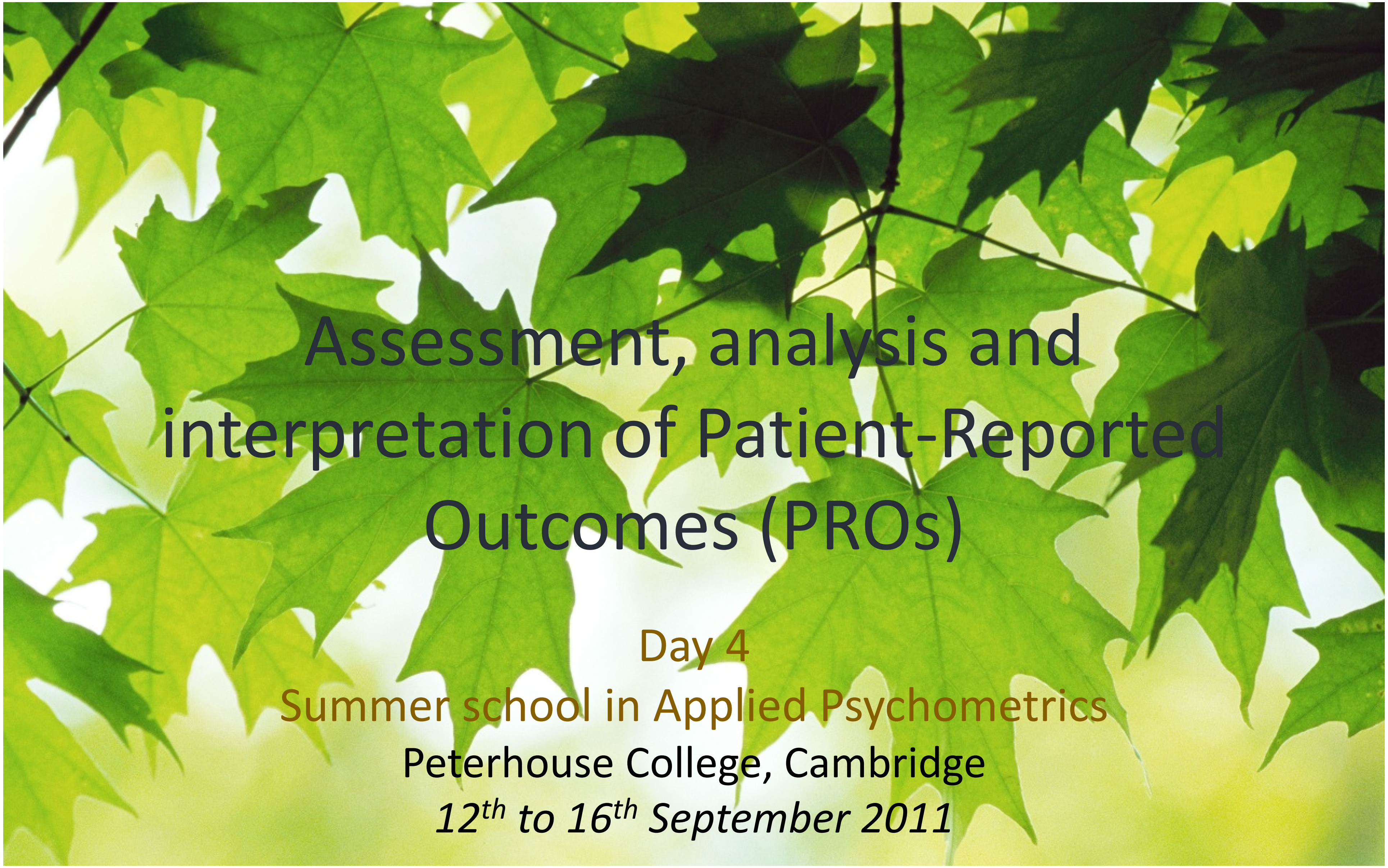


Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

Day 4

Summer school in Applied Psychometrics

Peterhouse College, Cambridge

12th to 16th September 2011

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

This course is prepared by

Anna Brown, PhD ab936@medschl.cam.ac.uk

Jan Stochl, PhD js883@cam.ac.uk

Tim Croudace, PhD tjc39@cam.ac.uk

(University of Cambridge, department of Psychiatry)

Jan Boehnke, PhD boehnke@uni-trier.de

(University of Trier, Department of Clinical Psychology and Psychotherapy)

The course is funded by the ESRC RDI and hosted by The Psychometrics Centre

Jan Boehnke

12. COMPUTER ADAPTIVE TESTING AKA "CAT"

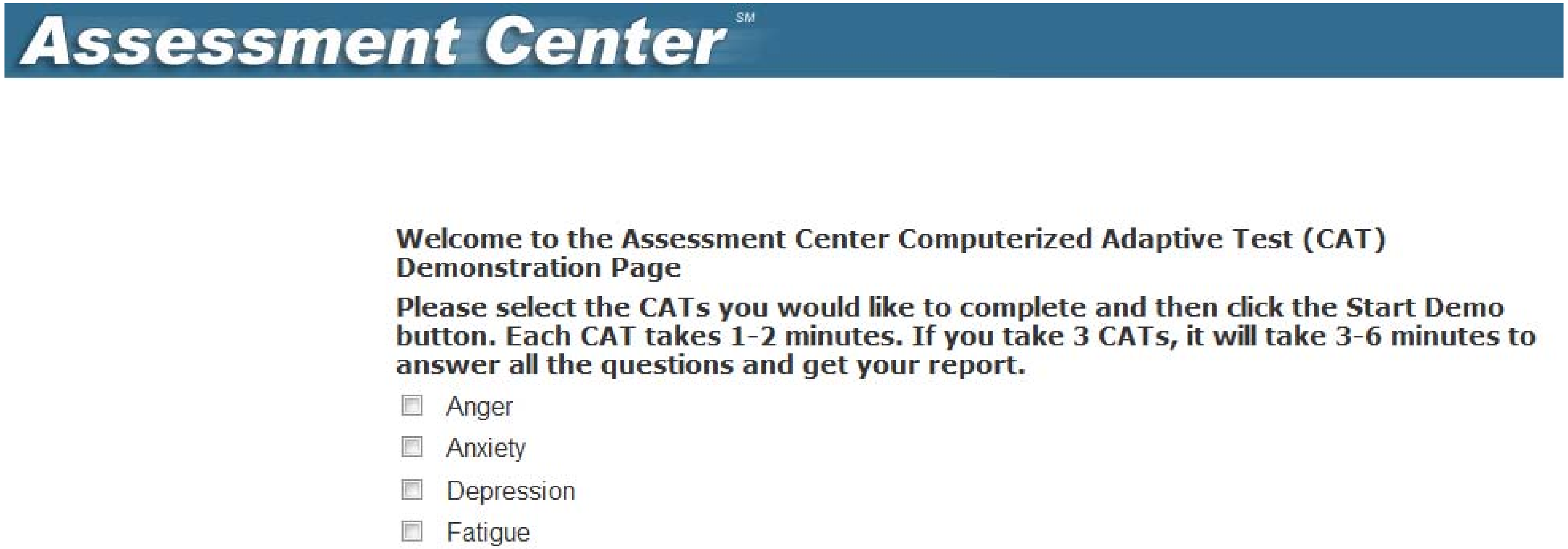
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

PROMIS practical

- together with neighbor

http://www.nihpromis.org/software/demonstration

- "Try a demonstration of the PROMIS CAT"



Assessment Center SM

Welcome to the Assessment Center Computerized Adaptive Test (CAT) Demonstration Page

Please select the CATs you would like to complete and then click the Start Demo button. Each CAT takes 1-2 minutes. If you take 3 CATs, it will take 3-6 minutes to answer all the questions and get your report.

