Open-source CAT software: R packages and Concerto

David Magis

Department of Psychology, KU Leuven, Belgium

Department of Education, University of Liège, Belgium

Joint work with

Juan Ramon Barrada

Universidad Zaragoza, Teruel, Spain

ISOQOL-NL, Utrecht, March 12, 2014

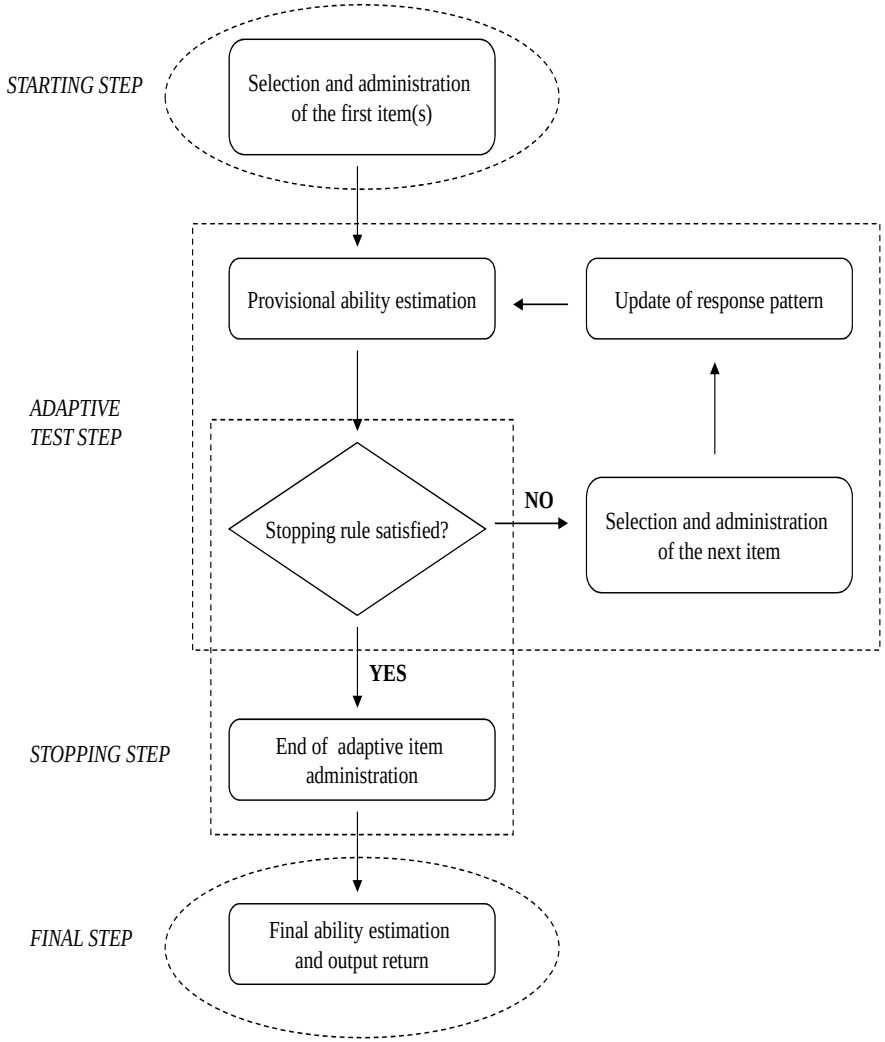
Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

1. CAT software



1. CAT software

Three main categories of CAT software:

1. Commercial software:
   - CATSim (Assessment Systems Corporation, 2012)
   - Adaptest (https://adaptest.vpgcentral.com/)
   - Assessment Center (https://www.assessmentcenter.net/)
   - ...
2. Open-source software (mostly based on R):
   - Firestar (Choi, 2009)
   - catIrt (Nydick, 2013)
   - catR (Magis & Raiche, 2012)
   - MAT (Choi, 2011)
   - ...

