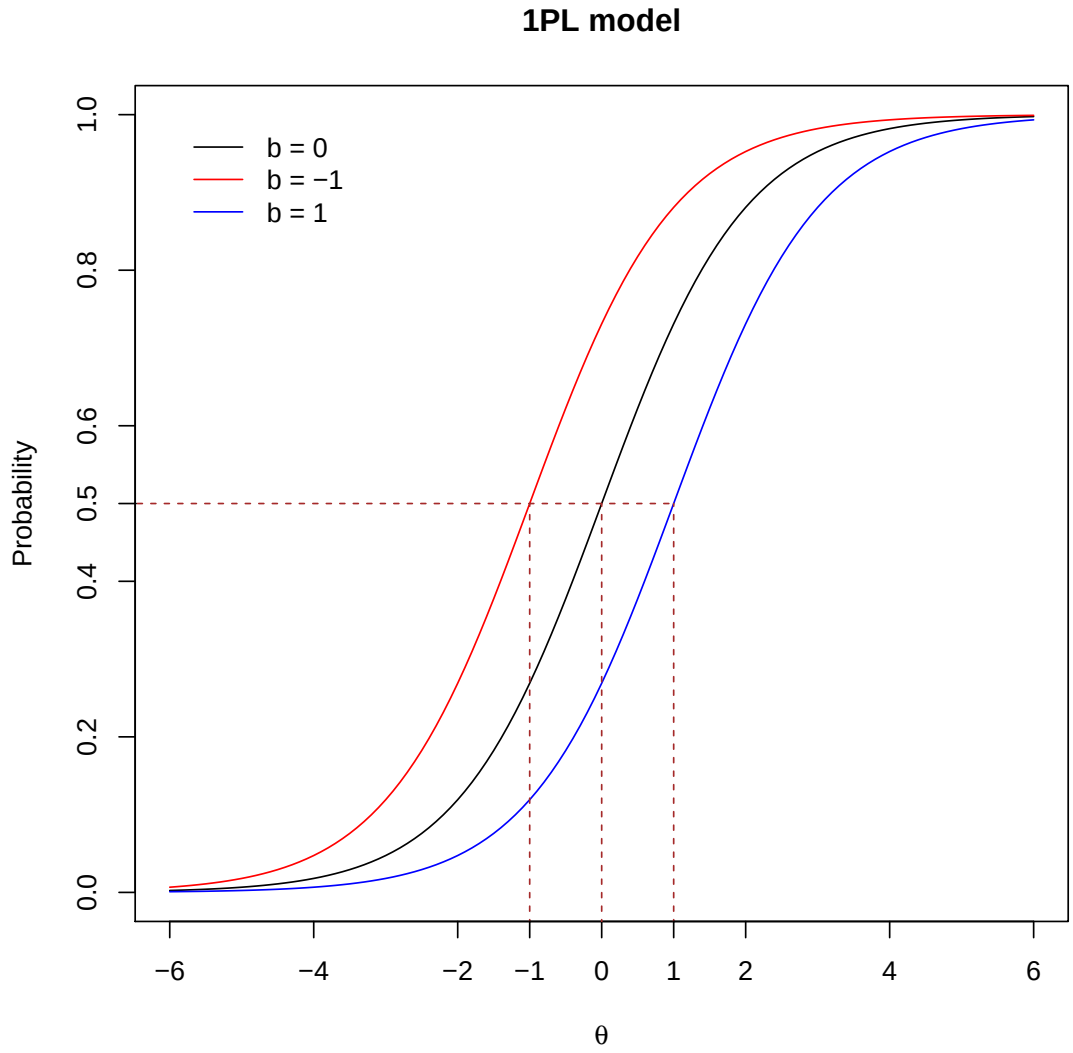
2. (b) IRT models



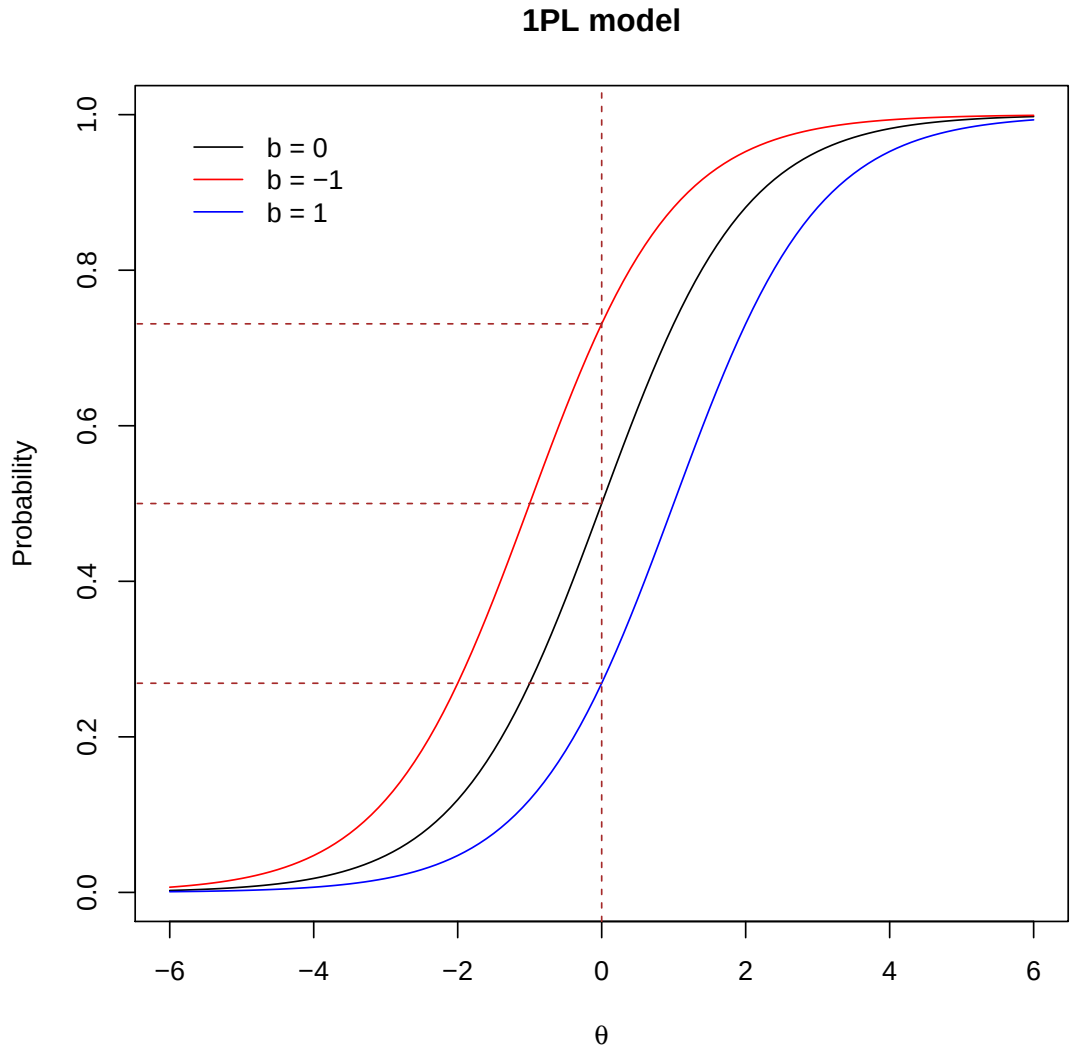
2. (b) IRT models



2. (b) IRT models



2. (b) IRT models



2. (b) IRT models

2PL model (Birnbaum, 1968):