□ Anger

□ Anxiety

□ Depression

□ Fatigue

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Comprehensive assessment in practice

- In many situations the use of lengthy tests is not possible or warranted:
  - repeated assessments during a trial or in therapy
  - patient population cannot be subjected to long tests or many repeated tests (e.g. cancer; Walker, Böhnke, Strasser & Cerny, 2010)
- but lengthy assessments are often needed (e.g. routine testing on several dimensions; each dimension should be estimated as accurate and as fast as possible)

Comprehensive assessment in practice

- In these contexts tests can be shortened adaptively to the situation at hand
- "situation" means (at least) two aspects:
  - – testing purposes of the investigator / provider
  - – variables on the side of the patient: how much can he/she take at the moment?

IRT

- IRT provides a straightforward way to select items
- when a scale is developed according to IRT standards:
  - it is already shown that all items measure one dimension/ construct (multi- is possible); *construct validity* should be established
  - it should be shown that relevant other variables are not influencing the outcome in the items (no DIF)

IRT

- Therefore any item taken from that scale should be a good representation of the trait
  - the items differ not in their content
  - in their clinical validity to assess the outcome / construct dimension in question
  - they only differ in their item parameters; for assessment purposes most important: their difficulty

Analogy

- A content based analogy (Fayers & Machin, 2007):
  - if we already know the answer to "Are you able to walk a short distance?"
  - it might not be interesting anymore to ask "Can you run a long distance?"

Building a CAT instrument: Item pool

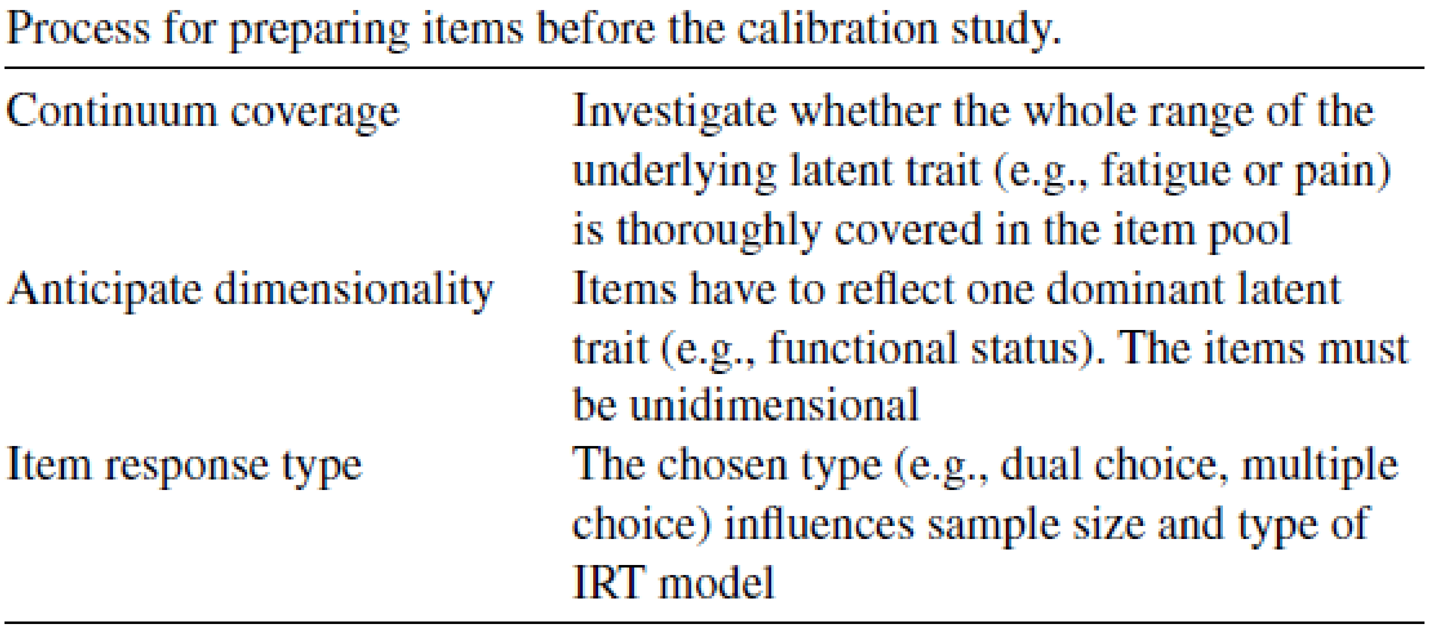
- The steps in item bank development are similiar to those for a usual measure
- main difference is that in an application you want a huge pool of items (called "item bank") to select your items from in practice and not only e.g. 40 like in a traditional test
- this makes it necessary that you in the very beginning have really many items to select from

Building a CAT instrument: Item pool

- this first step in developing a CAT application is called "item pool"
- it consists of all items that were seen as eligible for measuring the latent dimension
  - may stem from other questionnairs
  - qualitative work, interviews, focus groups
  - expert opinions, etc.
- like in any other test development!

Building a CAT instrument: Item pool

- some general suggestions on that:



Process for preparing items before the calibration study.