1. CAT software

Three main categories of CAT software:

3. Open-source web-based platform:
   - Concerto (Kosinski & Rust, 2011)

Commercial software designed for specific assessment purposes

R packages mostly designed for simulation studies

Web-based platform built for real online CAT assessments

1. CAT software

Purpose of this talk:

- to briefly present several CAT-related software
- to focus on their goals and options

Software to be presented (in detail): *catR* and *catIrt*

R package *MAT* and web-based platform *Concerto* briefly outlined

Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

2. R packages *catR* and *catIrt*

R packages to generate response patterns under CAT scenarios

Advantages:

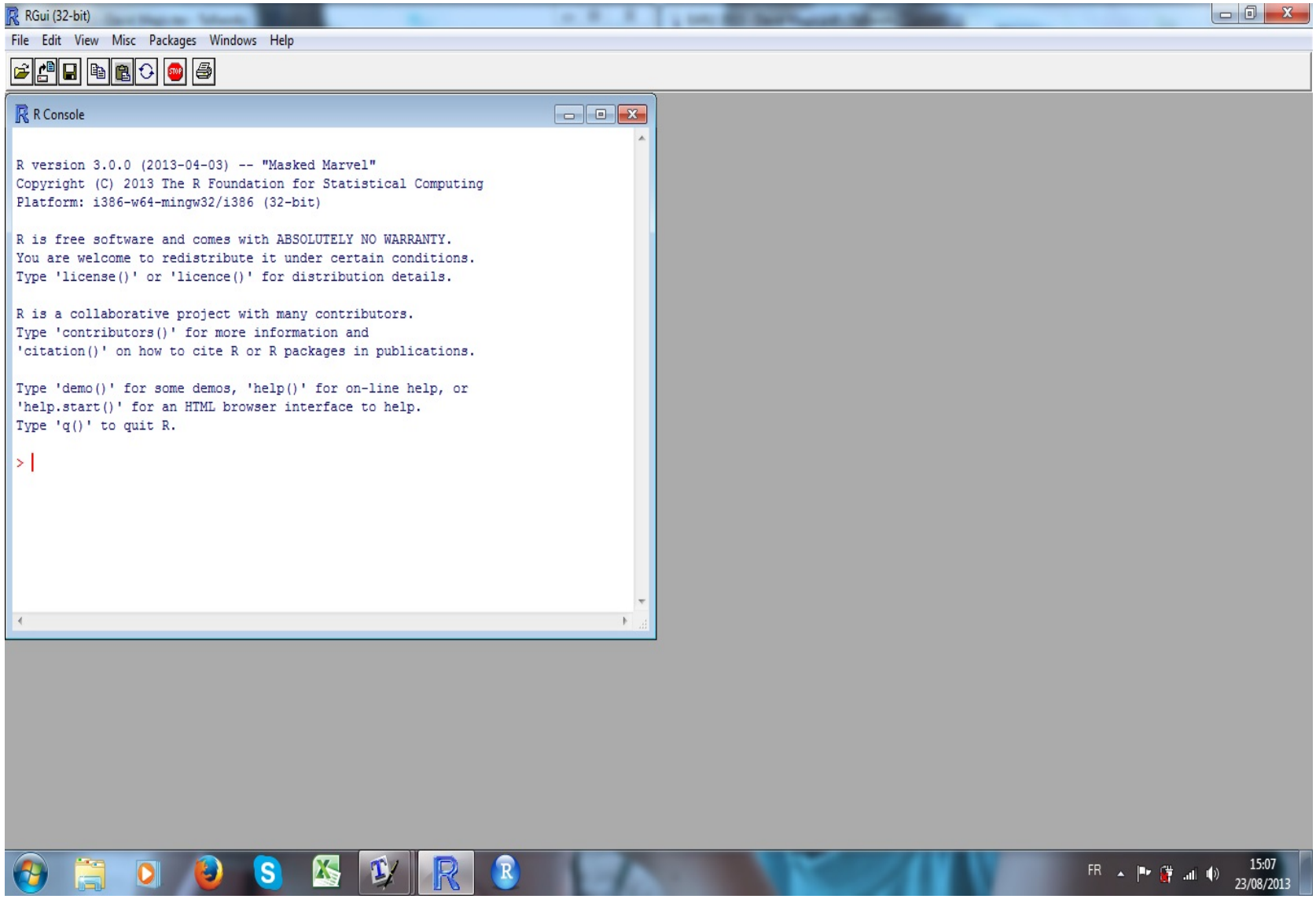
- Flexibility
- Absolute control of all CAT steps (selecting first items, next item selection, proficiency estimation, stopping rules, ...)
- Easy to generate lots of response patterns