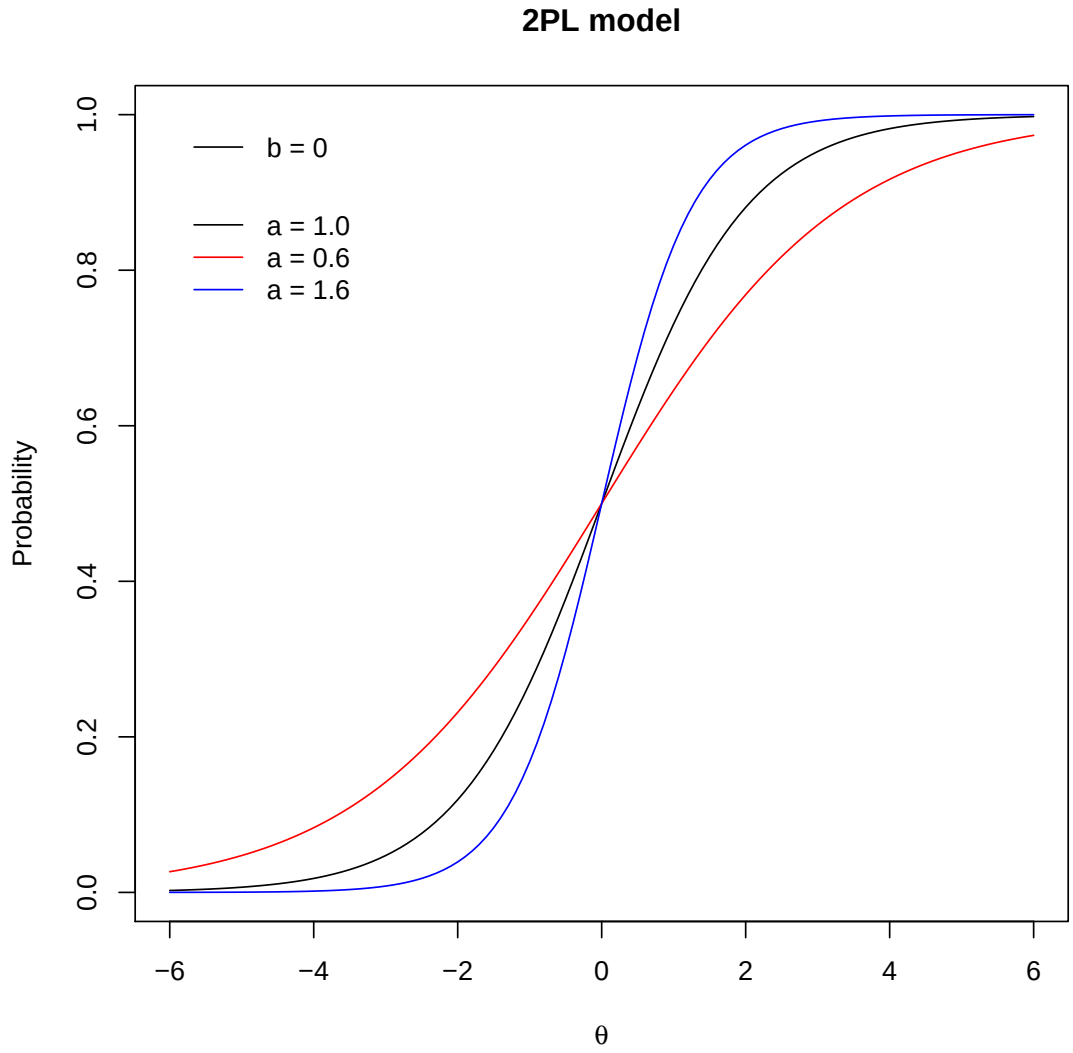
$$Pr(X_j = 1|\theta, a_j, b_j) = \frac{\exp\left[D\, a_j\, (\theta - b_j)\right]}{1 + \exp\left[D\, a_j\, (\theta - b_j)\right]}$$

a_j: item discrimination level

a_j controls for the slope of the logistic curve

All a_j values equal across items $\Rightarrow$ back to Rasch model

2. (b) IRT models



2. (b) IRT models

3PL model (Birnbaum, 1968):