Continuum coverage	Investigate whether the whole range of the underlying latent trait (e.g., fatigue or pain) is thoroughly covered in the item pool
Anticipate dimensionality	Items have to reflect one dominant latent trait (e.g., functional status). The items must be unidimensional
Item response type	The chosen type (e.g., dual choice, multiple choice) influences sample size and type of IRT model

Walker et al. (2010)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Building a CAT instrument: Item pool

- Example from Fliege et al., 2005:
  - goal was to develop a CAT application for "depression" as one of the major health relevant outcomes
  - 144 items were used as pool
  - these were administered to N = 3270 patients in two overlapping sets of items (linking; "random missing by design"; e.g. Holman et al., 2003)

Building a CAT instrument: Calibration study

- These items are then presented to a (big) sample of respondents from the target population
- with data from this population all kind of checks (see last days) are performed:
  - establishing validity, reliability, appropriate dimensionality etc.
  - getting rid of non-fitting items
  - analyzing DIF
  - and come to final estimates of the item parameters

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Building a CAT instrument: Calibration study

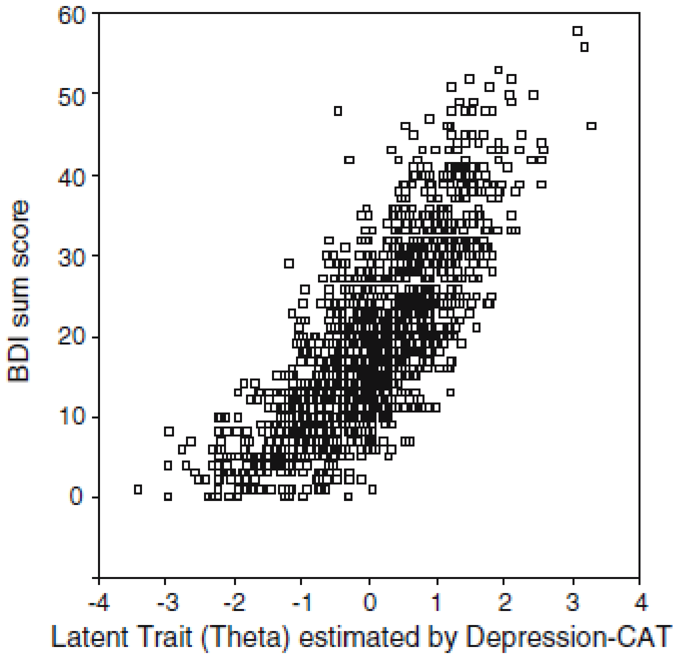
- in the Fliege et al. (2005) study:
  - face validity of items by experts; κ coefficients
  - IRT, dimensionality, monotonicity: GPCM, CFA, residual correlations
  - DIF with logistic regression approach
  - linking of items from the two versions
  - 64 items remained and formed the provisional "item pool"

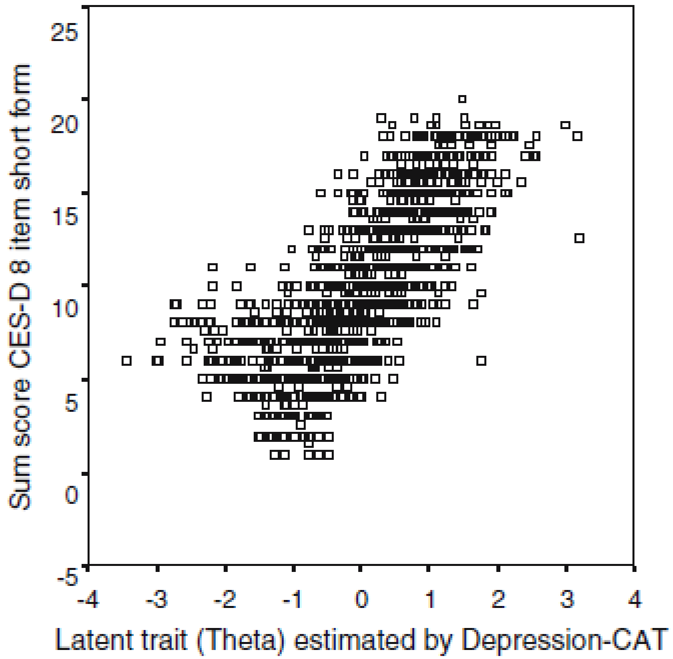
Building a CAT instrument: Calibration study

- also the properties of the items should be explored: does "going CAT" enhance effectiveness?
- example Fliege et al., 2005:
  - Simulation study A: simulated examinees with $-2 \leq \theta \leq 2$; the predefined precision of SE ≤ .32 ($\alpha = .9$) was reached after M = 7.15 (SD = 1.39); outside this range distinctly higher (M = 27.77; SD = 10.75)
  - Simulation study B: real patients data; M = 6.12 (SD = 2.11) for the same precision criterion; outside again distinctly higher

Building a CAT instrument: Calibration study

- concurrent validity of CAT-scores in Fliege et al., 2005:

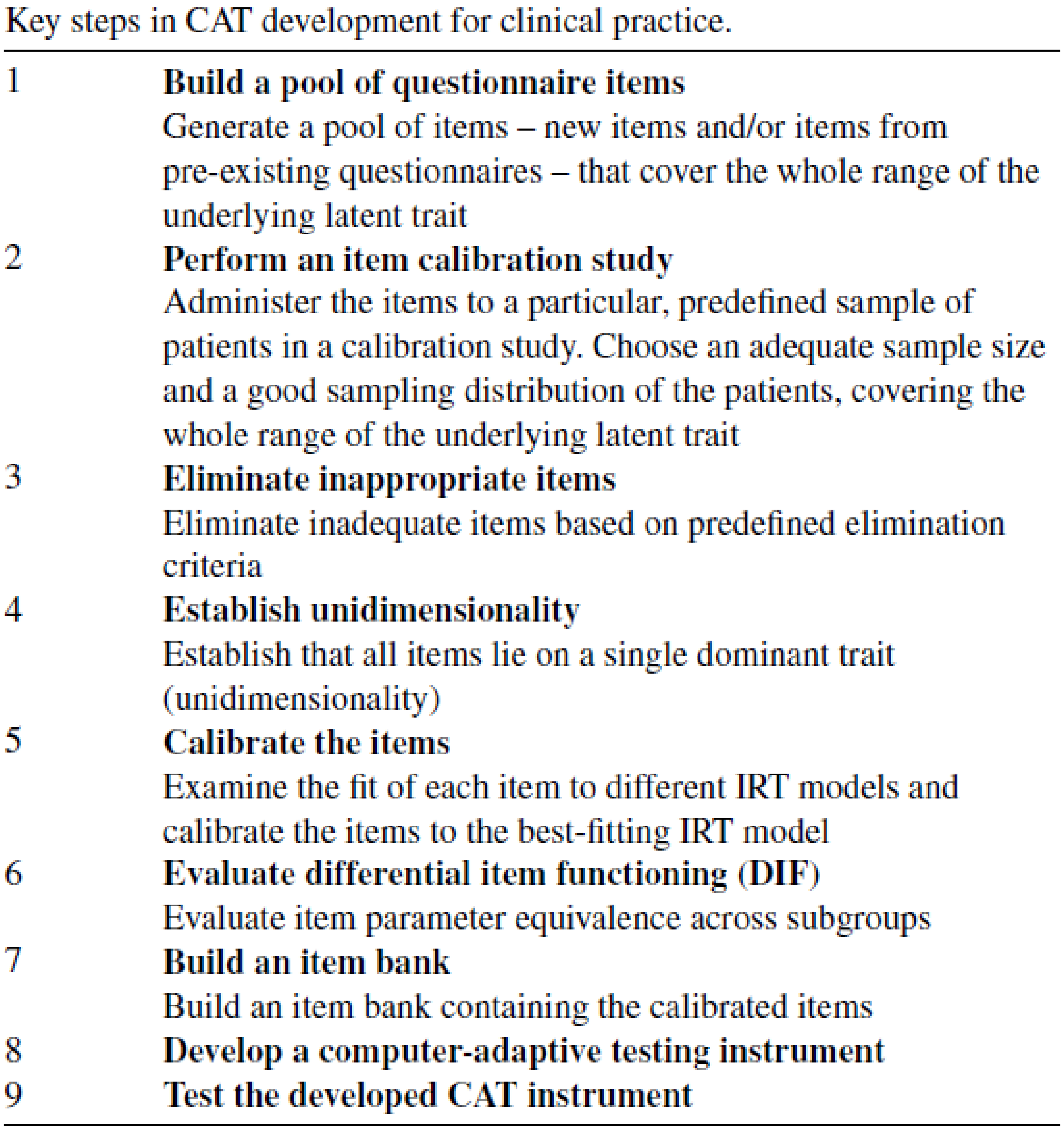




UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Building a CAT instrument: Validation study/ studies