Drawbacks:

- Some knowledge of R is necessary
- Limited to R

$\Rightarrow$ Most useful for intensive simulation studies

2. R packages *catR* and *catIrt*



2. R packages *catR* and *catIrt*

Currently available versions (on CRAN):

- *catR*: 2.6 (released: December 3rd, 2013)
- *catIrt*: 0.4-1 (released: May 31st, 2013)

Beta version of *catR* 3.0 (released: last week)

Not yet available on CRAN

Holds recent improvements

Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

3. Comparisons

Comparisons of *catR* (2.6), *catR* (3.0) and *catIrt*

Comparison criteria:

- IRT models
- Proficiency estimators
- First item(s) selection
- Next item selection
- Stopping rules
- Miscellaneous topics

3. Comparisons

IRT models:

Model	*catR 2.6*	*catR 3.0*	*catIrt*
1PL	X	X	X
2PL	X	X	X
3PL	X	X	X
4PL	X	X	
GRM		X	X
M-GRM		X	
PCM		X	X
GPCM		X	
RSM		X	
NRM		X	

3. Comparisons

Proficiency estimators:

- All packages can estimate proficiencies with
  - Maximum likelihood (ML)
  - Bayes modal or Maximum a posteriori (MAP)
  - Expected a posteriori (EAP)
  - Weighted likelihood (WL)
- Prior distributions for MAP and EAP:

Model	*catR 2.6*	*catR 3.0*	*catIrt*
Normal	X	X	X
Uniform	X	X	X
Jeffreys	X	X	

3. Comparisons

First item(s) selection:

- *catR* and *catIrt* can select one or several items to start the CAT
- Selection rules:
  - at random
  - selected item(s) by the user (*catR* only)
  - with respect to starting proficiency levels
  - with respect to optimal item selection
- Optimal selection based on maximum information (*catR*) or more sophisticated versions (*catIrt*)

3. Comparisons

Next item selection:

Model	*catR 2.6*	*catR 3.0*	*catIrt*
MFI	X	X	X
MLWI	X	X	X
MPWI	X	X	X
MEI	X	X	
MEPV	X	X	

- MFI: Maximum Fisher information
- MLWI: Maximum likelihood weighted information
- MPWI: Maximum posterior weighted information
- MEI: Maximum expected information
- MEPV: Minimum expected posterior variance

3. Comparisons

Next item selection:

Model	*catR 2.6*	*catR 3.0*	*catIrt*
KL		X	
KLP		X	
Progressive		X	
Proportional		X	
At random	X	X	X

- KL: Kullback-Leibler information
- KLP: Posterior Kullback-Leibler information

3. Comparisons

Stopping rules:

- Three kinds of stopping rules:
  - Length: stop after n items administered
  - Precision: stop when SEM of proficiency small enough
  - Classification: stop when classification is possible
- Classification rules:
  - Confidence interval method (Kingsbury & Weiss, 1983) in all packages
  - Sequential Probability Ratio Test (SPRT) (Eggen, 1999) only in *catIrt*
  - Generalized Likelihood Ratio (GLR) (Thompson, 2009) only in *catIrt*

3. Comparisons

Miscellaneous topics:

- All software allow for item exposure control:
  - Sympson-Hetter method (Sympson & Hetter, 1985) in *catIrt*
  - Randomesque method (Kingsbury & Zara, 1989) in *catR*
  - Progressive and proportional methods (for next item selection) in *catR 3.0*
- Content balancing (Kingsbury & Zara, 1989) is allowed in *catR*
- Several response patterns can be generated with one run in *catR 3.0* and *catIrt*
- Graphical output available

Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

4. A demo of *catR*

Small demo of the logic and functioning of *catR 3.0*

CAT design:

- Creation of a 200-item bank calibrated under Partial Credit Model (and at most 5 response categories per item)
- Generation of one CAT response pattern for true ability level $\theta = 0$
- 3 initial items selected randomly in the bank
- Next item selection by maximum Fisher information
- Interim ability estimation by weighted likelihood
- Stopping rule: stop when 10 items are administered
- Final ability estimation by maximum likelihood

4. A demo of *catR*

R code:

```
R> bank <- genPolyMatrix(nrItems = 200, nrCat = 5,
                         model = "PCM")
R> head(bank)

  deltaj1 deltaj2 deltaj3 deltaj4
1  -0.626   0.184      NA      NA
2  -0.836   1.595   0.330  -0.820
3   0.487   0.738   0.576      NA
4  -0.305   1.512   0.390  -0.621
5  -2.215   1.125      NA      NA
6  -0.045  -0.016   0.944      NA
```

4. A demo of *catR*

R code:

```
R> start <- list(nrItems = 3, seed = 1)
R> test <- list(method = "WL", itemSelect = "MFI")
R> stop <- list(rule = "length", thr = 10)
R> final <- list(method = "ML")

R> res <- randomCAT(trueTheta = 0, itemBank = bank,
                    model = "PCM", start = start,
                    test = test, stop = stop,
                    final = final)
```

4. A demo of *catR*

Output:

```
Random generation of a CAT response pattern

 Item bank calibrated under Partial Credit Model

 True ability level: 0

 Starting parameters:
   Number of early items: 3
   Early items selection: Random selection in item bank
   Items administered: 54, 75 and 114

 Adaptive test parameters:
   Next item selection method: maximum Fisher information
   Provisional ability estimator: Weighted likelihood estimator
```

4. A demo of *catR*

Output:

```
Stopping rule:
  Stopping criterion: length of test
  Maximum test length: 10 items

Item exposure control:
  Method: 'randomesque'
  Number of 'randomesque' items: 1

Content balancing control:
  No control for content balancing
```

4. A demo of *catR*

Output:

```
Adaptive test details:

Nr          1      2      3      4      5      6      7
Item       54     75    114    159    100     59    110
Resp.       2      0      1      0      0      3      2
Est.       NA     NA  0.264  0.111 -0.066 -0.027 -0.012
SE         NA     NA  0.332  0.279  0.266  0.232  0.208

Nr          8      9     10
Item      175     34    140
Resp.       2      1      3
Est.    0.027  0.029  0.082
SE       0.19  0.179  0.168
```