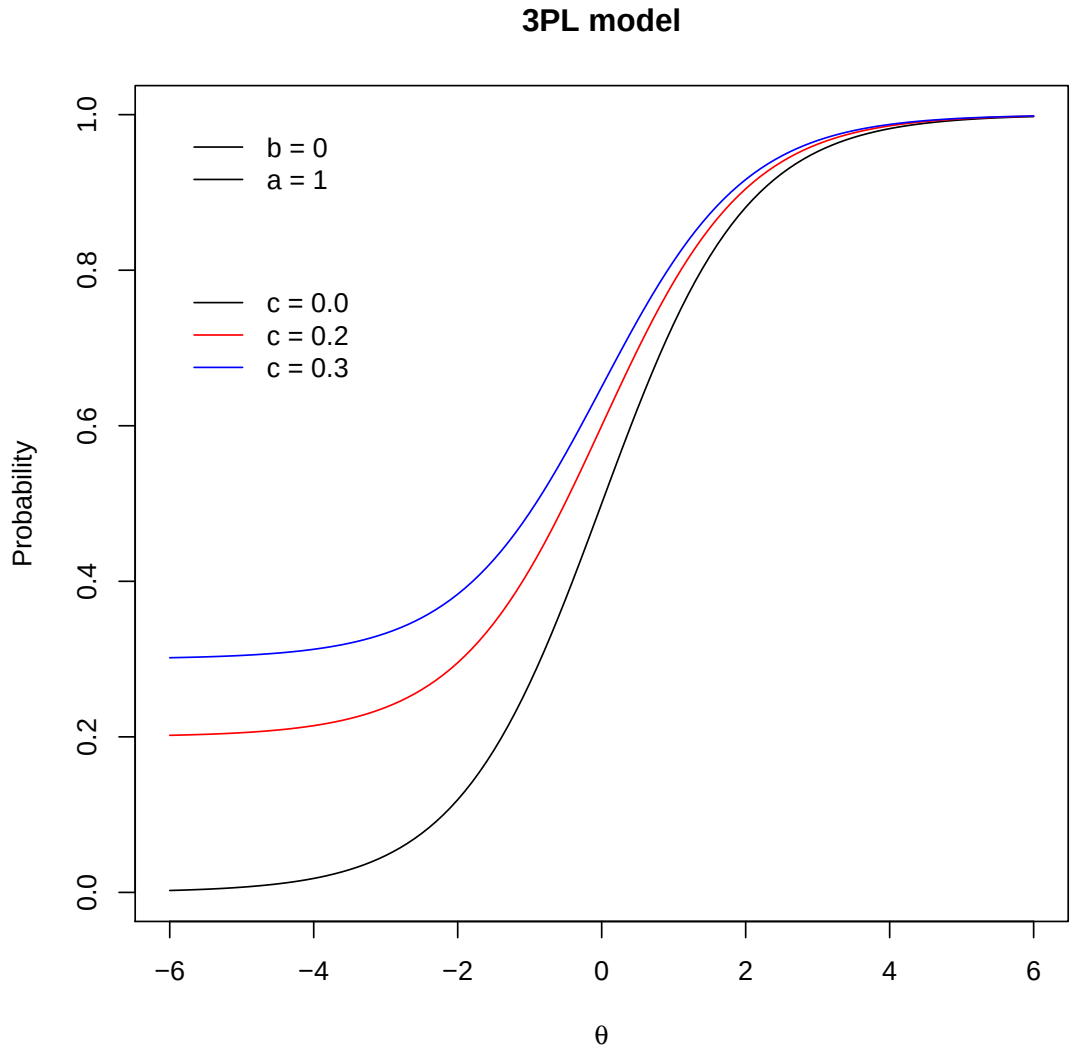
$$Pr(X_j = 1|\theta, a_j, b_j, c_j) = c_j + (1 - c_j)\,\frac{\exp\left[D\,a_j\,(\theta - b_j)\right]}{1 + \exp\left[D\,a_j\,(\theta - b_j)\right]}$$

c_j: pseudo-guessing level (lower asymptote)

Idea: when $c_j > 0$, even test takers with low proficiency have non-zero probability of answering the item correctly

All c_j equal to zero $\Rightarrow$ back to the 2PL model

2. (b) IRT models



2. (b) IRT models

4PL model (Barton & Lord, 1980):