- Since any of the item selection steps could simply be optimizing the test on sample variation, VALIDATION is needed:
  - simplest with a subset of the total sample
  - better with newly collected data
- since people might get used to the items, the population or the meaning of the items changes, etc. a recalibration has to be done every now and then



Key steps in CAT development for clinical practice.

1	**Build a pool of questionnaire items** Generate a pool of items – new items and/or items from pre-existing questionnaires – that cover the whole range of the underlying latent trait
2	**Perform an item calibration study** Administer the items to a particular, predefined sample of patients in a calibration study. Choose an adequate sample size and a good sampling distribution of the patients, covering the whole range of the underlying latent trait
3	**Eliminate inappropriate items** Eliminate inadequate items based on predefined elimination criteria
4	**Establish unidimensionality** Establish that all items lie on a single dominant trait (unidimensionality)
5	**Calibrate the items** Examine the fit of each item to different IRT models and calibrate the items to the best-fitting IRT model
6	**Evaluate differential item functioning (DIF)** Evaluate item parameter equivalence across subgroups
7	**Build an item bank** Build an item bank containing the calibrated items
8	**Develop a computer-adaptive testing instrument**
9	**Test the developed CAT instrument**

e.g. Walker, Böhnke, Strasser & Cerny (2010)

Cross cutting thoughts on "short scales"

- Recommendations for the selection of items for short scales in clinical research (Meier, 1997):

1. Items should be grounded in theory
2. several items should be used
3. ceiling and floor effects should be avoided
4. in intervention research: items should be able to detect change: in expected direction and compared to a group that does not change and is supposed not to change
5. an item should not discriminate between treatment groups at pre-test
6. cross-validation

- All these conditions are fulfilled for the CAT application; interestingly Meier wrote with view to CTT!

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Cross cutting thoughts

- the same questions apply to the development process of a CAT as to any other test
- main reason why this seems too much work is: for other purposes it is usually ignored!

CAT application

- After the calibration and the validation it can be said that the "item bank" exists
- both previous studies do not need any soft-/ hardware in particular – but it surely helps if at least part of the items are administered in the way they will be in the CAT application
- now the item bank can be put into use

CAT application

- CAT application means: an examinee should be tested
- and this with a shorter than usual instrument
- so, the examinee sits in front of a computer / gets the hendheld device and...

Steps of CAT application

1. Choose appropriate set of **start**ing items (one or several) for the examinee
2. based on the estimate new items are administered until the "stopping rule" is fulfilled (**test**)
3. **stop** presentation of new items when "stopping rule" is fulfilled
4. present **final** estimate of latent construct (plus additional information)

Starting step

- originally presentation of an item in the middle of the scale
- can be improved:
  - several items (esp. when test dichotomous)
  - incorporating prior information on the examinee (e.g. results from other tests; information that was relevant to DIF items)

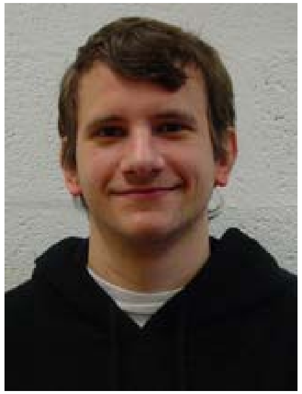
Test step

- estimate the construct from the starting step and then select item(s) for next presentation (from the pool of not yet administered ones):
  - those items with the highest information function for the provisional persons estimate
  - those items that reduce the variance of the estimate maximally after the response was added to the current response pattern
  - those items whose difficulty levels are closest to the provisional estimate
  - random draw

Stopping step

- the administration of new items is stopped e.g. when:
  - length: a fixed / maximal number of items is administered
  - precision: the SE / CI of the estimate falls below a specific criterion
  - classification: when a specific diagnostic value on the latent trait can be excluded (e.g. 0.5 as cut-off between clinical and non-clinical populations)