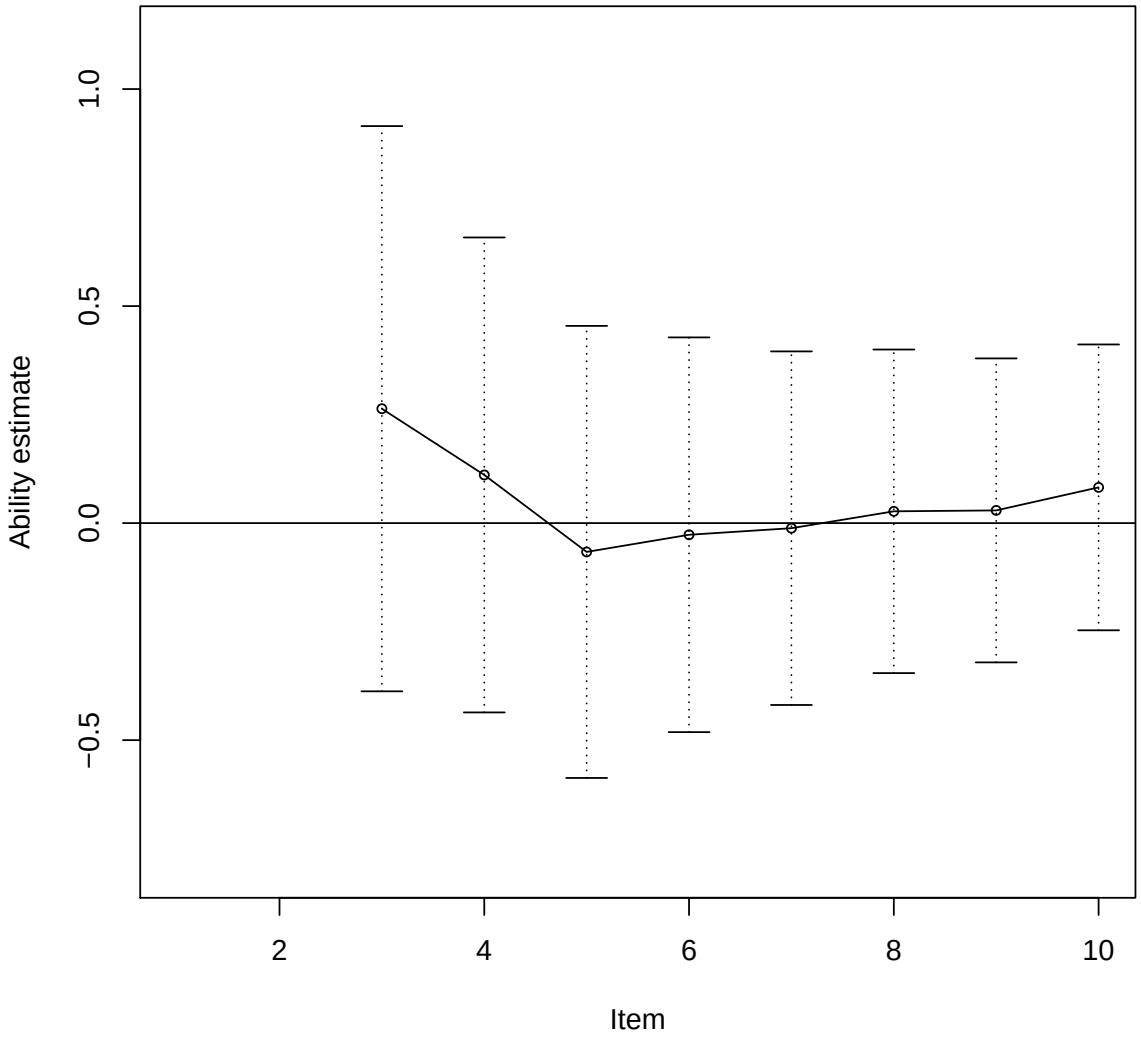
4. A demo of *catR*

Output:

```
Final results:
  Length of adaptive test: 10 items
  Final ability estimator: Maximum likelihood estimator
  Final range of ability values: [-4,4]
  Final ability estimate (SE): 0.079 (0.173)
  95% confidence interval: [-0.259,0.418]

Output was not captured!
```

4. A demo of *catR*



Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

5. R package *MAT*

Packages *catR* and *catIrt* make use of usual (dichotomous and polytomous) IRT models

Package *MAT* (for *Multidimensional Adaptive Testing*; Choi, 2011) incorporates the multidimensional 3-parameter logistic model (Reckase, 2009)

Promising for administering multidimensional CATs

(Apparently) not yet fully operational:

- Currently less options than *catR* and *catIrt*
- Item response matrix to be provided $\Rightarrow$ post-hoc simulations?

$\Rightarrow$ To be followed up

Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

6. Platform *Concerto*

catR, *catIrt* and *MAT* mostly designed to generate response patterns with given calibrated item bank