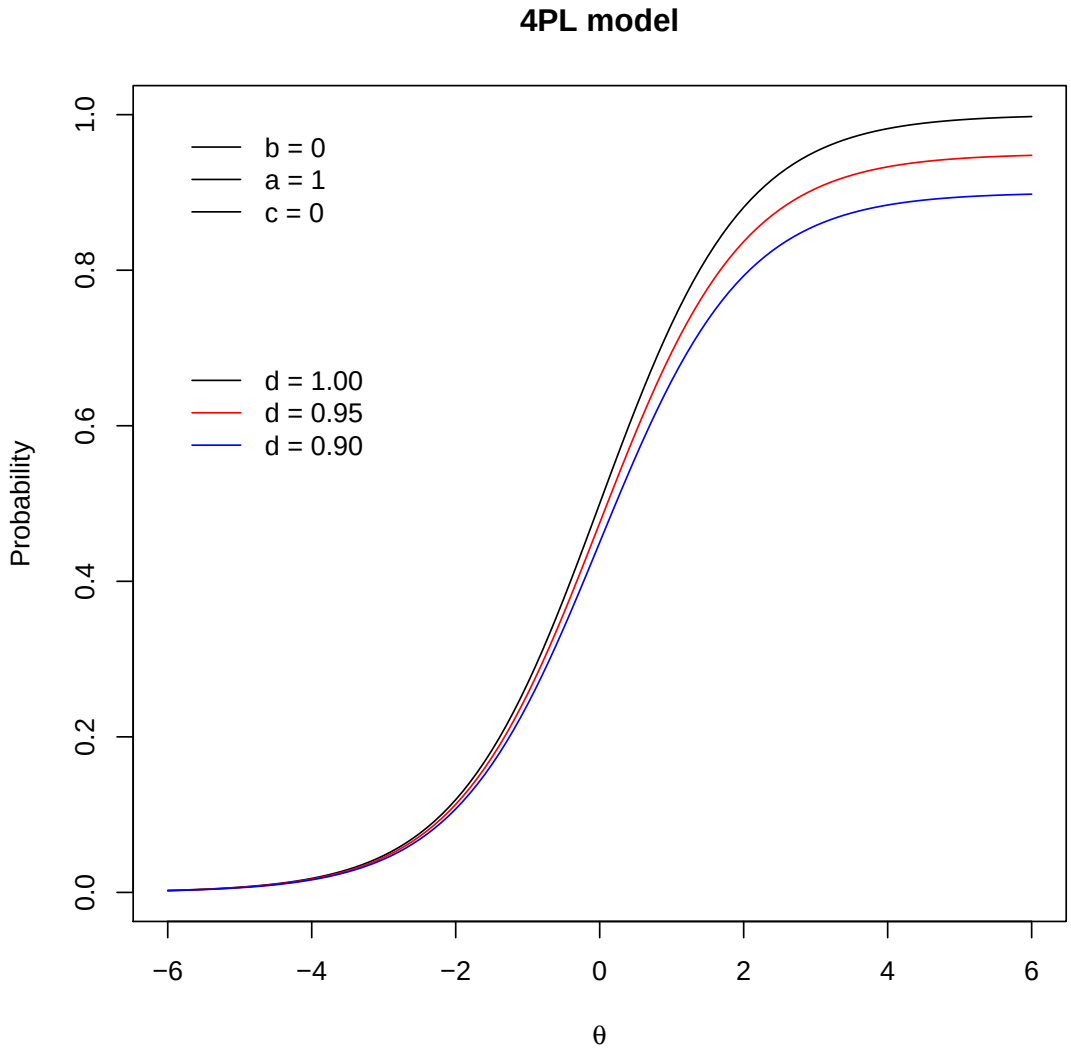
$$Pr(X_j = 1|\theta, a_j, b_j, c_j, d_j) = c_j + (d_j - c_j) \frac{\exp\left[D\, a_j\, (\theta - b_j)\right]}{1 + \exp\left[D\, a_j\, (\theta - b_j)\right]}$$

d_j: inattention level (upper asymptote)

Idea: when $d_j < 1$, even test takers with high proficiency have non-zero probability of answering the item incorrectly

All d_j equal to one $\Rightarrow$ back to the 3PL model

2. (b) IRT models



2. (c) Model calibration

Calibration of an IRT model: estimation of item parameters from a data set of collected item responses

Requires advanced computer software and routines

Usual calibration methods:

- joint maximum likelihood (Lord & Novick, 1968)
- conditional maximum likelihood (Andersen, 1970) - only for Rasch model
- marginal maximum likelihood (Bock & Aitkin, 1981)

Possible to test for accuracy of item calibration

In the following, item parameters are assumed to be known (i.e. item banks are calibrated)

2. (d) Proficiency estimation

Several ability estimation methods exist with dichotomous item responses:

- maximum likelihood (ML; Lord, 1980)
- Bayes modal (BM) or maximum a posteriori (MAP; Birnbaum, 1969)
- expected a posteriori (EAP; Bock & Mislevy, 1982)
- weighted likelihood (WL; Warm, 1989)

2. (d) Proficiency estimation

Maximum likelihood estimation: to find out the value of θ that is most likely (given item responses and item parameters)