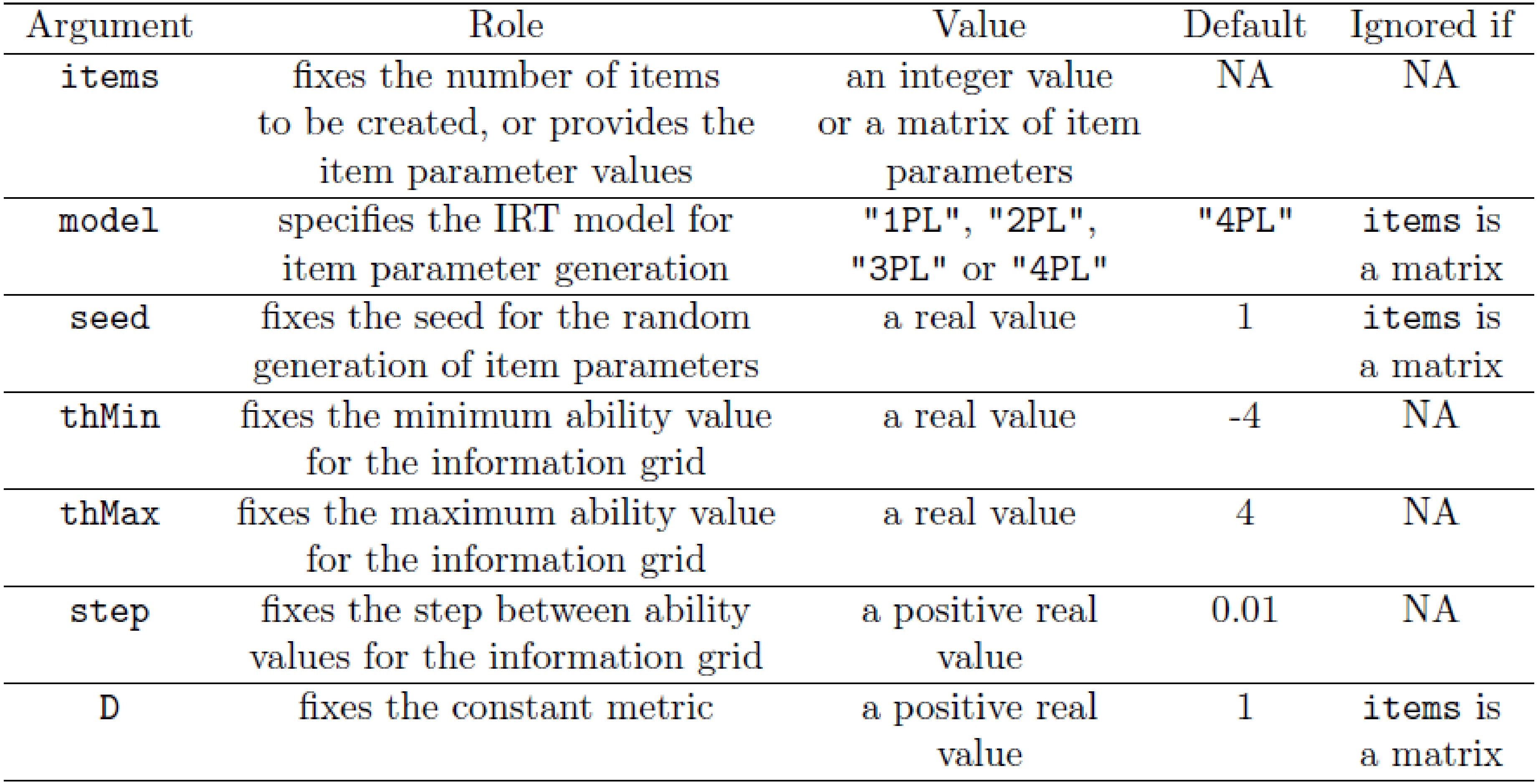
catR practical



- again we will use a package developed by David Magis (also difR; together with Gilles Raîche, UQAM)
- package is developed to estimate relevant statistics for CAT with an existing item bank
- for the package and tables on following slides: Magis & Raîche (2011, manuscript under review) & Magis & Raîche (2011)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

`createItemBank`



Argument	Role	Value	Default	Ignored if
`items`	fixes the number of items to be created, or provides the item parameter values	an integer value or a matrix of item parameters	NA	NA
`model`	specifies the IRT model for item parameter generation	`"1PL"`, `"2PL"`, `"3PL"` or `"4PL"`	`"4PL"`	`items` is a matrix
`seed`	fixes the seed for the random generation of item parameters	a real value	1	`items` is a matrix
`thMin`	fixes the minimum ability value for the information grid	a real value	-4	NA
`thMax`	fixes the maximum ability value for the information grid	a real value	4	NA
`step`	fixes the step between ability values for the information grid	a positive real value	0.01	NA
`D`	fixes the constant metric	a positive real value	1	`items` is a matrix

Magis & Raîche, 2011 in prep.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

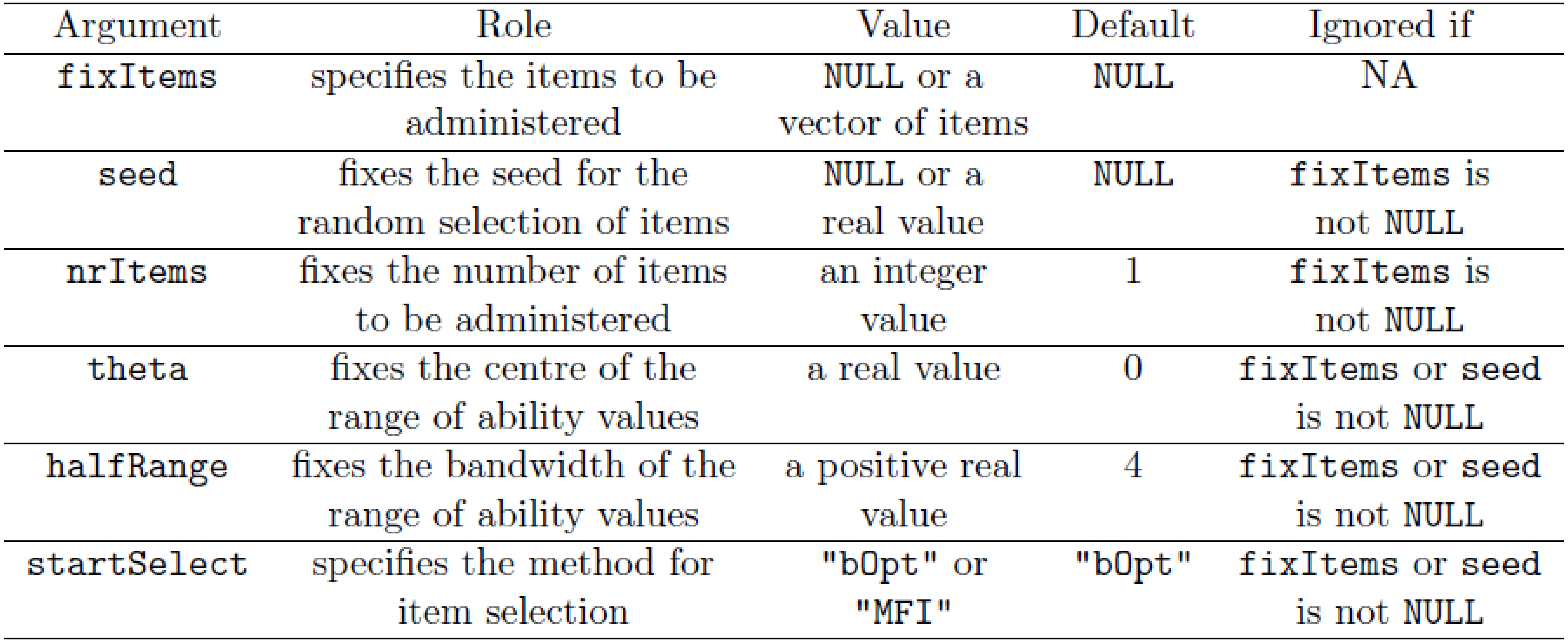
createItemBank

- Now we create our own item bank with 500 items, 2PL and in the range between -4 and 4 on the latent construct

```
Bank <- createItemBank(items =
  500, model = "2PL", thMin = -4,
  thMax = 4,step = 0.04)
```

- instead of creating one in this step also a real item bank can be read into catR!

start



Argument	Role	Value	Default	Ignored if
`fixItems`	specifies the items to be administered	`NULL` or a vector of items	`NULL`	NA
`seed`	fixes the seed for the random selection of items	`NULL` or a real value	`NULL`	`fixItems` is not `NULL`
`nrItems`	fixes the number of items to be administered	an integer value	1	`fixItems` is not `NULL`
`theta`	fixes the centre of the range of ability values	a real value	0	`fixItems` or `seed` is not `NULL`
`halfRange`	fixes the bandwidth of the range of ability values	a positive real value	4	`fixItems` or `seed` is not `NULL`
`startSelect`	specifies the method for item selection	`"bOpt"` or `"MFI"`	`"bOpt"`	`fixItems` or `seed` is not `NULL`