Item bank must be pre-calibrated and provided only in terms of item parameters

No formal item administration with these software

$\Rightarrow$ Need for software for "real" CAT administrations

6. Platform *Concerto*

Concerto is a web-based platform for CAT building and administration

Open-source platform with *catR* package as underworking routine

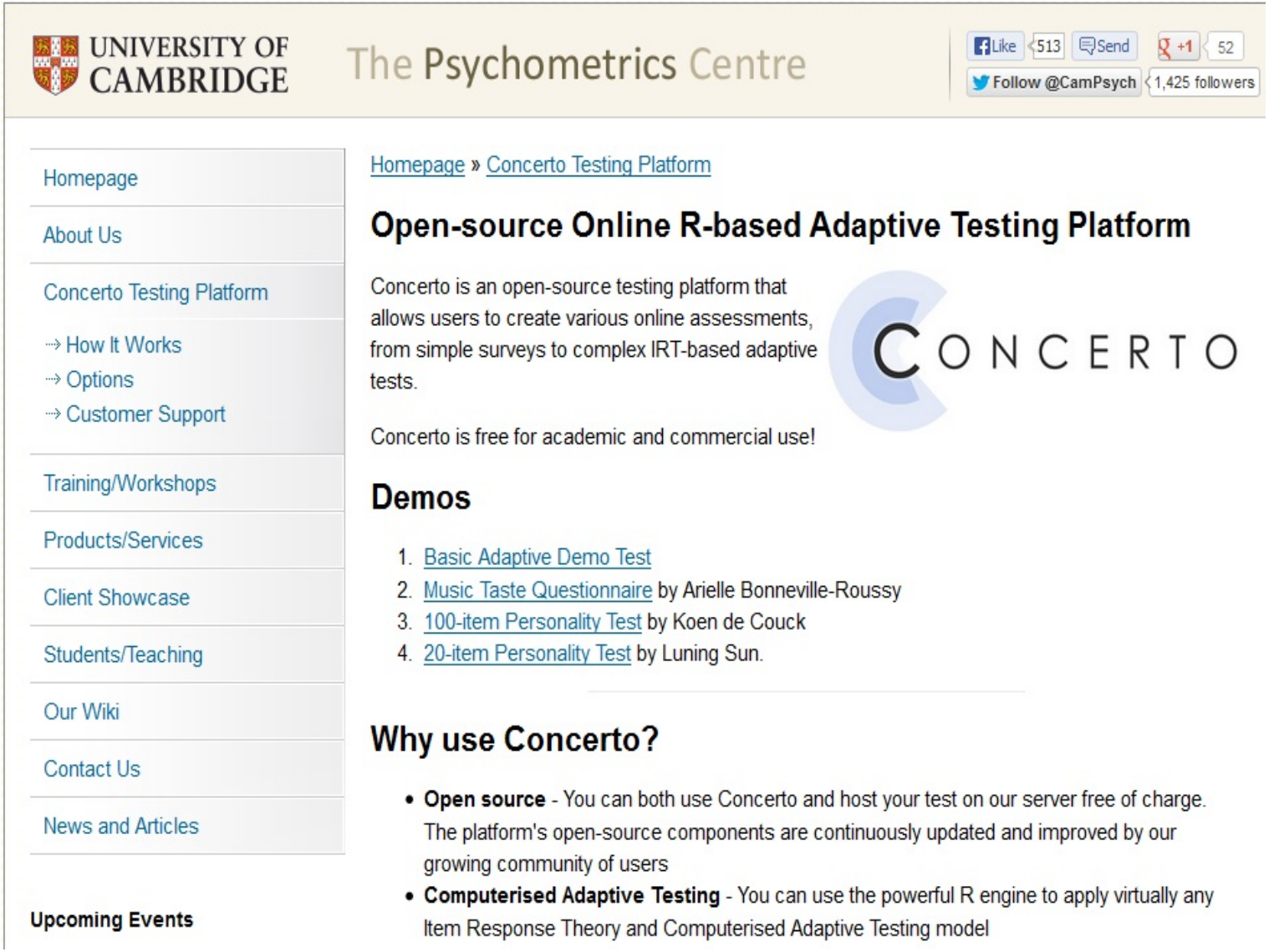
catR used for proficiency level estimation, first item(s) and next item selection

Concerto provides html menus and webpages to

- build items and create item banks
- incorporate additional item material (e.g. pictures, videos)
- specify the CAT options

CAT can be delivered through a variety of web browsers and on mobile devices (smartphones, tablets, ...)