Likelihood function with n items:

$$L(\theta) = \prod_{j=1}^{n} Pr(X_j = 1|\theta, \boldsymbol{p}_j)^{X_j} \left[1 - Pr(X_j = 1|\theta, \boldsymbol{p}_j)\right]^{1-X_j}$$

with $\boldsymbol{p}_j$ the set of item parameters (varies from one model to another)

Value $\hat{\theta}_{ML}$ of θ that maximizes $L(\theta)$ is the ML estimate of proficiency

Optimization routine (e.g. Newton-Raphson) is necessary

2. (d) Proficiency estimation

Maximum a posteriori estimation: to find out the value of θ that maximizes the posterior distribution of θ:

$$g(\theta) = f(\theta) \times L(\theta)$$

with $f(\theta)$ the prior distribution (or density) of θ

$L(\theta)$ is data (and test) driven but $f(\theta)$ must be specified *a priori*

$f(\theta)$ reflects some prior belief on the distribution of proficiency levels

Usual choices for $f(\theta)$: uniform distribution, normal distribution, Jeffreys' prior (Jeffreys, 1946)

Value $\hat{\theta}_{MAP}$ of θ that maximizes $g(\theta)$ is the MAP estimate of proficiency

2. (d) Proficiency estimation

Expected a posteriori estimation: compute the posterior mean of θ:

$$\hat{\theta}_{EAP} = \frac{\int_{-\infty}^{+\infty} \theta \, f(\theta) \, L(\theta) \, d\theta}{\int_{-\infty}^{+\infty} f(\theta) \, L(\theta) \, d\theta}$$

with $f(\theta)$ the prior distribution (or density) of θ

Usual choices for $f(\theta)$: uniform distribution, normal distribution, Jeffreys' prior (Jeffreys, 1946)

Integrals can be approximated by numerical techniques (e.g. gaussian quadrature)

2. (d) Proficiency estimation

Weighted likelihood estimation: correct for bias in ML estimation by introducing some weighted correction to the likelihood function

$\hat{\theta}_{WL}$ value of θ that maximizes the weighted likelihood function is the WL estimate of proficiency

2. (d) Proficiency estimation

Together with (point) proficiency estimates, standard errors (SEs) can be computed

SE is a measure of precision of the proficiency estimator

The smaller the SE the more precise the estimation of proficiency

SE usually decreases with test length...

... and often directly relates to the test information function

2. (e) Item and test information

Each item is most informative for specific range of proficiency

Item information function (IIF) for item j:

$$I_j(\theta) = \frac{P_j'(\theta)^2}{P_j(\theta)\left[1 - P_j(\theta)\right]}$$

with $P_j(\theta) = Pr(X_j = 1|\theta, \boldsymbol{p}_j)$ and $P_j'(\theta)$ is the first derivative of $P_j(\theta)$ w.r.t. θ

With Rasch model, $I_j(\theta) = P_j(\theta)\left[1 - P_j(\theta)\right]$ and is maximized whenever

$$P_j(\theta) = \frac{\exp\left[D\left(\theta - b_j\right)\right]}{1 + \exp\left[D\left(\theta - b_j\right)\right]} = 0.5 \quad \text{or} \quad \theta = b_j$$

$\Rightarrow$ Item most informative for proficiencies close to difficulty level

2. (e) Item and test information

Test information function (TIF) is the sum of all IIF:

$$TIF(\theta) = \sum_{j=1}^{n} I_j(\theta)$$

TIF indicates how informative the whole test is

Since IIF are always positive, TIF increases with test length

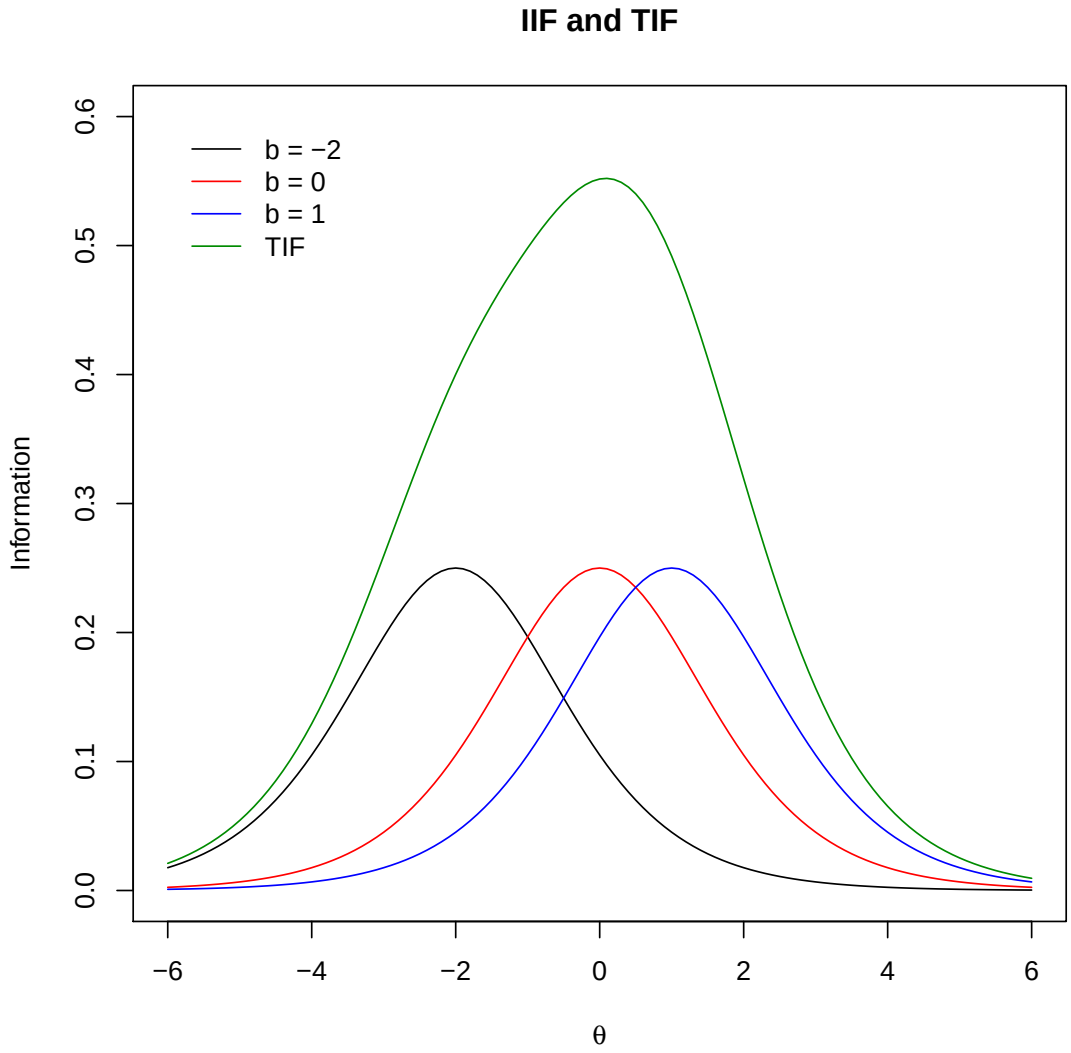
SE of ML estimator $\hat{\theta}_{ML}$ is given by

$$SE(\hat{\theta}_{ML}) = \frac{1}{\sqrt{TIF(\hat{\theta}_{ML})}}$$

$\Rightarrow$ The longer the test, the larger the TIF, the smaller the SE, the more precise the ML estimate

Specific SE formulas for other estimators

2. (e) Item and test information



3. Back to CAT

Six aspects will be further looked at:

- selection of first item(s)
- ad-interim proficiency estimation
- next item selection
- stopping rules
- item exposure control
- content balancing control

3. Back to CAT

Selection of first item(s):

- at random
- fixed by the administrator
- selected as the most informative in the bank for pre-specified proficiency level(s) (i.e. such that $I_j(\theta)$ is maximal for given θ)
- such that difficulty level(s) is (are) closest to pre-specified proficiency level(s) (i.e. such that $|b_j - \theta|$ is minimal)
- ...

3. Back to CAT

Ad-interim proficiency estimation:

- Maximum likelihood
- Maximum a posteriori
- Expected a posteriori
- Weighted likelihood
- ...
- Prior distributions fort MAP and EAP:
  - uniform
  - normal
  - Jeffreys (non-informative prior based on IIF)
  - ...

3. Back to CAT

Next item selection: plenty of methods (at least 12)!