Magis & Raîche, 2011 in prep.

start

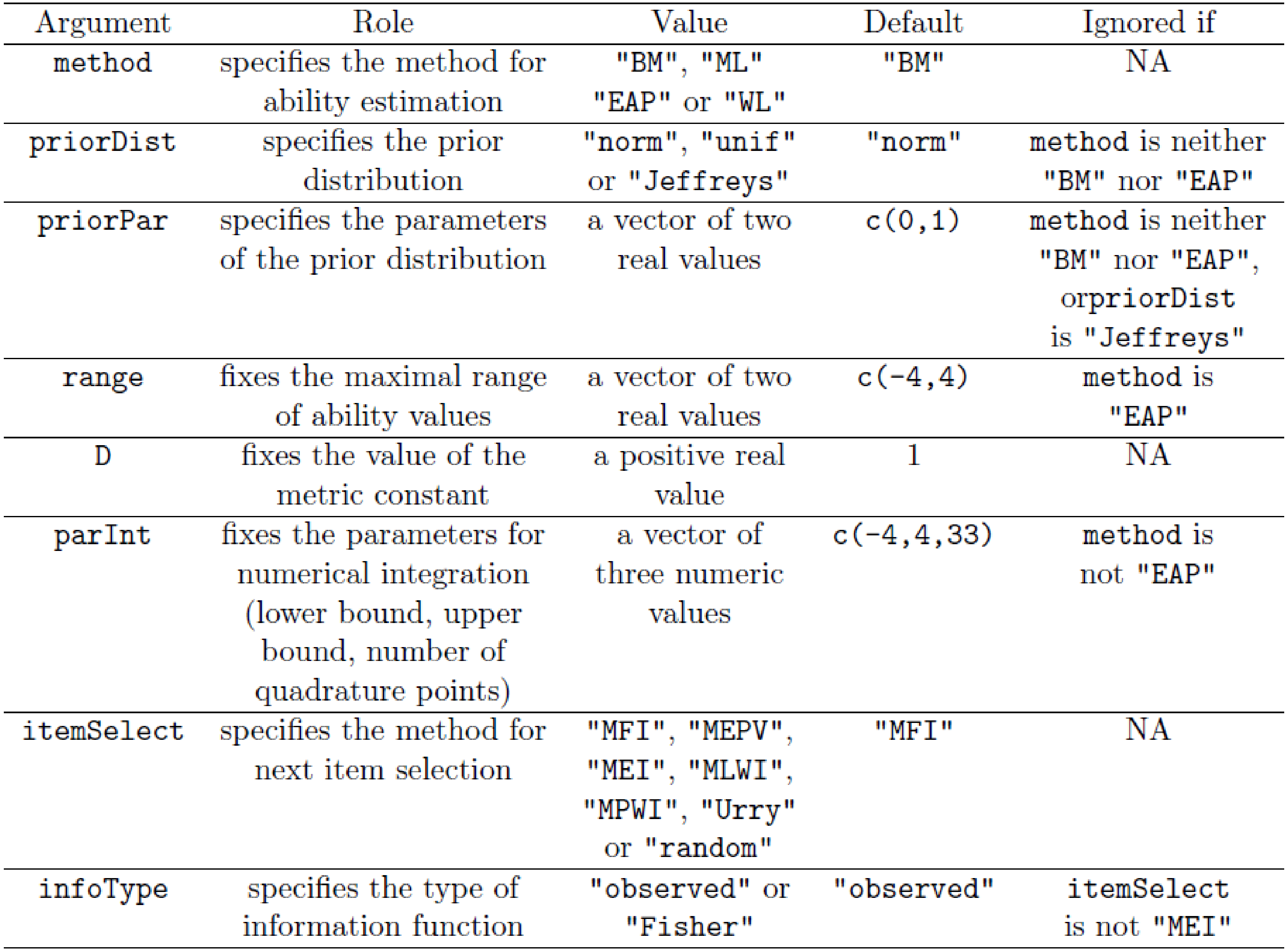
- now we define our own starting strategy: presenting 3 randomly chosen items (seed is set) and maximizing the information function

```
Start <- list(seed=1284, nrItems
  = 3, startSelect = "MFI")
```

- (this only defines options to be read in the actial analysis!)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

test



Argument	Role	Value	Default	Ignored if
method	specifies the method for ability estimation	"BM", "ML" "EAP" or "WL"	"BM"	NA
priorDist	specifies the prior distribution	"norm", "unif" or "Jeffreys"	"norm"	method is neither "BM" nor "EAP"
priorPar	specifies the parameters of the prior distribution	a vector of two real values	c(0,1)	method is neither "BM" nor "EAP", orpriorDist is "Jeffreys"
range	fixes the maximal range of ability values	a vector of two real values	c(-4,4)	method is "EAP"
D	fixes the value of the metric constant	a positive real value	1	NA
parInt	fixes the parameters for numerical integration (lower bound, upper bound, number of quadrature points)	a vector of three numeric values	c(-4,4,33)	method is not "EAP"
itemSelect	specifies the method for next item selection	"MFI", "MEPV", "MEI", "MLWI", "MPWI", "Urry" or "random"	"MFI"	NA
infoType	specifies the type of information function	"observed" or "Fisher"	"observed"	itemSelect is not "MEI"

Magis & Raîche, 2011 in prep.