6. Platform *Concerto*

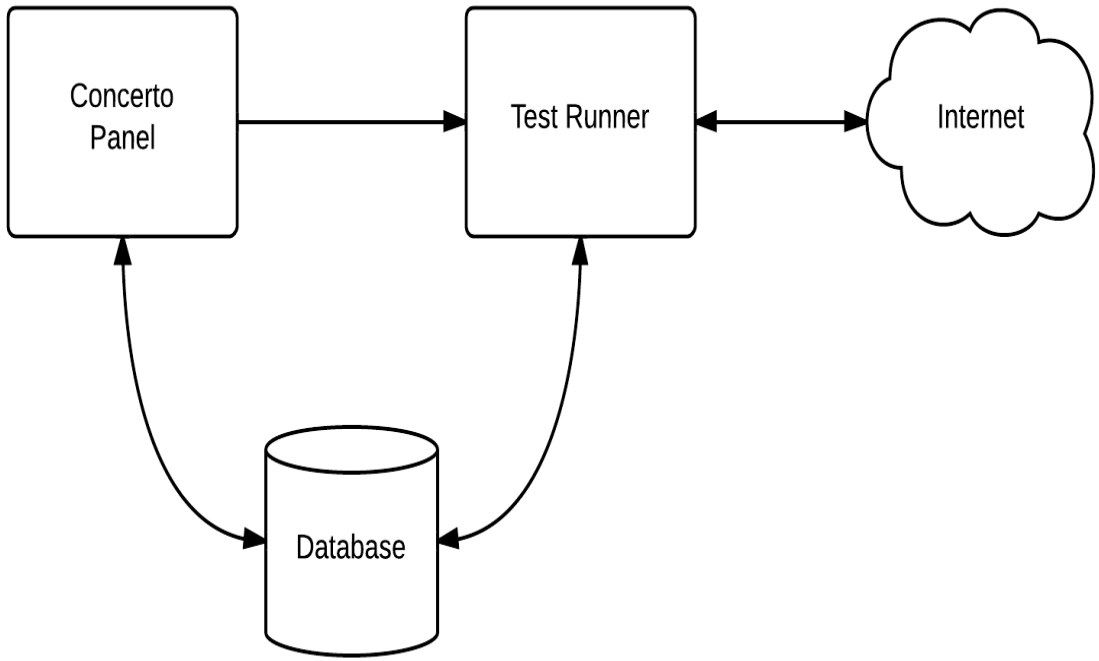


6. Platform *Concerto*

Concerto is built in two parts:

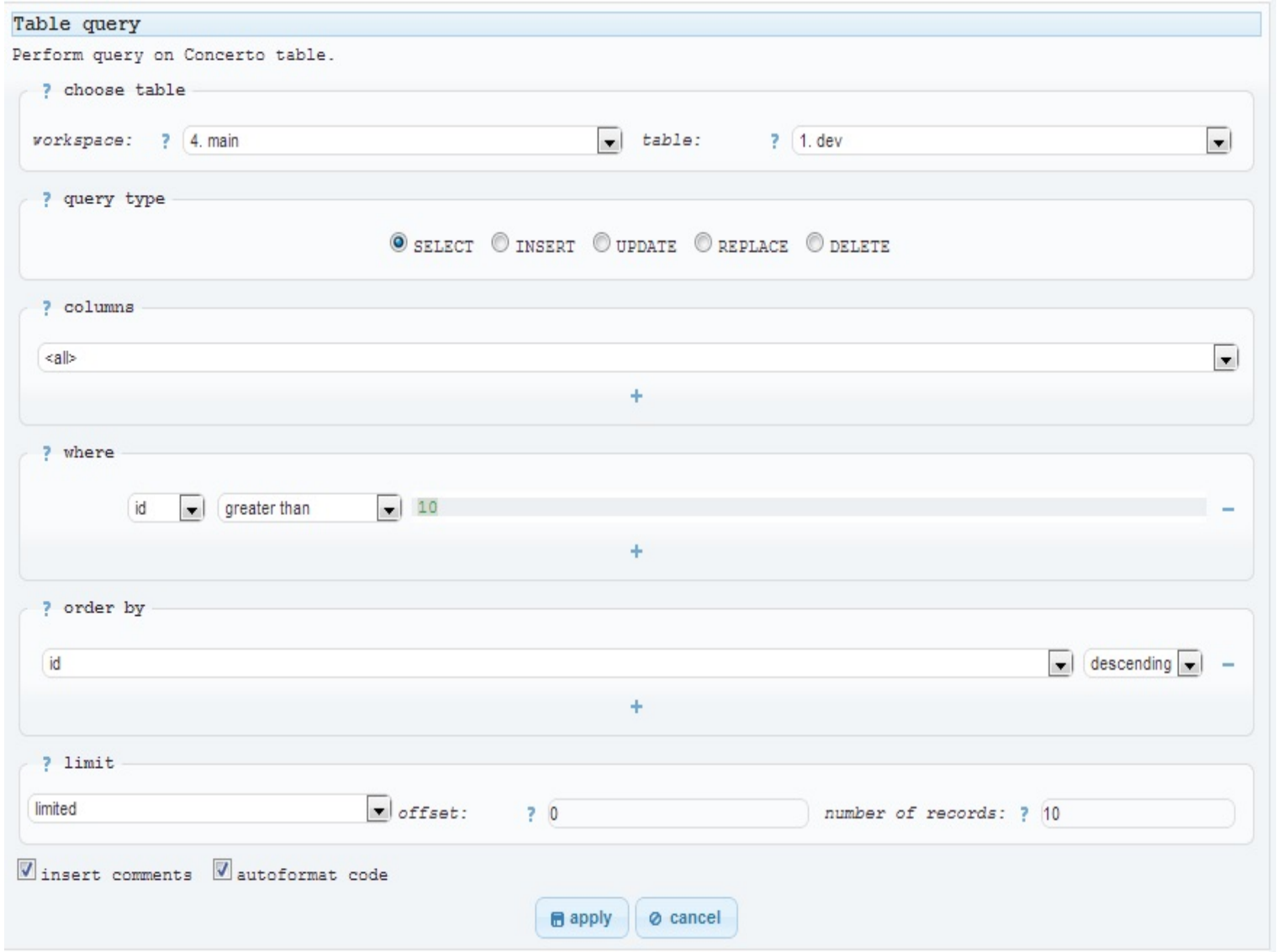
- *Concerto web panel*: online application
  - to prepare R-based scripts containing test logic,
  - that provides rich text editor,
  - interacting with databases for storing test items and their results,
  - that exports prepared tests to the Test Runner
- *Test Runner*: responsible for executing prepared tests and interacting with the user:
  - keeps the test state for the duration of the session, enabling resuming the test (in case of communication failure),
  - to be deployed on multiple machines to serve large numbers of participants