- Maximum Fisher information (MFI): select item j that maximizes $I_j(\hat{\theta})$ (with $\hat{\theta}$ ad-interim proficiency estimate)
- bOpt criterion (Urry, 1970): select item j such that $|b_j - \hat{\theta}|$ is minimal (equivalent to MFI with Rasch model)
- Maximum likelihood weighted information (MLWI; Veerkamp & Berger, 1997): select item j that maximizes $L(\hat{\theta})\, I_j(\hat{\theta})$ (with $L(\hat{\theta})$ computed with currently administered items)
- Maximum posterior weighted information (MPWI; van der Linden, 1998): select item j that maximizes $f(\hat{\theta})\, L(\hat{\theta})\, I_j(\hat{\theta})$ (with $f(\theta)$ prior distribution)

3. Back to CAT

Next item selection:

- Maximum expected information (MEI; van der Linden, 1998)
- Minimum expected posterior variance (MEPV)
- Kullback-Leibler divergency criterion (KL; Chang & Ying, 1996): select item j that minimizes a weighted form of the KL information
- Posterior Kullback-Leibler divergency criterion (KLP; Chang & Ying, 1996): select item j that minimizes the posterior weighted form of the KL information
- Random selection

3. Back to CAT

Next item selection:

- Progressive method (Barrada, Olea, Ponsoda, & Abad, 2008, 2010): select item j that maximizes a weighted sum of two elements, a random selection component and an item information component:
  - at early stage of CAT, random component is most weighted
  - during CAT, random component gets downweighted and item information becomes more weighted
  - at end of CAT, only item information selection
- Proportional method (Barrada, Olea, Ponsoda, & Abad, 2008, 2010): random selection of items with selection probability related to their information at current $\hat{\theta}$

3. Back to CAT

Stopping rules:

- Length: stop when K items were administered
- Precision: stop when the precision on ad-interim proficiency estimate $\hat{\theta}$ is good enough, or when $SE(\hat{\theta}) < t$
- Classification: stop when proficiency level can be accurately classified as below or above some threshold T, i.e., when

$$\hat{\theta} - z_{1-\alpha/2}\, SE(\hat{\theta}) > T \quad \text{or} \quad \hat{\theta} + z_{1-\alpha/2}\, SE(\hat{\theta}) < T$$

  and z_α the z-score with lower tail probability α (confidence interval method; Kingsbury & Weiss, 1983)
- Other classification rules: sequential probability ratio test (SPRT; Eggen, 1999) and generalized likelihood ratio (GLR; Thompson, 2009)

3. Back to CAT

Item exposure control:

- Important to ensure that items are not too often administered (content security issue)
- Possible to control for item exposure with several techniques:
  - Randomesque method (Kingsbury & Zara, 1989): select first a small set of optimal items, then randomly pick up and administer one of them
  - Sympson-Hetter method (Sympson & Hetter, 1985)
  - Progressive and proportional methods (for next item selection)

3. Back to CAT

Content balancing control:

- Possible to balance the administered items by selecting them from pre-specified subgroups within the bank
- Selection made to satisfy some predefined ratios of administrations for each subgroup (e.g. 25%, 30%, 15%, 30% for four subgroups)
- Simple method proposed by Kingsbury and Zara (1989):
  - first identify the subgroup from which the next item must be administered (to match as closely as possible the pre-specified rates)
  - then apply the item selection rule into this subgroup only

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society (Series B), 32*, 283-301.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATS. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model.* Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

References

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*, 258-276.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of the EM algorithm. *Psychometrika, 37*, 29-51.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 34*, 438-452.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249–261.

References

Haley, D.C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error.* Technical report no 15. Palo Alto, CA: Applied Mathematics and Statistics Laboratory, Stanford University.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 186*, 453-461.

Kingsbury, G. G., & Weiss (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York, NY: Academic Press.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359–375.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

References

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testingwith the R package catR. *Journal of Statistical Software, 48 (8)*, 1–31.

Reckase, M. D. (2009). *Multidimensional item response models*. New York: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

References

Thompson, N. A. (2009). Using the generalized likelihood ratio as a termination criterion. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing.*

Urry, V. W. (1970). *A Monte Carlo investigation of logistic test models.* Unpublished doctoral dissertation. West Lafayette, IN: Purdue University.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

Van der Linden, W., & Glas, C. A. W. (2009). *Elements of adaptive testing.* New York: Springer.

Van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

References

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, UK: Cambridge University Press.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika, 54*, 427-450.