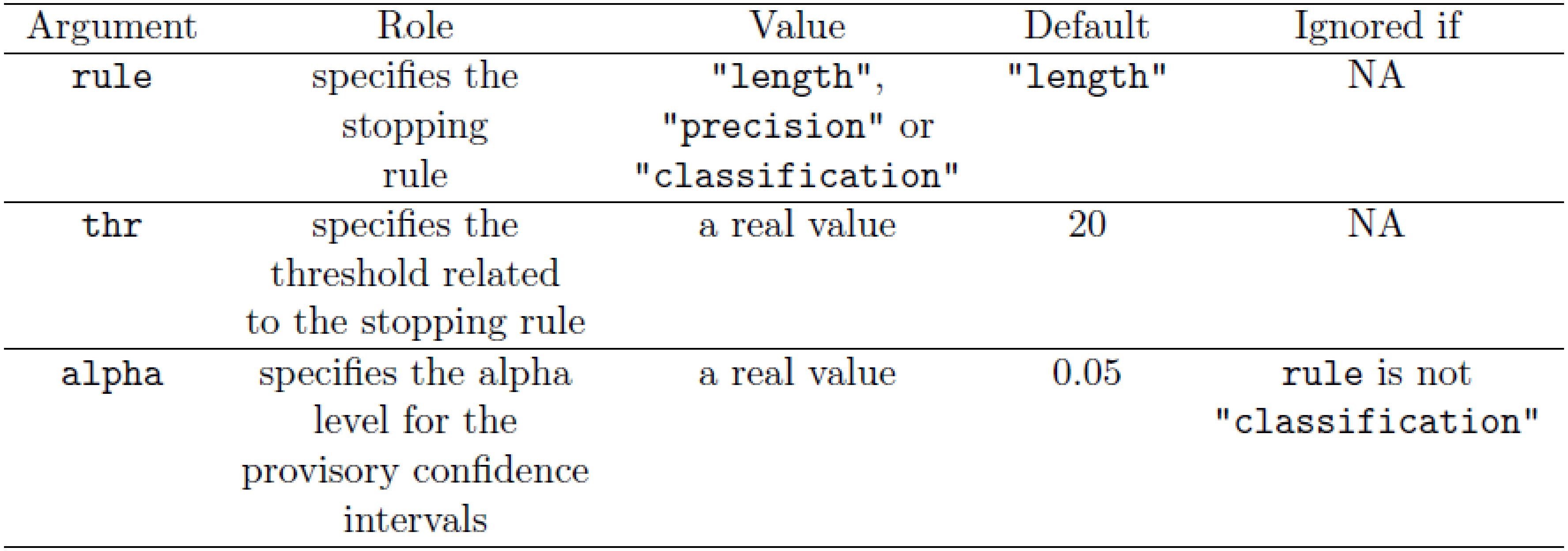
test

- "test" defines how the actual test administration would be handled: which items are presented, which selection criteria are applied etc...
- our simple rule will contain the following:
- Baysian Modal estimation of the ability with a uniform prior (-1,1)

```
Test <- list(method = "BM",
  priorDist="unif", priorPar=c(-5,5),
  itemSelect = "MFI", range=c(-5,5))
```

- (this only defines options to be read in the actial analysis!)

stop



Argument	Role	Value	Default	Ignored if
`rule`	specifies the stopping rule	`"length"`, `"precision"` or `"classification"`	`"length"`	NA
`thr`	specifies the threshold related to the stopping rule	a real value	20	NA
`alpha`	specifies the alpha level for the provisory confidence intervals	a real value	0.05	`rule` is not `"classification"`

Magis & Raîche, 2011 in prep.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

stop

- Creating our own rule when to stop the presentation of items:
- we use the measurement accuracy with a threshold of SE=.315 (alpha = .90)

```
Stop <- list(rule = "precision", thr = 0.315)
```

- (this only defines options to be read in the actial analysis!)
- (Babcock & Weiss, 2009, GMAC conference)

$$\mathrm{SEM} = s_{\mathrm{obs}}\,(1 - \rho_{xx})^{½}$$

UNIVERSITY OF
CAMBRIDGE
The Psychometrics Centre

final

- the command for the final estimation can contain any commands of the "test" part
- and additionally an argument `alpha` for the CI around the final estimation
- we use:

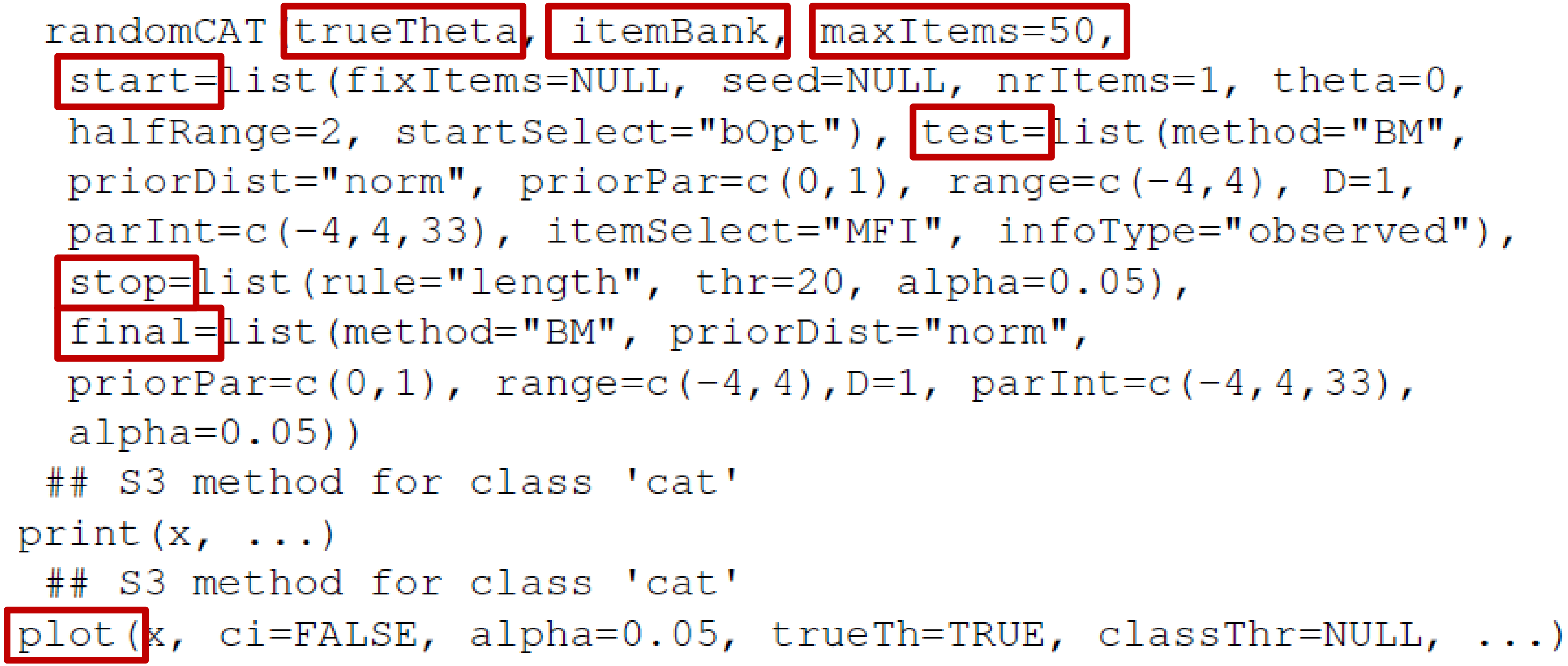
```
Final <- list(method = "WL", alpha = 0.05,
  range=c(-5,5))
```

- (this only defines options to be read in the actial analysis!)