6. Platform *Concerto*



6. Platform *Concerto*

6. Platform *Concerto*



Outline:

1. CAT software
2. R packages *catR* and *catIrt*
3. Comparisons
4. A demo of *catR*
5. R package *MAT*
6. Platform *Concerto*
7. Final comments

7. Final comments

Recent CAT software mostly based on R development

- as packages (*catR* and *catIrt*)
- as front end-user interfaces with with back-end R routines such as *Firestar* (Choi, 2009)
- as web-based platforms for CAT delivery with back-end R routines (*Concerto*)

Both *catR* and *Concerto* are under development

Further benefitial updates of *Concerto*: inclusion of user interfaces to simplify item bank and test developments

ERC starting grant application related to *Concerto* updates (soon) submitted

7. Final comments

catR and *catIrt* available (freely) from CRAN:

http://cran.r-project.org

Concerto can be accessed from

http://www.psychometrics.cam.ac.uk/page/338/concerto-testing-platform.htm

References

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33*, 644–645.

Choi, S. W. (2011). *MAT: Multidimensional adaptive testing (MAT)*. R package version 0.1-3.

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

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.

References

Kosinski, M. & Rust, J. (2011). *The development of Concerto: An open source online adaptive testing platform.* Paper poresenetd at the International Association for Computerized Adaptive Testing, Pacific Grove, CA.

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.

Nydick, S. W. (2013). *catIrt: An R package for simulating IRT-based computerized adaptive tests.* R package version 0.4-1.

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

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.*

To get the slides:

http://hdl.handle.net/2268/163595

To get the beta version of *catR 3.0*:

david.magis@ppw.kuleuven.be

or

david.magis@ulg.ac.be

To get the slides:

http://hdl.handle.net/2268/163595

To get the beta version of *catR 3.0*:

david.magis@ppw.kuleuven.be
or
david.magis@ulg.ac.be

THANK YOU!

To get the slides:

http://hdl.handle.net/2268/163595

To get the beta version of *catR 3.0*:

david.magis@ppw.kuleuven.be
or
david.magis@ulg.ac.be

BEDANKT!