Run catR

- after these specifications we can actually let the first examinee "take" the questionnaire
- for this the `randomCAT()` command is used

randomCAT



```
 randomCAT(trueTheta, itemBank, maxItems=50,
  start=list(fixItems=NULL, seed=NULL, nrItems=1, theta=0,
  halfRange=2, startSelect="bOpt"), test=list(method="BM",
  priorDist="norm", priorPar=c(0,1), range=c(-4,4), D=1,
  parInt=c(-4,4,33), itemSelect="MFI", infoType="observed"),
  stop=list(rule="length", thr=20, alpha=0.05),
  final=list(method="BM", priorDist="norm",
  priorPar=c(0,1), range=c(-4,4),D=1, parInt=c(-4,4,33),
  alpha=0.05))
 ## S3 method for class 'cat'
print(x, ...)
 ## S3 method for class 'cat'
plot(x, ci=FALSE, alpha=0.05, trueTh=TRUE, classThr=NULL, ...)
```

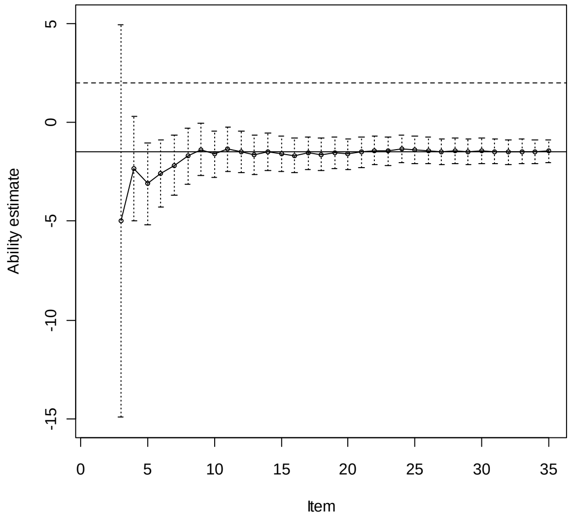
randomCAT

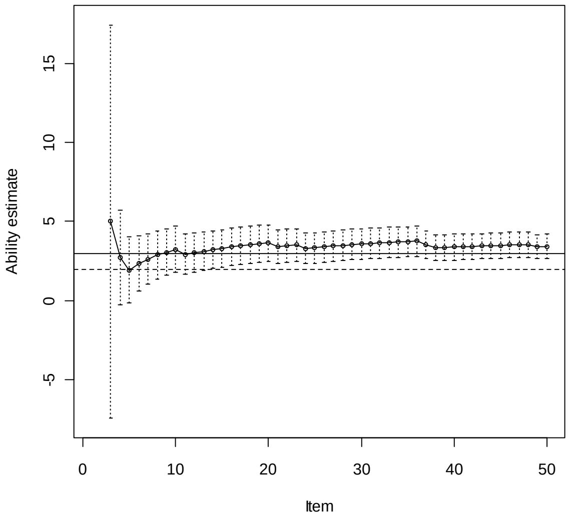
- Enter first test respondent with an actual Theta = -1.5
- the test uses our prior specifications:

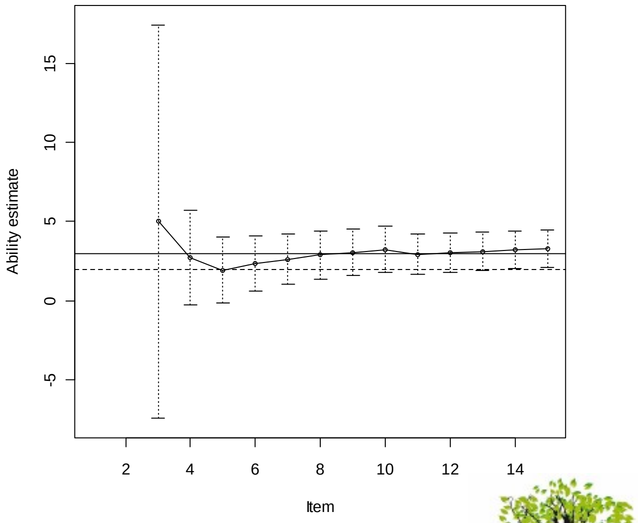
```
res <- randomCAT(trueTheta = -1.5, maxItems=50,
  itemBank = Bank, start = Start,test = Test,
  stop = Stop, final = Final)

plot(res,ci = TRUE, trueTh = TRUE, classThr =
  2)
```

randomCAT







UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

All commands toether

```
Bank <- createItemBank(items = 500, model = "2PL",
    thMin = -4, thMax = 4,step = 0.04)
Start <- list(seed=1284, nrItems = 3, startSelect = "MFI")
Test <- list(method = "BM", priorDist="unif", priorPar=c(-
    5,5), itemSelect = "MFI", range=c(-5,5))
Stop <- list(rule = "precision", thr = 0.3)
Final <- list(method = "WL", alpha = 0.05, range=c(-5,5))
res <- randomCAT(trueTheta = -1.5, maxItems=50,
    itemBank = Bank, start = Start,test = Test, stop = Stop,
    final = Final)
plot(res,ci = TRUE, trueTh = TRUE, classThr = 2)
```

Practical

- Please explore the different options / settings

catR

- The package contains everything that is necessary to base a real application on
  - Item bank can be read in instead of simulated
  - `nextItem(Bank, theat, criterion="MFI")` would actually generate the next item from the bank as output (which could be used as input for presentation software)

CAT at Psychometrics Center

- this was actually done by the Psychometrics Center (Michal Kosinski, John Rust and others):

http://www.psychometrics.cam.ac.uk/page/300/concerto-testing-platform.htm

Some final thoughts

- Advantages:
  - fewer items needed
  - items presented maybe more relevant to the examinee
  - minimizing floor & ceiling effects
  - flexible precision

- Disadvantages
  - comprehensive item bank (& before that: pool) has to be generated
  - large number of patients in calibration
  - implementation may be more difficult

Some final thoughts

- think about patients / examinees' computer skills and familiarity; dexterity? eyesight?
- Spend time and energy on hardware choice!!!!
- How will the data be saved and which other possible uses of the data will be made (e.g. patient-oriented psychotherapy research; Lutz, 2002)?

References

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Presented at the 2009 GMAC (R) Conference on CAT. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09babcock.pdf

Fliege, H., Becker, J., Walter, O., Bjorner, J., Klapp, B., & Rose, M. (2005). Development of a Computer-adaptive Test for Depression (D-CAT). *Quality of Life Research*, *14*, 2277-2291.

Holman, R., Lindeboom, R., Glas, C. A. W., Vermeulen, M., & de Haan, R. J. (2003). Constructing an item bank using item response theory: The AMC linear disability score project. *Health Services & Outcomes Research Methodology*, *4*, 19-33.

Lutz, W. (2002). Patient-focused psychotherapy research and individual treatment progress as scientific groundwork for an empirical based clinical practice. *Psychotherapy Research*, *12*, 251-273.

Magis, D., & Raîche, G. (2011). catR: An R package for computerized adaptive testing. Applied Psychological Measurement. in press/online first

Meier, S. (1997). Nomothetic Item Selection Rules for Tests of Psychological Interventions. *Psychotherapy Research*, *7*, 419-427.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, *35*, 5-25.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates: Hillsdale, New Jersey.

Walker, J., Böhnke, J. R., Cerny, T., & Strasser, F. (2010). Development of symptom assessments utilising item response theory and computer-adaptive testing—A practical method based on a systematic review. *Critical reviews in Oncology/Hematology*, *73*, 47-67.